The fickle nature of similarity change as a result of categorization

Emmanuel M. Pothos & Irene Reppa

Please address correspondence regarding this article to Emmanuel M. Pothos, Department of Psychology, City University London EC1V 0HB or to Irene Reppa, Department of Psychology, Swansea University, Swansea SA2 8PP, UK. Electronic mail can be sent to e.m.pothos@gmail.com or to i.reppa@swansea.ac.uk.

Running head: similarity changes **Text word count**: 5,891

Abstract

Several researchers have reported that learning a particular categorization leads to compatible changes in the similarity structure of the categorized stimuli. The purpose of this study is to examine whether different category structures may lead to greater or less corresponding similarity change. We created six category structures and examined changes in similarity within categories or between categories, as a result of categorization, in between-participant conditions. The best supported hypothesis was that the ease of learning a categorization affects change in within categories similarity, so that greater (within categories) similarity change was observed for category structures which were harder to learn.

There is widespread evidence that learning to categorize stimuli in a particular way leads to corresponding changes in the similarity structure of the stimuli (for brevity, we will henceforth refer to such changes just as ‘similarity changes’). For example, Goldstone (1994) found increased perceptual sensitivity for schematic, meaningless stimuli categorized in different categories and, in some cases, decreased perceptual sensitivity for stimuli categorized in the same category. Schyns, Goldstone, and Thibaut (1997) argued that category learning can lead to the development of new features, which may alter the similarity of the categorized items. Introducing a category boundary in a continuum of stimulus variation often results in enhanced discriminability on either side of the boundary (Harnad, 1987). Some researchers have reported differences in color perception across linguistically different communities, and this is another facet of the influence of categories on similarity (Roberson et al., 2005). There have been several reports of broadly analogous effects, across diverse category learning paradigms (e.g., Gureckis & Goldstone, 2008; Goldstone & Steyvers, 2001; Lupyan, 2012; Ozgen & Davies, 2002; Schyns & Oliva, 1999; Stevenage, 1998) and even in the animal cognition literature (e.g., Delamater, 1998, 2012).

Such research has flourished for several reasons. It is theoretically important, since it is at the heart of answering core issues regarding representation and the processing of sensory input. Does the cognitive system aim to construct a faithful representation of sensory input? Or does it aim for representations, which achieve a compromise between information from sensory input and functional considerations relating to representation/ categorization? Moreover, do the factors which moderate similarity change also influence the kind of feature transfer effects, identified by Goldstone (1995)? He reported that, e.g., stimuli assigned to a category of

predominantly blue objects were perceived as more blue than they were really were. This finding has clear applications in practical domains (e.g., prejudice and social stereotyping).

Understanding the nature of similarity changes is also relevant to formal models of categorization. Exemplar and prototype models both employ a sophisticated computational machinery for altering the representation of the studied stimuli, to accommodate requirements from the learned categorizations (Minda & Smith, 2001; Nosofsky, 1984). From such models, we know that, if the required classification assumes one stimulus dimension to be more diagnostic than others, then there would be increased attentional weight for this dimension. However, this research does not discriminate between the possibilities that such changes are changes in the representation of the stimuli or are simply moderators of stimulus information at the point of classification decisions. The latter appears the standard assumption, with stimulus representations implied static throughout the categorization process (see Pothos and Wills, 2011, for a comprehensive overview).

There has been controversy regarding the exact nature of similarity changes as a result of categorization. For example, it is possible that categorizing stimuli in a certain way does not alter our perception of the stimuli but, rather, makes certain stimuli more or less similar because of the augmentation of the stimulus representations with an additional feature corresponding to category labels (e.g., Goldