A dimensional summation account of polymorphous category learning

Andy J. Wills · **Lyn Ellett** · **Fraser Milton** · **Gareth Croft** · **Tom Beesley**

Abstract Polymorphous concepts are hard to learn, and this is perhaps surprising because they, like many natural concepts, have an overall similarity structure. However, the dimensional summation hypothesis (Milton & Wills, 2004) predicts this difficulty. It also makes a number of other predictions about polymorphous concept formation, which are tested here. In Experiment 1 we confirm the theory's prediction that polymorphous concept formation should be facilitated by deterministic pretraining on the constituent features of the stimulus. This facilitation is relative to an equivalent amount of training on the polymorphous concept itself. In with some other accounts, including a more general strategic account (Experiment 2), seriality of training and stimulus decomposition accounts (Experiment 3), and the role of errors (Experiment 4). The dimensional summation hypothesis provides the best account of these data. In Experiment 5, a further prediction is confirmed — the single feature pretraining effect is eliminated by a concurrent counting task. The current experiments suggest the hypothesis that natural concepts might be acquired by the deliberate serial summation of evidence. This idea has testable implications for classroom learning.

A polymorphous concept is one defined by an *n-out-of-m* rule (Ryle, 1951). For example, in a set of geometric shapes that vary in the three stimulus dimensions of size (large or small), shape (square or triangle), and shade (black or white), one polymorphous category would be defined by the rule "category A is at least two of small, square, and white". In an undergraduate project subsequently published in *Nature*, Stephen Lea and colleagues demonstrated that people found polymorphous concepts harder to acquire than either conjunctive (e.g. "large AND square") or disjunctive (e.g. "black OR triangular") concepts (Dennis, Hampton, & Lea, 1973).

In a similar vein, Shepard, Hovland, and Jenkins (1961) had previously demonstrated that three-dimension polymorphous concepts (a.k.a. Type IV problems) were harder to acquire than single dimension concepts (Type I problems, e.g. "category A is square"), and two dimension exclusive-or concepts (Type II problems, e.g. "white squares OR black triangles"). The former of Shepard's results is easily replicated (e.g. Lewandowsky, 2011; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Rehder & Hoffman, 2005); the latter is perhaps more elusive (Kurtz, Stanton, Romero, & Morris, 2013). Polymorphous concept formation has also been studied with five dimension stimuli in both people and pigeons. People took on average 440 trials to reach 85% accuracy on these problems (Wills, Noury, Moberly, & Newport, 2006); pigeons never came under control of all five features (Lea, Lohmann, & Ryan, 1993).

In summary, polymorphous concepts are hard to learn, for both humans and pigeons. This is perhaps somewhat surprising because, as Dennis et al. (1973) pointed out, many

Andy J. Wills · Gareth Croft
School of Psychology, UK
University of Plymouth
E-mail: andy@willslab.co.uk

Lyn Ellett
Department of Psychology,
Royal Holloway University of London, UK

Fraser Milton
School of Psychology,
University of Exeter, UK

Tom Beesley
Department of Psychology,
Lancaster University, UK

everyday concepts seem to be polymorphous in nature. For example, Wittgenstein (1958) argued that many concepts have a polymorphous structure, and he described natural categories as being characterized by a set of "family resemblances". In psychology, a similar point was made and evidenced by Rosch (e.g. Rosch & Mervis, 1975). It is perhaps odd that the human, or pigeon, brain, which has presumably adapted to learn concepts of the form occurring in their environment, should find polymorphous concepts so hard to acquire.

The puzzle appeared to deepen in the 1980s when evidence emerged that people switch from classification by dimensional rules, to classification by overall similarity, when time or cognitive resources are scarce (e.g. Kemler Nelson, 1984; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983). A polymorphous classification is a form of overall similarity classification, so it seemed striking that classification under time pressure, concurrent load, or incidental conditions, should take a form that people apparently found so difficult to master when not under those constraints. These results, among others, led to a number of theorists proposing dual-process accounts of category learning, in which an implicit system learns by overall similarity and an explicit system attempts to extract simple rules (e.g. Ashby, Alfonso-Reese, Turken, & Waldron, 1998).

By the 21st century, this dual-process view of category learning became sufficiently popular that for some it was viewed more as an established fact than as a theory (Ashby & Maddox, 2011). Against this context, our lab reported a number of cases where overall similarity classification was *more* effortful than single dimension classification. Time pressure and concurrent load reduced overall similarity classification, and increased single dimension classification (Milton, Longmore, & Wills, 2008; Wills, Milton, Longmore, Hester, & Robinson, 2013). Instructions to respond meticulously increased overall similarity classification, and decreased single dimension classification (Wills et al., 2013). Those who employed overall similarity classification had more frontal lobe activation and larger working memory capacities than those employing single dimension classification (Milton, Wills, & Hodgson, 2009; Wills et al., 2013).

One interpretation of our results is that the relationship between effort and overall similarity classification depends on details of the experimental procedure. In other words, there were differences in our procedures, relative to the earlier work, and these differences led to us failing to observe low effort overall similarity classification. An alternative interpretation, which we favor, is that results appearing to show overall similarity classification is a low effort "fall back" mechanism arise from methodological or analytic confounds. For example, the conclusions of the Kemler Nelson, Smith, and Ward procedures cited above can be shown to be artefacts of their analysis technique (Wills, Inkster, & Milton, 2015). A range of other results appearing to support overall similarity classification as a low effort classification mechanism (Filoteo, Lauritzen, & Maddox, 2010; Nomura et al., 2007; Smith et al., 2014; Spiering & Ashby, 2008; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006), also turn out to be flawed (Carpenter, Wills, Benattayallah, & Milton, 2016; Edmunds, Milton, & Wills, 2018; Edmunds, Wills, & Milton, 2019; Le Pelley, Newell, & Nosofsky, 2019; Milton & Pothos, 2011; Newell, Dunn, & Kalish, 2010; Newell, Moore, Wills, & Milton, 2013; Tharp & Pickering, 2009; Wills et al., 2019). In summary, the existing evidence is largely compatible with the idea that overall similarity classification is more effortful than single dimension classification.

We have previously proposed the dimensional summation theory (Milton & Wills, 2004) as an account of why overall similarity classification is so effortful. In brief, the theory says that when participants classify by overall similarity, they engage in an explicit, serial, counting process. They count up the number of dimensions in the stimulus that are characteristic of each of the candidate categories, and pick the category with the highest total. For example, imagine [black, square, large] are characteristic of category A, while [white, triangle, small] are characteristic of category B. The participant is presented with a small black triangle. They note that one dimension (color) is characteristic of category A, while two dimensions (shape, size) are characteristic of category B. They therefore conclude that the stimulus belongs to category B. This dimensional summation account explains why overall similarity classification is more effortful and time consuming than single-dimension classification—overall similarity classification is in effect the summation of several single dimension classifications.

Dimensional summation theory also predicts that accurate polymorphous classification should be difficult. In part, this is because polymorphous concepts are a type of overall similarity category structure, and hence require summation across multiple dimensions. Dimensional summation theory predicts that polymorphous concepts should be particularly difficult to acquire because it is hard to determine which stimulus features are characteristic of which categories. For example, in a five dimension, two category polymorphous classification problem, such as the one depicted in Figure 1 and Table 1, any given feature occurs 11 times in one category and 5 times in the other category. It will therefore take considerable exposure to the category structure to reliably determine which category each feature is more characteristic of. This information is required by a dimensional summation strategy to classify polymorphous concepts accurately.

Dimensional summation theory's account of the difficulty of polymorphous classification leads to a prediction — polymorphous classification should become easier if one

Category A						Category B				
lines	*color*	*shape*	*trapezium*	*flankers*		*lines*	*color*	*shape*	*trapezium*	*flankers*
horizontal	yellow	stars	long-base	fine		vertical	blue	blobs	long-top	coarse
horizontal	yellow	stars	long-base	coarse		vertical	blue	blobs	long-top	fine
horizontal	yellow	stars	long-top	fine		vertical	blue	blobs	long-base	coarse
horizontal	yellow	blobs	long-base	fine		vertical	blue	stars	long-top	coarse
horizontal	blue	stars	long-base	fine		vertical	yellow	blobs	long-top	coarse
vertical	yellow	stars	long-base	fine		horizontal	blue	blobs	long-top	coarse
horizontal	yellow	stars	long-top	coarse		vertical	blue	blobs	long-base	fine
horizontal	yellow	blobs	long-base	coarse		vertical	blue	stars	long-top	fine
horizontal	blue	stars	long-base	coarse		vertical	yellow	blobs	long-top	fine
vertical	yellow	stars	long-base	coarse		horizontal	blue	blobs	long-top	fine
horizontal	yellow	blobs	long-top	fine		vertical	blue	stars	long-base	coarse
horizontal	blue	stars	long-top	fine		vertical	yellow	blobs	long-base	coarse
vertical	yellow	stars	long-top	fine		horizontal	blue	blobs	long-base	coarse
horizontal	blue	blobs	long-base	fine		vertical	yellow	stars	long-top	coarse
vertical	yellow	blobs	long-base	fine		horizontal	blue	stars	long-top	coarse
vertical	blue	stars	long-base	fine		horizontal	yellow	blobs	long-top	coarse

Table 1 The five dimension two category polymorphous classification employed in the current experiments. Stimulus dimensions are as shown in Figure 1. The most typical stimuli are shown above the first dotted line. Stimuli that differ by one feature from these are shown between the first and second dotted line. The remainder of the stimuli, below the second dotted line, differ from the most typical examples of the category by two features.

first receives deterministic training on each of the stimulus dimensions. For example, one is first trained that horizontal bars indicate category A, and vertical bars indicate category B. Having mastered this simple discrimination, one is then trained on each of the other four dimensions in turn, before commencing the polymorphous classification problem. A participant who is pretrained in this way is then in a position to immediately apply a dimensional summation strategy, using the knowledge they have already acquired about each of the constituent features.

This prediction receives informal support from the results of Wills et al. (2009) where, with this kind of single feature pretraining, around two thirds of participants (both pigeon and human) successfully classified a polymorphous category structure on the basis of overall similarity (see also Lea et al., 2018). However, these previously reported experiments had no control group against which to compare the pretrained group. In the current set of experiments, control participants received an equivalent amount of training on the polymorphous concept itself.

Outline of the paper

In Experiment 1 we demonstrate that single feature pretraining increases accuracy on a subsequent five dimension polymorphous classification, relative to an equivalent amount of training on the polymorphous classification itself. It also increases reaction time. Experiment 2 shows, through a partial reversal procedure, that the pretraining advantage is specific

to the feature-category information learned in pretraining, and not to some more general strategic or motivational factor. In Experiments 3 and 4, we demonstrate that, as predicted by dimensional summation theory, it is the deterministic structure of the single feature pretraining that leads to the advantage, and not a range of other coincidental differences. Finally, in Experiment 5, we test a prediction of dimensional summation theory that the single feature pretraining advantage should be reduced or eliminated by a concurrent counting task — a task that should interfere with the counting operations assumed to underlie high-accuracy polymorphous classification. This prediction is confirmed.

Experiment 1

In Experiment 1, we compared single feature pretraining on a five dimension polymorphous concept to an equivalent amount of training on the polymorphous problem itself. We employed a between subjects manipulation, with a yoking procedure to match overall training length across randomly-selected pairs of participants.

Method

Participants and apparatus

Twenty-four people participated in this experiment, randomly allocated to two between subject conditions, with the constraint that each condition had twelve participants. The sam-
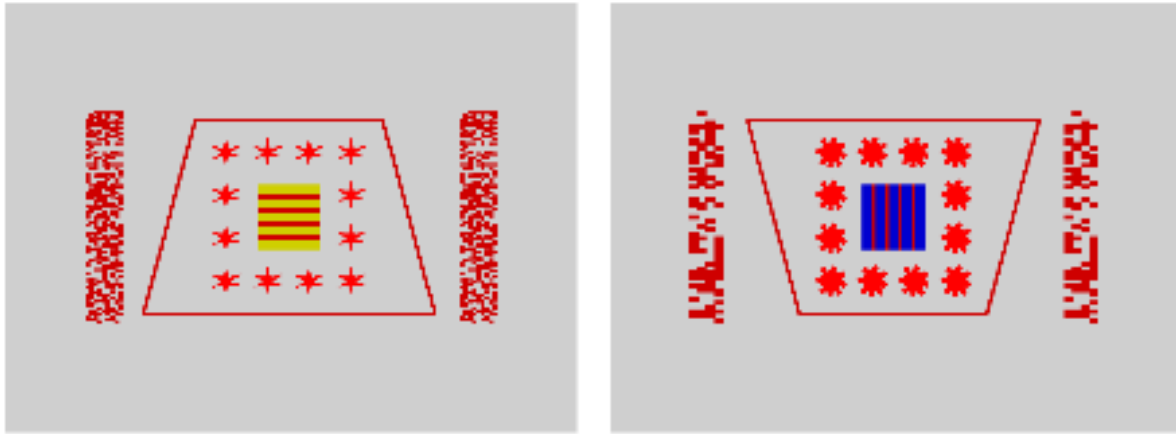
Fig. 1 Examples of the stimuli employed in the current experiments. The stimulus on the left is the most typical member of category A; on the right is the most typical member of category B. Image credit: Andy J. Wills, CC-BY-SA 4.0. https://osf.io/4g6az/

ple size was decided before data collection on the basis that is was sufficient to detect large between-subject effects ($d = 1.2$ at 80% power). All participants in the current paper were undergraduate students from the University of Exeter. In all experiments, the stimuli were presented on a 17-inch monitor, placed approximately 1 meter from the participant at eye level. Responses were collected via a standard keyboard.

Stimuli

Two types of stimuli were employed in the current experiment: five feature stimuli and single feature stimuli. Figure 1 shows two five feature stimuli, and Table 1 shows the category structure. Each five feature stimulus had five binary stimulus dimensions: (1) orientation of the center stripes (horizontal or vertical), (2) background color of the center stripes (yellow or blue), (3) icon shape ("stars" or "blobs"), (4) trapezium shape (long-base or long-top), and (5) flanker texture (fine or coarse). These stimuli were similar to those used in related experiments in pigeons (Lea, Wills, & Ryan, 2006). Each single feature stimulus comprised just one feature, selected from the five feature stimuli (e.g. a long-base trapezium). All stimuli were presented on a mid-gray background, and the five feature stimuli were approximately 14×7 degrees of visual angle in size, excluding that background. The size and location of features in the single feature stimuli were the same as the corresponding features in the five feature stimuli, with the remaining four features absent. In the single feature presentation of the center-stripes' background color, the color was depicted as a solid square of the same dimensions as the center stripes.

Ten single feature stimuli, and thirty-two five feature stimuli (see Table 1), are possible within this stimulus set. Five of the single feature stimuli were assigned to category A (horizontal, yellow, stars, long-base, fine); the remain-

ing five single feature stimuli were assigned to category B (vertical, blue, blobs, long-top, coarse). The category membership of the five feature stimuli was determined by the number of category A and category B features presented — where the number of category A features exceeded the number of category B features, the item was assigned to category A, otherwise it was assigned to category B. For example the stimulus [`horizontal, blue, stars, long top, fine`] was assigned to category A because it has three category A features [`horizontal, stars, fine`] and two category B features [`blue, long top`].

Procedure

Participants were randomly paired for the purposes of yoked training. In each yoked pair of participants, one was randomly assigned to the *single feature* condition, and the other to the *control* condition. The single feature participant was trained on the single feature stimuli to an errorless criterion, and then trained on the five feature stimuli. The yoked control participant received the same total number of training trials as the single feature participant, but trained on the five feature stimuli from the outset.

Each training trial began with the presentation of the to-be-categorized stimulus, which remained on the screen until the participant classified it as either a member of category A (by pressing the "X" key) or a member of category B (by pressing the ">" key). A feedback message immediately followed, informing the participant whether they were correct or incorrect, and also giving the correct category membership (e.g. "Correct. It was category A" or "Wrong! It was category B"). The next trial began 2000 ms after the onset of the feedback message. In every block, each presented stimulus occurred with equal frequency and the order of presentation was randomized. Participants were trained in blocks

of 32 trials, with the opportunity to rest for a few seconds at the end of each block. The participant's percent correct score for the current block was presented to them at the end of the block, along with a statement that the target accuracy was 100%.

Each single feature participant trained on one stimulus dimension at a time, reaching a criterion of one errorless block on that stimulus dimension before being trained on the next. The stimulus dimensions were trained in a different random order for each participant. After reaching criterion on the fifth stimulus dimension, single feature participants were given a 30-item test on all the single feature stimuli (each feature was presented three times, in a random order). The procedure for these test trials was identical to the training procedure above, except that no feedback was given. One or more errors on this test led to the single feature training phase being restarted. Once the single feature participant had achieved errorless performance on each of the dimensions and on the 30-item test, they were trained on the five feature stimuli for 8 blocks (receiving feedback, as before). The yoked control participant received the same number of blocks of training on the five feature stimuli as the single feature participant had received on the single feature stimuli, and then continued to train on the five feature stimuli for a further 8 blocks.

Results and discussion

The raw data for this experiment are archived along with the analysis scripts at `https://osf.io/4g6az/`. All data analysis was conducted with R (R Core Team, 2019), using principal packages *dplyr* (Wickham, Francois, Henry, & Muller, 2019), *effsize* (Torchiano, 2019), *ggplot2* (Wickham, 2016), and *pwr* (Champely, 2018). Bayesian analysis was conducted following the procedure described by Dienes (2011), with priors as described below. Following Jeffreys (1961), Bayes Factors exceeding three were considered as evidence for the experimental hypothesis, while Bayes Factors smaller than one-third were considered as evidence for the null.

In this first experiment, the absence of any closely related previous study led us to adopt quite broad priors for the effect of single feature pretraining on accuracy and on reaction time. For accuracy differences (i.e. probability of a correct response after single feature pretraining minus probability of a correct response in the control condition), we assumed a uniform prior ranging for -.5 to .5. Chance performance is .5 on this task, and perfect performance is 1. It thus seemed unlikely that the difference between conditions would fall outside this range. Reaction time is not bounded in this way, but in practice it is very rare for conditions in lab-based category learning studies to differ in mean reaction time by more than 5 seconds. We thus adopted a uni-

form prior from -5 to +5 seconds for our reaction time analysis.

Two participants in the single feature condition did not complete the experiment in the time available. They were excluded on this basis, along with their yoked participants in the control condition. The remaining participants took a mean of 9.60 blocks (range: 7–16, SD = 2.46) to complete phase one. Median reaction times across phase one were 0.78 s (IQR = 0.16) in the single feature condition, and 2.49 s (IQR = 1.64) in the control condition.

Figure 2 shows the results of principal interest, which are the phase two accuracy and reaction times. As predicted, single feature pretraining substantially increased accuracy on the polymorphmous classification, relative to an equivalent amount of training on the polymorphous problem itself. Both the unstandardized and standardised effect sizes were large; there was an increase in accuracy of about .18 on average, and Cohen's *d* (Cohen, 1992) was 1.46. There was strong Bayesian evidence for the experimental hypothesis, *BF* = 29.9. Figure 2B shows that this increase in accuracy was accompanied by a substantial increase in reaction time, *mean* = 1.79 seconds, *d* = 1.12, *BF* = 4.03.

In summary, single feature pretraining increased accuracy on a polymorphous classification with a large effect size. This increase in accuracy was accompanied by an increase in reaction time. This pattern of results is consistent with the idea that single feature pretraining encourages a dimensional summation strategy, which takes substantial time to complete. However, it's also compatible with a more general strategic or motivational account, that the pretraining causes a general slow down of responding, and this results in better performance at test. Under this latter account, the specific content of the pretraining phase should not matter too much. Under the former account, it's crucial that the information gained in training is compatible with the test phase.

In Experiment 2, we compared these two accounts by introducing a partial reversal between the two phases, where the valence of three dimensions, but not the other two, was reversed. For example, in pretraining, horizontal lines indicated category A and vertical lines indicated category B. In the polymorphous classification phase this might be reversed, so that horizontal lines were now characteristic of category B and vertical lines characteristic of category A. Two further dimensions (e.g. flankers and trapezium) would be reversed, while the other two dimensions maintained their pretraining valence. Under a dimensional summation account, reaction time in this partial reversal condition should remain high, as participants attempted to employ a dimensional summation strategy, but the accuracy benefit observed under regular single feature pretraining should be reduced or eliminated. Partial reversal is a superior design to full reversal, as humans are readily able to detect a full reversal of polymorphous concepts and adapt by reversing the cate-
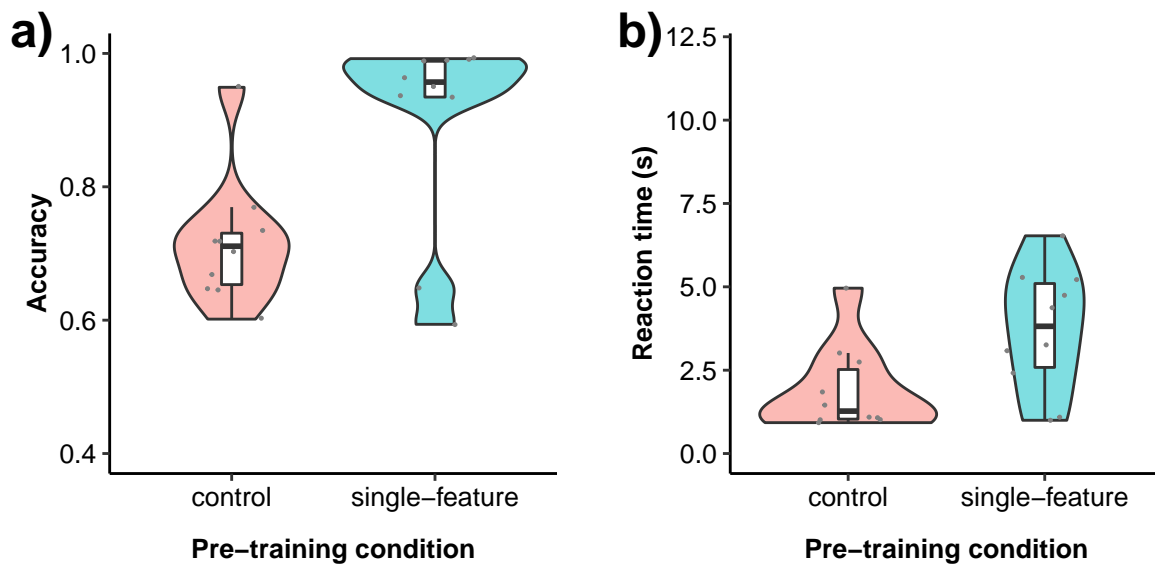
a)

b)

Fig. 2 Accuracy and reaction time in phase two of Experiment 1, as a function of pretraining condition. Distributional information is shown as a boxplot, a violin plot, and individual data points. The box plot shows median performance and interquartile range. The violin plot is a density plot, rotated through ninety degrees, and mirror copied to produce the symmetrical pattern shown; see Hintze and Nelson (1998) for details. The small gray plot symbols are scores for individuals. Image credit: Andy J. Wills, CC-BY-SA 4.0. https://osf.io/4g6az/

gory labels, leaving the underlying feature-category knowledge intact (Kruschke, 1996; Wills et al., 2006). This strategy does not work when only some of the dimensions have been reversed.

Experiment 2

In Experiment 2, we conducted a between subjects comparison of (a) single feature pretraining, (b) partially-reversed single feature pretraining, and (c) pretraining on the polymorphous problem itself.

Method

Participants

Sixty people participated in the current experiment, randomly allocated across three between subject conditions, with the constraint that each condition had twenty participants. The sample size for this and all subsequent experiments in this manuscript was determined after analysis of Experiment 1, and prior to data collection for Experiments 2–5. It was chosen to provide good statistical power (greater than 90%) to detect effects of the size seen in Experiment 1.

Procedure

The procedure was identical to Experiment 1, apart from four changes. First, a partial reversal condition was added.

Participants in this condition received the same pretraining as participants in the single feature condition. However, at the end of phase one, and unknown to the participants, the valence of three of the dimensions was reversed. For example, horizontal stripes might become characteristic of category B, with vertical stripes now characteristic of category A (the opposite to phase one). For each participant, three of the five stimulus dimensions, randomly selected, were reversed.

Second, phase one now involved a fixed amount of training, rather than training to criterion. All participants received twenty blocks of training in phase one. This change was made to eliminate the possibility that the yoked training procedure of Experiment 1 might have over-estimated the size of the single feature pretraining effect, due to the Church effect (Church, 1964). For those unfamiliar with the Church effect, a brief summary is provided in the Appendix. In the single feature and partial reversal conditions, participants now completed four blocks of training on each of the five stimulus dimensions before moving on to the next (as in Experiment 1, the order in which the dimensions were trained was randomized). In the control condition, participants received twenty blocks of training on the five feature stimuli.

The third change to the procedure, relative to Experiment 1, was that phase two was shortened from eight blocks, to four blocks. This was for practical reasons; shortening phase two reduced the length of the experiment, and hence allowed us to test participants more efficiently. Analysis of Experiment 1 indicated that the first four blocks were sufficient to detect the effect of pretraining.

The fourth and final change was that the single feature tests right at the end of single feature pretraining in Experiment 1 did not appear in this or any subsequent experiment. We considered such tests to be superfluous, given that training in phase one was now fixed in length, rather than to criterion. Participants moved directly from the end of single feature pretraining to the beginning of phase two without any intervening tests.

Results and discussion

The raw data for this experiment are archived along with the analysis scripts at https://osf.io/8nyfw/. Following Dienes (2011), Bayesian priors for the effects of single feature pretraining on accuracy and reaction time were based on the results of Experiment 1. Specifically, a normally distributed prior was used, with a mean of the observed effect in Experiment 1, and a standard deviation of half that. Due to the lack of previous data on the effects of a partial reversal on single feature pretraining, we employed the same broad priors described in Experiment 1 for comparisons involving this condition (i.e.-.5 to +.5 for accuracy, -5 to +5 for reaction time).

Fixed-length training in phase one of the single feature and partial reversal conditions elicited accurate, rapid responding from participants; for both conditions mean accuracy was 98% (SD = .01), and median reaction time was around 0.67 s (IQR = 0.23). Unsurprisingly, phase one performance for participants in the control condition was worse (mean accuracy = 0.65, SD = 0.05) and slower (median RT = 1.33, IQR = 0.46)

Figure 3 shows the results of phase two. As observed in Experiment 1, single feature pretraining substantially increased accuracy on the polymorphous classification, relative to an equivalent amount of training on the polymorphous problem itself. Both the unstandardized and standardised effect sizes were large, albeit not quite as large as in Experiment 1 (perhaps indicating that our concerns about the Church effect were well founded). The mean increase in accuracy was 0.12, $d = 1.01$, $BF = 45.4$. In contrast, the partial reversal pretraining condition *reduced* accuracy on the polymorphous classification, relative to the control condition. The mean decrease was 0.19, $d = 2.2$, $BF = 9 \times 10^8$. This pattern of results is not consistent with a general strategic or motivational account of the single feature pretraining effect. It is, however, predicted by the dimensional summation hypothesis.

Figure 3B shows the reaction times in phase two. As in Experiment 1, the increase in accuracy produced by single feature pretraining was accompanied by an increase in reaction time, mean increase = 2.04 s, $d = 1.96$, $BF = 4 \times 10^7$. The partial reversal condition also induced an increase in reaction time relative to the control condition, mean increase = 1.44 s, $d = 0.79$, $BF = 3.02$, which suggests that participants were still attempting to apply the information from phase one. The reaction time results are predicted by the dimensional summation hypothesis

Experiment 3

Having established the presence of a single feature pretraining (SFPT) effect in polymorphous classification, we then set about trying to understand why it occurs. One possibility, already mentioned, is that SFPT helps polymorphous classification primarily because it trains the feature dimensions deterministically. This provides, for the participant, clear knowledge of the feature-category relationships upon which a dimensional summation strategy can be enacted. However, there are a number of other possible explanations, two of which we examined in Experiment 3.

The first possibility is that SFPT allows participants to focus on one stimulus dimension at a time, and it is this seriality of the learning of the feature-category relationships which leads to the SFPT advantage, rather than the deterministic presentation of these relationships. One prediction of this *seriality* hypothesis is that single feature pretraining should also increase polymorphous classification accuracy if dimensions are trained sequentially, but with the feature $\rightarrow$ category associations presented in the same probabilistic manner as they occur in the polymorphous concept itself. For example, in one block of single feature training, horizontal stripes would be followed by category A on 11 occasions and category B on 5 occasions. In contrast to the seriality hypothesis, a dimensional summation hypothesis predicts that probabilistic single feature pretraining should be less effective than deterministic training. This is because probabilistic training gives a less clear (albeit more accurate) indication of the relationship between features and category labels in a polymorphous concept.

Another prediction of the *seriality* hypothesis is that the effectiveness of single feature pretraining should be reduced if dimensional training is intermixed. So, for example, if all ten features occur on different, randomly ordered, trials in the same block then, under a seriality hypothesis, participants must either attempt to learn several feature-category relationships concurrently, or ignore all trials but the ones for the features they have chosen to focus on. Either way, intermixed single feature pretraining should be less effective than serial single feature pretraining. Under the dimensional summation hypothesis, intermixed deterministic training can be approximately as effective as sequential deterministic training, because it is the deterministic nature of training, not its precise sequencing, that is crucial.

In Experiment 3, we added two further single feature pretraining conditions. In one condition, single feature pretraining is intermixed but deterministic. In the other, single
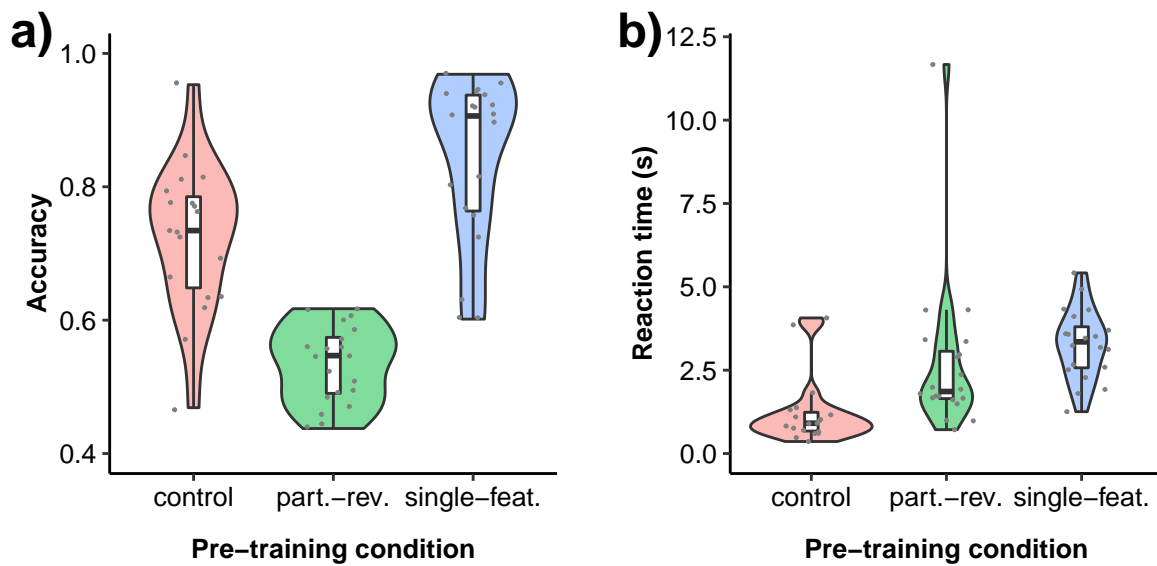
Fig. 3 Accuracy and reaction time in phase two of Experiment 2, as a function of pretraining condition. Image credit: Andy J. Wills, CC-BY-SA 4.0. https://osf.io/8nyfw/

feature pretraining is sequentially-ordered, but probabilistic. The addition of these two conditions allows a comparison of the dimensional summation and seriality explanations of the single feature pretraining effect. The seriality hypothesis was generated prior to data collection in Experiments 2–5, and came from a simple enumeration of the procedural differences between the single-feature pre-training and control conditions. In other words, single feature pretraining is not only deterministic, it is also serial, and thus it is a logical possibility that it is the seriality of single feature pretraining, rather than its deterministic nature, that drives the effect. Thus, the seriality hypothesis was not inspired by any particular formal theory of classification learning, simply by the several procedural differences between the phase one conditions of Experiment 1.

We also examined further a priori, procedurally-inspired account of the single feature pretraining effect in Experiment 3. Perhaps, by presenting features singly during pretraining, the dimensional structure of the polymorphous stimuli becomes more obvious in phase two, and it is this *decomposition* of the stimuli during pretraining that primarily leads to the single feature pretraining advantage. If this is the case, it should be possible to produce similar levels of accuracy on the phase two polymorphous classification by making the dimensional structure more obvious during phase one in some other way. In Experiment 3, we did this through *constructed polymorphous* pretraining. In this procedure, the polymorphous stimuli in phase one were constructed on screen on each trial, with each dimension being added every few hundred milliseconds until the stimulus was complete.

Method

Participants

One hundred people participated in the current experiment, randomly allocated across five between subject conditions, with the constraint that each condition had twenty participants. The sample size was set prior to data collection (see Experiment 2 for details).

Procedure

The procedure of the single feature and control conditions was identical to Experiment 2. Three further between subject conditions were added. All conditions received the standard polymorphous classification task in phase two, but differed in the nature of the phase one pretraining. All conditions contained the same number of trials in each phase. The three additional conditions were as follows:

Single feature probabilistic pretraining Training in this condition was identical to the single feature pretraining of Experiment 2, except that the feedback was probabilistic rather than deterministic. In each block, each feature was presented 16 times. On 11 of those occasions, the feedback was given consistent with the mapping described in the Experiment 1 procedure (e.g. vertical lines → category A). On the remaining 5 occasions, the opposite feedback was given (e.g. vertical lines → category B). This probabilistic relationship between a single cue and the category label is identical to that observed in the polymorphous classification task. The end-of-block message still reported percent correct, but partici-

pants were told the target accuracy was "over 65%" (rather than 100%, as in the other conditions). The maximum sustained accuracy achievable in this pretraining task is 68.75%.

Single feature intermixed pretraining Pretraining in this condition was identical to the single feature pretraining of Experiment 2, except that in each block, all ten features were presented, in a random order. In each 32-trial block, eight features were presented six times, and two features were presented eight times. The two features to be presented with slightly higher frequency were randomly selected on each block, with the constraint that both features came from the same dimension (e.g. bar orientation). This randomization procedure allowed us to match the block and phase length of the other conditions exactly, at the cost of slight random variations in presentation frequency of each feature across blocks and participants.

Constructed polymorphous pretraining Pretraining in this condition was identical to polymorphous pretraining, except that, on each trial, the stimulus was constructed on screen over time. Each stimulus presentation began with the central lines feature (horizontal or vertical stripes). After 300 ms, the background color was added (blue/yellow), and after subsequent delays of 300 ms each, the other features were added in the order: icons (stars/blobs), trapezium, flankers. Responses made before the stimulus was complete were ignored.

Results and Discussion

The raw data for this experiment are archived along with the analysis scripts at `https://osf.io/vu8ms/`. The data for two participants were lost due to technical errors. Bayesian priors for the effects of single feature pretraining on accuracy and reaction time were based on the results of Experiment 2, in the manner described in Experiment 2. Due to the lack of previous data on the other hypothesized effects, we employed the same broad priors described in Experiments 1 and 2 for the remaining comparisons (i.e.-.5 to +.5 for accuracy, -5 to +5 for reaction time).

Deterministic single feature pretraining elicited high accuracy, both with sequential ordering (mean accuracy = .98, SD = .01), and intermixed ordering (mean accuracy = .92, SD = .10). Probabilistic single feature pretraining led to lower performance (mean accuracy = 0.56, SD = .05); this is to be expected given the probabilistic nature of the feedback. Accuracy in the probabilistic single feature condition was higher than could be achieved by random responding, $BF = 1.6 \times 10^7$ (calculated against a uniform prior of accuracy being somewhere between 0.5 and 1). Mean accuracy in the polymorphous pretraining condition was 0.62 (SD = .05), comparable to the results of Experiment 2. Similar accuracy

was observed under constructed polymorphous pretraining (mean accuracy = 0.64, SD = .07).

Reaction times in phase one were also as expected on the basis of Experiment 2, with a median of 0.68 s (IQR = .13) and 1.23 s (IQR = 0.62) in the single feature and polymorphous control conditions, respectively. Reaction times for the single feature intermixed and single feature probabilistic conditions were in the same range as the two previously-mentioned conditions, with median times of 1.03 s (IQR = 0.18) and 0.95 s (IQR = 0.37) respectively. Reaction times were substantially longer in the constructed polymorphous condition (median = 2.71 s, IQR = 0.64), but this is to be expected as the stimulus took time to construct on screen, and responses prior to the completion of the construction were ignored. Reaction times as measured from the completion of the stimulus were similar to the polymorphous control condition (median = 1.21 s, IQR = .64).

Turning to phase two, Figure 4 shows the key results of the current experiment:

Deterministic single feature pretraining As previously observed in Experiments 1 and 2, single feature pretraining substantially increased accuracy and reaction time on the polymorphous classification, relative to an equivalent amount of training on the polymorphous problem itself. Mean accuracy increased by 0.18, $d = 1.67$, $BF = 2.4 \times 10^5$, and mean reaction time increased by 1.84 s, $d = 1.84$, $BF = 6.8 \times 10^{10}$.

Probabilistic single feature pretraining A single feature pretraining advantage was not observed when the feature $\rightarrow$ category associations in pretraining were probabilistic. Accuracy was slightly lower (mean = -.04) after probabilistic pretraining, than after polymorphous pretraining, $d = .44$, but with Bayesian evidence for the null, $BF = 0.18$. Reaction time was slightly higher, mean = 0.10 s, $d = .20$, again with Bayesian evidence for the null, $BF = 0.05$.

Given these and the deterministic single feature pretraining results, one would expect accuracy and reaction time to be lower after probabilistic pretraining than after deterministic pretraining. Comparison of the deterministic and probabilistic single feature conditions confirms this expectation; accuracy is lower, mean = .22, $d = 1.79$, $BF = 6.3 \times 10^5$, and reaction time is higher, mean = 1.74 s, $d = 1.98$, $BF = 1.5 \times 10^7$. The results concerning the probabilistic single feature pretraining condition are predicted by the dimensional summation hypothesis, but not the seriality hypothesis.

Intermixed single feature pretraining A single feature pretraining advantage was observed when the single feature pretraining was deterministic, but with the features presented in an intermixed order. Accuracy was higher after single feature intermixed pretraining than polymorphous pretraining,
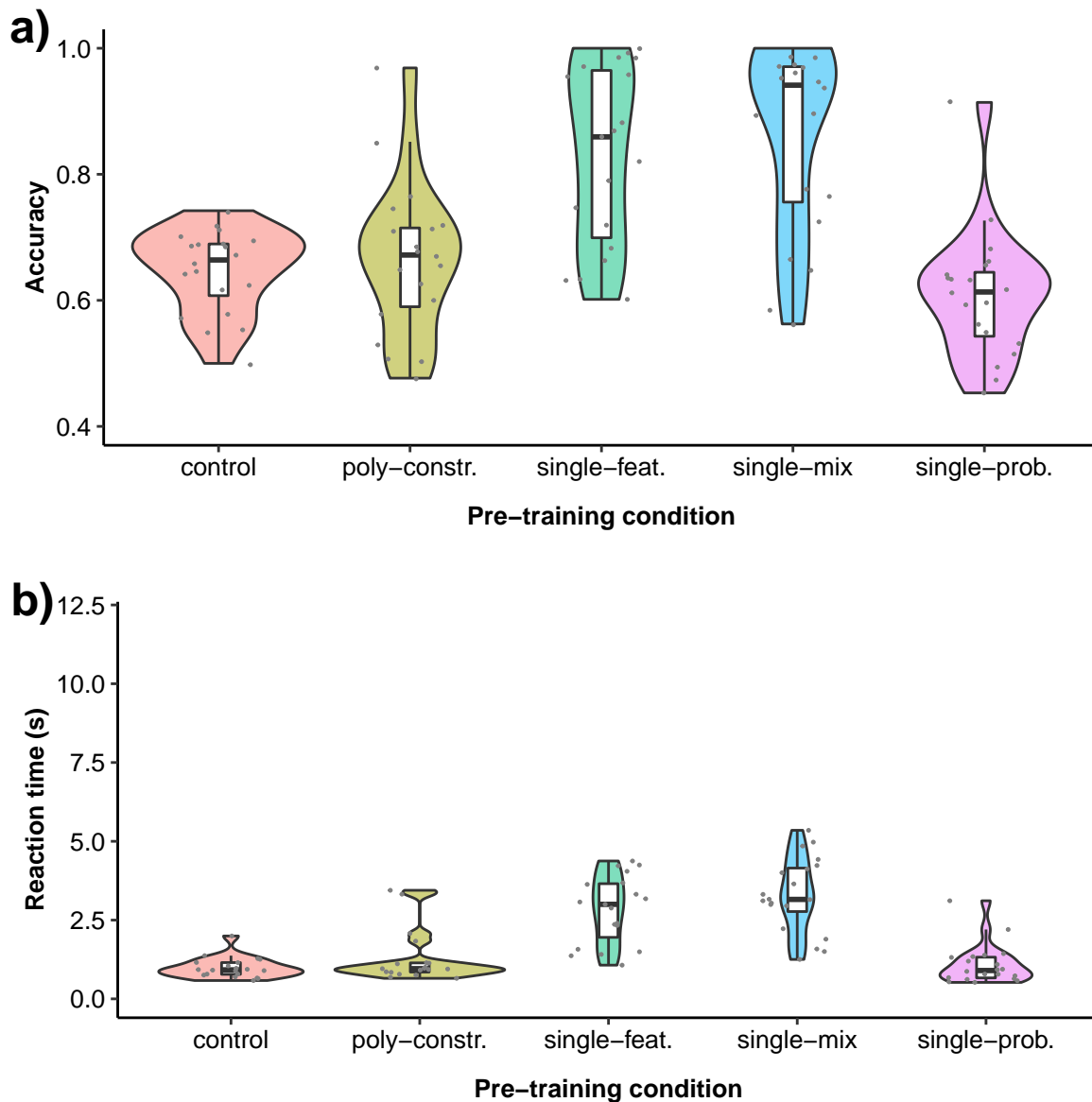
Fig. 4 Accuracy and reaction time in phase two of Experiment 3, as a function of pretraining condition. Image credit: Andy J. Wills. CC-BY-SA 4.0. https://osf.io/vu8ms/

mean = .21, $d = 1.85$, $BF = 2.4 \times 10^6$, as was reaction time, mean = 2.30 s, $d = 2.65$, $BF = 1.2 \times 10^{14}$. Sequential and intermixed single feature pretraining produced comparable levels of accuracy on subsequent polymorphous classification, mean accuracy difference = .03, $d = 0.21$, $BF = 0.15$, with comparable reaction times, mean reaction time difference = 0.47 s, $d = 0.41$, $BF = 0.21$. These results seem contrary to the seriality hypothesis, but can be accommodated by the dimensional summation hypothesis.

Constructed polymorphous pretraining Constructing the polymorphous stimulus on the screen one feature at a time did not improve accuracy on a subsequent standard polymorph-

ous classification, relative to an equivalent amount of training on that standard classification, mean = .02, $d = .19$, $BF = 0.09$; nor did it affect reaction time, mean = .30 s, $d = .48$, $BF = .15$. These results provide no support for the *decomposition* hypothesis. They are compatible with the dimensional summation hypothesis, because the relationships between features and category labels are probabilistic, not deterministic, in the constructed polymorphous pretraining condition.

Summary The results of Experiment 3 provide further support for the dimensional summation explanation of the SFPT

effect, but not for two alternative explanations of this effect (the seriality, and decomposition, hypotheses).

Experiment 4

In Experiment 4, we considered another alternative explanation of the SFPT effect — the effect of errors during pretraining. A number of authors have previously argued that learning can be enhanced through the avoidance of errors during training (Baddeley & Wilson, 1994; Terrace, 1963). Although the issue of whether making errors is beneficial or detrimental to learning remains controversial (Ashby, Maddox, & Bohil, 2002; Edmunds, Milton, & Wills, 2015; Potts & Shanks, 2014; Seabrooke, Hollins, Kent, Wills, & Mitchell, 2019), it is undeniably the case that, in the current experiments, single feature pretraining is not well matched to the polymorphous control condition in terms of the number of classification errors participants produce. It is therefore possible that the SFPT advantage comes from this difference in error frequency, rather than from the provision of deterministic feature-category associations assumed to be the cause by the dimensional summation hypothesis.

In Experiment 4, we tested this possibility by introducing two further between subject conditions to the basic single feature pretraining design. In these additional conditions, classification errors were essentially eliminated during phase one by presenting the correct category label a few hundred milliseconds after the stimulus was presented. This procedure is sometimes referred to as observational training (cf. Ashby et al., 2002; Edmunds et al., 2015; Wills & McLaren, 1997), as opposed to the more common feedback training used throughout Experiments 1–3. Phase two of Experiment 4 employed standard feedback training in all conditions, for comparability with these previous experiments. If the SFPT advantage is primarily due to the lower rate of participant errors in single feature pretraining, compared to the polymorphous classification control condition, then the use of observational training during phase one should largely eliminate the SFPT advantage. In contrast, the dimensional summation hypothesis predicts a SFPT effect should be seen under both observational and feedback pretraining, because only in the cases of single feature pretraining is there a deterministic relationship between the features and the category membership.

Method

Participants and apparatus

Our intention was to test eighty participants, randomly allocated across four conditions, with the constraint that each condition had twenty participants. The sample size was set prior to data collection (see Experiment 2 for details). Due to an administrative error, data collection in two conditions was terminated at twenty one rather than twenty participants. Prior to the analysis of Experiments 2–5, we chose to retain rather than discard these two additional data sets, on the grounds that this minor and accidental over-collection of data was unlikely to inflate statistical error rates.

In addition to the apparatus of Experiments 1–3, stereo on ear headphones were used to present auditory stimuli to participants. The sounds were digitized and their presentation was controlled by computer.

Procedure

The procedure of the single feature and control conditions was identical to Experiments 2 and 3, with the exception that feature → category association ratings were taken at the end of each block. On each rating trial, a single feature was presented in the center of the screen, with a rating scale directly above it. At the top of the screen appeared a question asking how likely it was that category A (or B) would occur when this feature was present. Participants then had to respond on a scale from -10 (will never appear) to +10 (will always appear). Participants were asked to rate each feature presented in the preceding block for both categories. The ratings part of the experiment was exploratory in nature, with no clear predictions considered ahead of data collection. The results turned out to be largely unremarkable and, for brevity, are not discussed in the current paper. The interested reader may find the ratings data in the OSF repository cited below.

The procedure in the single feature observation and control observation conditions was the same as the corresponding conditions described above, with the following exception. In phase one, presentation of the stimulus was followed, after 200 ms, by the category label spoken over headphones ("A" or "B"). A manual categorization response was still required, and feedback still given, but with the expectation that very few errors would be made under such conditions. The procedure in phase two was identical to the single feature and control conditions described above; in other words there was no spoken label, participants had to guess, and received feedback.

Results and discussion

The raw data for this experiment are archived along with the analysis scripts at `https://osf.io/g3w8a/`. Bayesian priors were determined in the same manner as previous experiments, using mean effect sizes from Experiments 2 and 3 where appropriate, and broad priors otherwise.

As in previous experiments, single feature training in phase one elicited fast and accurate responding (mean accuracy = .98, SD = .02, median RT = 0.73 s, IQR = 0.20), while
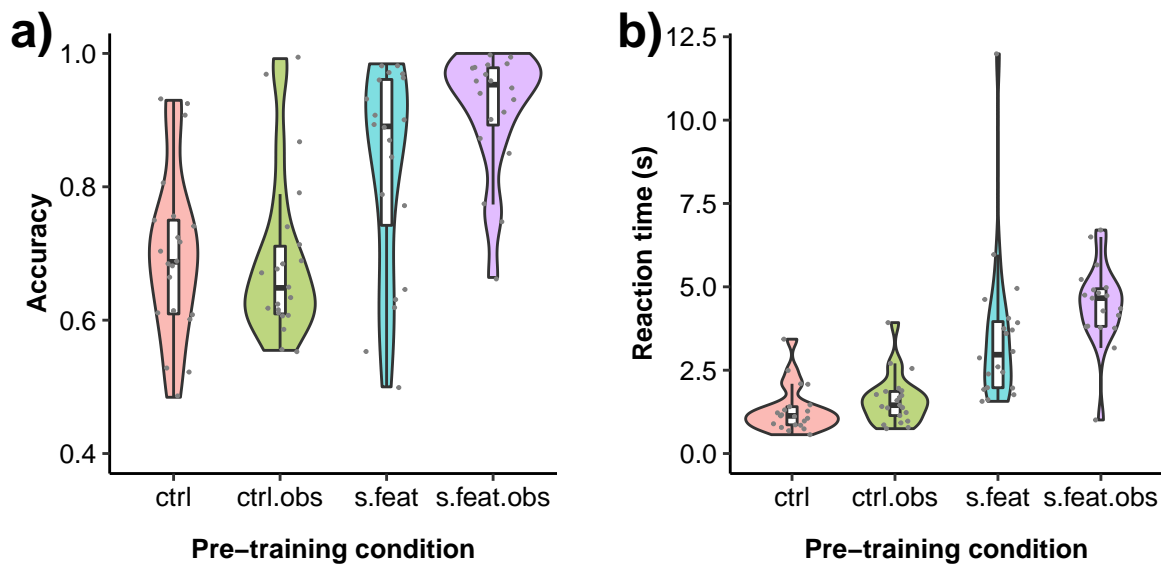
Fig. 5 Accuracy and reaction time in phase two of Experiment 4, as a function of training type and pretraining condition. Image credit: Andy J. Wills. CC-BY-SA 4.0. https://osf.io/g3w8a/

performance in the control condition was worse (mean accuracy = .67, SD = .11, median RT = 1.99 s, IQR = 1.94). Unsurprisingly, presenting the correct category label resulted in very few errors in both the single feature observation (accuracy = .98, SD = .02) and control observation (accuracy = .98, SD = .04) conditions. Reaction times for these conditions were in the same range as the other two conditions, 0.95 s (IQR = 0.35) for the single feature observation condition, and 1.77 s (IQR = .45) for the control observation condition.

Figure 5 shows the results of phase two. As previously observed in Experiments 1–3, single feature pretraining substantially increased accuracy and reaction time on the polymorphous classification, relative to an equivalent amount of training on the polymorphous problem itself. Mean accuracy increased by .13, $d = .93$, $BF = 43.0$, and mean reaction time increased by 2.22 s, $d = 1.31$, $BF = 2945$. Single feature pretraining also substantially increased accuracy and reaction time in the observation conditions, with accuracy increasing by .23, $d = 2.1$, $BF = 5.3 \times 10^8$, and reaction time increasing by 3.38 s, $d = 1.8$, $BF = 1.6 \times 10^6$.

Inspection of Figure 5 suggests the post-hoc hypothesis that observational single feature training might be more effective than standard single feature training, but the Bayesian evidence for this hypothesis is equivocal, $BF = 1.13$, as is the evidence for a difference in reaction times, $BF = 1.11$. In terms of the two control conditions, there was substantial evidence for the null hypothesis that the two conditions do not differ, both for accuracy, $BF = 0.10$, and for reaction time, $BF = 0.14$.

In summary, the results of Experiment 4 showed that the single feature pretraining advantage still occurs in the absence of any classification errors in phase one. This result is compatible with the dimensional summation hypothesis, but is problematic for the idea that the SFPT advantage is caused by the lower frequency of classification errors in SFPT, relative to a polymorphous control condition.

Experiment 5

Across Experiments 1–4, we established the presence of a single feature pretraining advantage in polymorphous category learning, and presented a range of findings that were supportive of a dimensional summation account of this effect. One further, striking, prediction of the dimensional summation account is that a relatively simple counting task, if conducted at the same time as polymorphous classification, should eliminate the SFPT advantage. This is because a counting task should interfere with the deliberate summation of characteristic features assumed to underlie accurate polymorphous classification. So, while SFPT should provide the classifier with the constituent knowledge required to subsequently classify a polymorphous category structure effectively, a concurrent counting load should stop the classifier from applying that knowledge. Such is the idea tested in Experiment 5.

Method

Participants

Eighty people participated in the current experiment, randomly allocated across four between subject conditions, with the constraint that each condition had twenty participants. The sample size was set prior to data collection (see Experiment 2 for details).

Procedure

The procedure in phase one was identical to the single feature and control conditions of Experiments 2 and 3. The procedure in phase two was also identical to these conditions, with the exception that each block was accompanied by an asynchronous stream of spoken two digit numbers. The numbers ranged from 11 to 98 and appeared in a random order. Each number had a spoken duration of approximately one second (achieved using voice synthesis), and there was a silent gap of 200 ms between the end of each number and the beginning of the next. The stream of numbers began simultaneously with the beginning of each block, and stopped at the end of each block, in phase two. Participants in the load conditions were told to keep an exact count of the number of even numbers in each block, and were asked to state their total at the end of each block. Participants in no load conditions experienced the same stimuli, but were told that we were interested in automatic processing, so they should ignore the numbers, and just enter their guess of the number of even numbers at the end of each block. All participants received feedback on the accuracy of their count/guess.

For participants in the load conditions, phase one was preceded by one block (thirty two trials) of practice on the counting task. Participants were asked to keep count of the number of even numbers, but the categorization stimuli were replaced by a large letter "A" or "B", making that part of the task very easy. Participants in the no-load conditions did not receive practice on the counting task, as this might have undermined the later instruction that they were to ignore the digits and just make a guess.

Results and Discussion

The raw data for this experiment are archived along with the analysis scripts at `https://osf.io/fdm8r/`. Bayesian priors were determined in the same manner as previous experiments, using mean effect sizes from Experiments 2–4 where appropriate, and broad priors otherwise. As in previous experiments, single feature training in phase one elicited fast and accurate responding (mean accuracy = .98, SD = .02, median RT = 0.57 s, IQR = 0.11), while performance in the control condition was worse (mean accuracy = .64, SD = .06, median RT = 1.52 s, IQR = 0.96).

Figure 6 shows the results of phase two. As previously observed in Experiments 1–4, single feature pretraining substantially increased accuracy and reaction time on the polymorphous classification, relative to an equivalent amount of training on the polymorphous problem itself. Mean accuracy increased by .11, $d = .97$, $BF = 46.3$, and mean reaction time increased by 1.29 s, $d = 1.78$, $BF = 1.4 \times 10^6$.

Concurrent load in phase two eliminated the single feature pretraining advantage; mean accuracy differed by less than .01, $d = .08$, with Bayesian evidence for the null, $BF = .08$. Reaction time increased by 0.49 s, $d = 0.58$, with weak Bayesian evidence for the null, $BF = 0.36$. Given these results, one would expect performance to be worse in the single feature load condition than the single feature no-load condition, and this is indeed the case; mean accuracy dropped by 0.15, $d = 1.18$, $BF = 101.7$. There was Bayesian evidence that the two conditions did not differ in reaction time; on average reaction time was 0.50 s shorter under load, $d = 0.49$, $BF = 0.27$. In the polymorphous pretraining conditions, load had no effect on accuracy, with Bayesian evidence for the null; mean accuracy drop = .04, $d = 0.48$, $BF = 0.22$. The effect of load on reaction times in the polymorphous conditions was inconclusive, mean increase = 0.30 s, $d = 0.69$, $BF = 0.38$.

In summary, Experiment 5 demonstrated that the single feature pretraining advantage in polymorphous classification was eliminated by the concurrent presence of a counting task. This result is predicted by the dimensional summation hypothesis. It is also congruent with another experiment from our lab, in which the prevalence of overall similarity classification in a different procedure (the match-to-standards task) was reduced by a concurrent counting load (Wills et al., 2013).

General Discussion

Polymorphous concepts are those defined by an *n-out-of-m* rule. Such concepts are hard to learn and, from some perspectives, this is surprising because polymorphous concepts have an overall similarity structure — a type of structure assumed to be commonplace in real-world categories. However, the difficulty of acquiring polymorphous concepts is predicted by the dimensional summation theory of overall similarity classification (Milton & Wills, 2004). This theory states that overall similarity classification is achieved through the deliberate counting of the number of stimulus features characteristic of each of the candidate categories. This strategy is hard to apply to polymorphous concepts because each feature, taken individually, only probabilistically predicts the category label. This explanation, in turn, leads
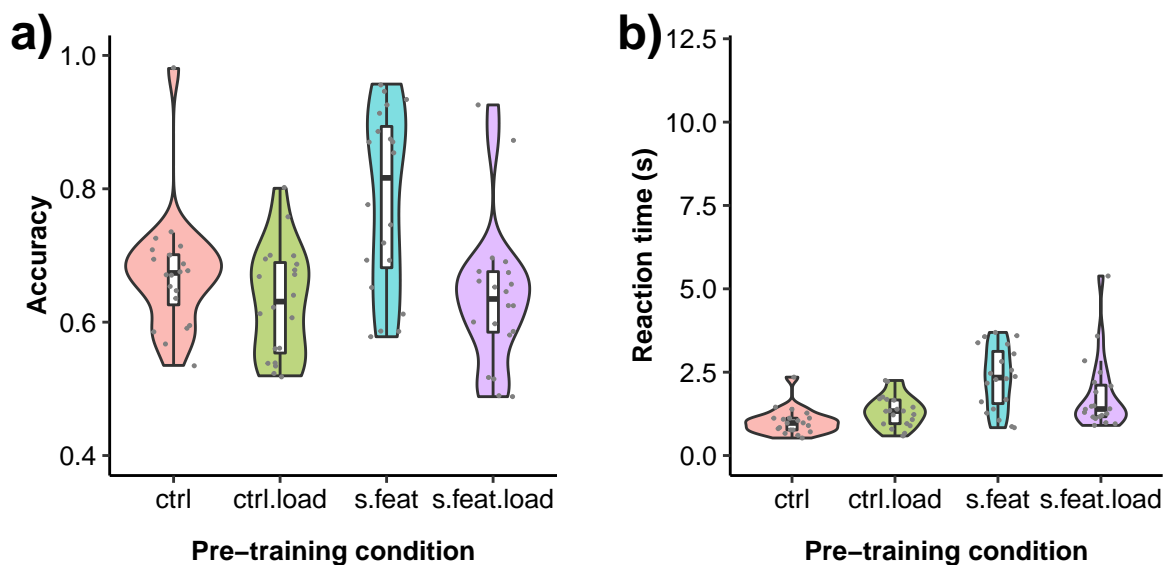
Fig. 6 Accuracy and reaction time in phase two of Experiment 5, as a function of load and pretraining condition. Image credit: Andy J. Wills. CC-BY-SA 4.0. https://osf.io/fdm8r/

to the prediction that deterministic pretraining of the characteristic feature-category associations should facilitate subsequent classification of a polymorphous category structure.

In Experiment 1, we demonstrated the presence of such a single feature pretraining advantage, relative to a control of an equivalent amount of training on the polymorphous problem itself. In Experiment 2, we employed a partial reversal manipulation to demonstrate this advantage critically depended on the information learned in pretraining being applicable to the subsequent polymorphous classification, and hence eliminated a more general strategic or motivational account of the effect. In Experiments 3 and 4, we eliminated a number of other possible accounts of the single feature pretraining effect, and in Experiment 5 we demonstrated that the single feature pretraining effect was eliminated by the presence of a concurrent counting load during the subsequent polymorphous classification phase. Overall, the results of these five experiments provide strong support for the dimensional summation account.

Future research

There are several ways in which the current investigations could be extended to further investigate the dimensional summation hypothesis. For example, the stimulus dimensions in the current experiments are easy to verbalize, highly perceptually discriminable, and mostly spatially separated. Ease of verbalization seems likely to facilitate rule-like processing (see e.g. Kurtz et al., 2013), which seems relevant given that the dimensional summation hypothesis is a rule-like account of behavior. Hence, future work might productively investi-

gate whether the effects of single feature pretraining (SFPT) persist where the stimulus dimensions are harder to verbalize.

It might also be informative to investigate the effects of SFPT with spatially integrated and/or perceptually subtle stimulus dimensions. Other work we have done demonstrates that such stimuli are less likely to be sorted spontaneously into overall similarity groups than the current stimuli (Milton, Copestake, Satherley, Stevens, & Wills, 2014; Milton, McLaren, Copestake, Satherley, & Wills, in press; Milton, Viika, Henderson, & Wills, 2011; Milton & Wills, 2004, 2008). Thus, our prediction is that SFPT of such stimuli would be particularly beneficial to subsequent polymorphous classification of them. This is because induction of an overall similarity strategy is particularly likely to be required for such stimuli and, under our account, SFPT is an effective way of inducing such a strategy.

The prevalence of spontaneous overall similarity sorting is also affected not only by stimulus properties, but also by the underlying category structure (Pothos & Close, 2008). It might therefore be worthwhile examining whether the benefits of SFPT generalize to category structures other than the polymorphous structure considered here. For example, in a polymorphous category structure, all category members are presented with equal frequency. This leads to participants encountering low typicality category members more frequently than high typicality category members. This is because, with a polymorphous category structure, there are more low typicality members than high typicality members (see Table 1). It might be argued that, in more naturalistic categories, the reverse is likely to be true — in other words,

that participants are more likely to encounter high typicality members than low typicality members. For this reason, one might wish to consider extending the current investigation to category structures where high typicality items dominate.

Dimensional summation theory predicts that the accuracy benefits of SFPT would be present but smaller for such structures. They would be smaller because single-dimension responding leads to high accuracy in such structures anyway, so there is less to be gained by switching to a dimensional summation strategy. For example, in a case where only the high-typicality members are presented (the first six rows of Table 1), participants can achieve .83 accuracy by using any one dimension and ignoring all the others. At limit SFPT, even if completely effective in inducing overall similarity responding in all participants, can only lead to an increase in accuracy of .17. This is substantially smaller than the upper limit improvement of .31 possible with the current category structure.

Another category structure that might be interesting to investigate is Sephard et al.'s Type VI structure, which is well known to be particularly hard to learn, and certainly harder than the polymorphous (Type IV) structure used in the current studies (Nosofsky et al., 1994). In a five dimensional version of the Type VI problem, and using the current stimulus set as an illustration, a stimulus belongs to category A if it contains one, three, or five of the features `[horizontal, yellow, stars, long base, fine]`, and category B otherwise. This results in two 16-item categories for which no feature, considered individually, has any diagnostic value. The dimensional summation hypothesis therefore correctly predicts that a Type VI problem is very hard to learn — no feature is characteristic of either category, and hence there are no single dimension rules that can be derived from Type VI training. With no single dimension rules, there is nothing to summate.

The core concepts of the dimensional-summation hypothesis also predict that SFPT could, in principle, be effective in increasing the accuracy of Type VI classifications, particularly if that pretraining involved different category labels to the subsequent Type VI phase. This is because such pretraining would facilitate the application of a parity rule (e.g. "It's category C if it has 1, 3, or 5 category A features, otherwise it's category D"). We describe this as an "in principle" accuracy benefit because it's an open question how often such a parity rule would spontaneously occur to participants under these conditions, and how effortful it would be to apply if it did occur to them. The counting rule for polymorphous problems, in contrast, seems very likely to occur to participants and to be rather easy for them to apply, as it is essentially the same as a simple majority voting system (or, equivalently, a simple "pros and cons" comparison). In the case of Type VI problems, it may be that some pretraining in parity rules would be required in order SFPT to be effective.

Future research might also investigate whether the current results are specific to the binary-valued stimulus dimensions we used (e.g. horizontal versus vertical). The dimensional summation hypothesis is also applicable to continuous dimensions (e.g. line orientation in degrees rather than just 0 versus 90 degrees). SFPT in which participants learn the boundary between category A and B on each continuously varying dimension separately, should be beneficial to subsequent polymorphous classification for the same reasons it is beneficial for binary valued dimensions.

However, perhaps the biggest question left unanswered by the dimensional summation hypothesis is how one reconciles the idea of a slow deliberative summation of evidence with the fact that natural categories, which are assumed to be polymorphous, can be classified extremely rapidly (Thorpe & Imbert, 1989). Of course, there may be a big difference between how we act when we are first learning a new category, compared to how we act after the thousands of hours of practice we all have on familiar real-world concepts. Clearly there is a set of methodologically difficult experiments that could be performed here to look at the effect of very extended practice on unfamiliar polymorphous concepts (c.f. Logan & Klapp, 1991; Soto, Waldschmidt, Helie, & Ashby, 2013). We suggest that such experiments might be an interesting topic for future research.

Alternative theoretical frameworks

It is possible to express the essence of the dimensional summation account in a variety of alternative theoretical frameworks. For example, an anonymous reviewer argued that the critical component of single-feature pretraining might be that it helped emphasize the within-category correlations between stimulus features. We agree. It is a central part of the dimensional summation account that, after SFPT, participants know that `[horizontal, yellow, stars, long base, fine]` are the characteristic features of category A. In other words, we argue that it is critical they discover that these features go together; that they discover that the features correlate within the category.

The same reviewer argued that single feature pretraining might encourage participants to move away from single feature rules in phase two, because the pretraining makes it clear that all of the features have diagnostic value, and because in phase two it becomes rapidly apparent that no single feature is a perfect predictor. Again, we agree — these are the reasons dimensional summation theory predicts single feature pretraining is effective. We note that the results of the probabilistic single feature condition in Experiment 3 directly support this interpretation.

Finally, one might question the centrality of counting in our dimensional summation hypothesis. The concurrent load manipulation of Experiment 5 provides good evidence that,

whatever is going on in phase two after SFPT, it is somewhat effortful. However, it is of course possible that this effortful process is something other than counting. Researchers may wish to specify an alternative account, and test it empirically.

Applications

By understanding the single feature pretraining effect in polymorphous classification, we perhaps better understand how people learn polymorphous concepts, and it is widely believed that many natural concepts are polymorphous. On that basis, our hypothesis is that natural concepts are learned by effortful combination of information, compiling evidence to support one classification over another.

Of course, as discussed, this hypothesis requires further investigation with a broader range of stimulus types than the single set employed in the current work. However, if our results turn out to have some generality, one potential application would be to use an analog of the single feature pretraining procedure to speed the training of natural concepts in the classroom. This idea goes beyond the old adage to split complex problems into simple ones, and adds that it might be productive to caricature probabilistic relationships as deterministic. In that regard, there are parallels to transfer along a continuum procedures, which exaggerate perceptual differences to speed learning, and which have been shown to be advantageous in the training of difficult real-world discriminations (Hornsby & Love, 2014; McClelland, Fiez, & McCandliss, 2002).

Concluding remarks

Although the dimensional summation account of overall similarity classification is at variance with some older ideas about how overall similarity classification works (e.g. Ashby et al., 1998; Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983), it is fully consistent with a substantial body of more recent work, across several different procedures. For example, it is consistent with results from the match-to-standards procedure (Milton et al., 2008; Milton & Wills, 2004; Milton et al., 2009; Wills et al., 2013), the triad procedure (Wills et al., 2015), the criterial-attribute procedure (Wills et al., 2015), and information-integration category learning procedure (Carpenter et al., 2016; Edmunds et al., 2015, 2018, 2019; Newell et al., 2013). It finds support from not only human behavioral data, but also from comparative work with rats and pigeons (Lea et al., 2018, 2006; Wills et al., 2009) and from functional imaging data in humans (Carpenter et al., 2016; Milton et al., 2009). In conclusion, the dimensional summation hypothesis is a plausible account of overall similarity classification in a wide variety of lab-based

conditions, and the current experiments add the acquisition of polymorphous concepts to that growing evidence base.

Author contributions

AJW: Rationale, design, programming, analysis, interpretation, and write up. **LE**: Ran and performed initial analyses for Experiments 1–3; contributed to write up. **FM**: Ran and performed initial analyses for Experiment 4; contirbuted to write up. **GC**: Ran and performed initial analyses for Experiment 5; contributed to write up. **TB**: Contributed to the theoretical interpretation of results.

Open practices

The data and code for all analyses for all experiments are available at the OSF addresses given in each Results section. The stimuli are available at the same locations.

Appendix: Explanation of the Church effect

The Church effect (Church, 1964) concerns a potential confound in yoked designs. The problem is that yoked designs can lead to artifactual differences between groups if there are individual differences across participants. For the purposes of illustration, imagine a simpler version of Experiment 1 in which *both* conditions receive single feature pretraining. One would not expect to find a difference between these two conditions, as they differ only in whether they contain yoked or non-yoked participants. In the Experimental group, each participant is trained to an errorless criterion (as we did in Experiment 1, SFPT condition). In the Control group, each participant receives exactly the same number of trials of single feature pretraining as one randomly selected participant in the Experimental group (a yoking procedure). Now, consider those pairs of participant, and imagine that the two members of the pair learn at different rates. If the yoked participant learns faster than the experimental participant, the yoked participant will also reach an errorless criterion, and so their terminal performance as measured by percent correct will be the same. However, if the yoked participant learns slower than the experimental participant, then they may not reach an errorless criterion and their terminal percent correct will be lower than the participant to which they are yoked. So, when we work out the mean performance of the yoked group, we will average together some people who are worse than the Experimental group, and some who are the same. This will of course lead to a lower mean in the yoked condition. So, the yoking procedure has been ineffective in matching learning in the two conditions, as measured by percent correct. This problem can be avoided by giving everyone the same number

of training trials, assuming random allocation of these individual differences across the two conditions. This is the procedure we use from Experiment 2 onwards.

References

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

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–61.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory and Cognition*, *30*, 666–677.

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: amnesia and the problem of error elimination. *Neuropsychologia*, *32*, 53–68.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*, 3557–3574.

Champely, S. (2018). pwr: Basic functions for power analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=pwr` (R package version 1.2-2)

Church, R. M. (1964). Systematic effect of random error in the yoked control design. *Psychological Bulletin*, *62*, 122–131.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Dennis, I., Hampton, J. A., & Lea, S. E. G. (1973). New problem in concept formation. *Nature*, *243*, 101–102.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *Quarterly Journal of Experimental Psychology*, *68*, 1203–1222.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2018). Due process in dual process: Model-recovery simulations of decision-bound strategy analysis in category learning. *Cognitive Science*, *42*, 833–860.

Edmunds, C. E. R., Wills, A. J., & Milton, F. (2019). Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology*, *72*, 151–167.

Filoteo, J., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, *21*, 415–423.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician*, *52*, 181–184.

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*, 72–76.

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

Kemler Nelson, D. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*, 734–759.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*, 225–247.

Kurtz, K. J., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 552–72.

Lea, S. E. G., Lohmann, A., & Ryan, C. M. E. (1993). Discrimination of 5-dimensional stimuli by pigeons: Limitations of feature analysis. *Quarterly Journal of Experimental Psychology*, *46B*, 19–42.

Lea, S. E. G., Pothos, E. M., Wills, A. J., Leaver, L. A., Ryan, C. M. E., & Meier, C. (2018). Multiple feature use in pigeons' category discrimination: The influence of stimulus set structure and the salience of stimulus differences. *Journal of Experimental Psychology: Animal Learning and Cognition*, *44*, 114–127.

Lea, S. E. G., Wills, A. J., & Ryan, C. M. E. (2006). Why are artificial polymorphous concepts so hard for birds to learn? *Quarterly Journal of Experimental Psychology*, *59*, 251–67.

Le Pelley, M. E., Newell, B. R., & Nosofsky, R. M. (2019). Deferred feedback does not dissociate implicit and explicit category learning systems: Commentary on Smith et al. (2014). *Psychological Science*.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*, 720–738.

Logan, G. D., & Klapp, S. T. (1991). Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*, 179–195.

McClelland, J., Fiez, J., & McCandliss, B. D. (2002). Teaching the /r/-/l/ discrimination to Japanese adults: behav-

ioral and neural aspects. *Physiology and Behavior*, *77*, 657–662.

Milton, F., Copestake, E., Satherley, D., Stevens, T., & Wills, A. J. (2014). The effect of pre-exposure on family resemblance categorization for stimuli of varying levels of perceptual difficulty. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1018–1023). Cognitive Science Society.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 407–415.

Milton, F., McLaren, I. P. L., Copestake, E., Satherley, D., & Wills, A. J. (in press). The effect of pre-exposure on overall similarity categorization. *Journal of Experimental Psychology: Animal Learning and Cognition*.

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*, 1326–1336.

Milton, F., Viika, L., Henderson, H., & Wills, A. J. (2011). The effect of time pressure and the spatial integration of the stimulus dimensions on overall similarity categorization. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 795–800). Austin, TX: Cognitive Science Society.

Milton, F., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 407–415.

Milton, F., & Wills, A. J. (2008). The influence of perceptual difficulty on family resemblance sorting. In B. Love, K. McRae, & M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2273–2278). Austin, TX: Cognitive Science Society.

Milton, F., Wills, A. J., & Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. *NeuroImage*, *46*, 319–326.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: a state-trace analysis. *Memory and Cognition*, *38*, 563–81.

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves "implicit" perceptual categorization. A comment on Filoteo, Lauritzen and Maddox, 2010. *Psychological Science*, *24*, 386–3389.

Nomura, E. M., Maddox, W. T., Filoteo, J., Gitelman, D., Parrish, T. B., Mesulam, M. M., & Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*, 37–43.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*, 352–369.

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, *107*, 581–602.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/ (R version 3.6.1)

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 574–605.

Ryle, G. (1951). Thinking and language. *Proceedings of the Aristotelian Society*, *65*, 65–82.

Seabrooke, T., Hollins, T. J., Kent, C., Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, *104*, 70–82.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13), Whole No. 517.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, A., Barbara, & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*, 447–57.

Smith, J. D., & Kemler Nelson, D. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General*, *113*, 137–159.

Smith, J. D., & Shapiro, J. (1989). The occurrence of holistic categorization. *Journal of Memory and Language*, *28*, 386–399.

Soto, F. A., Waldschmidt, J. G., Helie, S., & Ashby, F. G. (2013). Brain activity across the development of automatic categorization: a comparison of categorization tasks using multi-voxel pattern analysis. *NeuroImage*, *71*, 284–297.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information-integration but not rule-based category learning. *Psychological Science*, *19*, 1169–1177.

Terrace, J. S. (1963). Discrimination learning with and without "errors". *Journal of the Experimental Analysis of Behavior*, *6*, 1–27.

Tharp, I. J., & Pickering, A. D. (2009). A note on DeCaro,

Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information-integration category learning. *Cognition*, *111*, 411–5.

Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist modeling. In Z. Pfeifer, F. Schreter, F. Fogelman-Soulie, & L. Steels (Eds.), *Connectionism in perspective* (pp. 63–92). Amsterdam: Elsevier.

Torchiano, M. (2019). effsize: Efficient effect size computation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=effsize` (R package version 0.7.6)

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

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 103–112.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from `https://ggplot2.tidyverse.org`

Wickham, H., Francois, R., Henry, L., & Muller, K. (2019). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=dplyr` (R package version 0.8.3)

Wills, A. J., Edmunds, C. E. R., Le Pelley, M. E., Milton, F., Newell, B. R., Dwyer, D. M., & Shanks, D. R. (2019). Dissociable learning processes, associative theory, and testimonial reviews: A comment on Smith and Church (2018). *Psychonomic Bulletin and Review*.

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33.

Wills, A. J., Lea, S. E. G., Leaver, L. A., Osthaus, B., Ryan, C. M. E., Suret, M. B., . . . Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: I. Unidimensional classification does not necessarily imply analytic processing; evidence from pigeons (Columba livia), squirrels (Sciurus carolinensis), and humans (Homo sapiens). *Journal of Comparative Psychology*, *123*, 391–405.

Wills, A. J., & McLaren, I. P. L. (1997). Generalization in human category learning: A connectionist explanation of differences in gradient after discriminative and non-discriminative training. *Quarterly Journal of Experimental Psychology*, *50A*, 607–630.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology*, *66*, 299–318.

Wills, A. J., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory and Cognition*, *34*, 17–27.

Wittgenstein, L. (1958). *Philosophical investigations*. Oxford, UK: Blackwell.

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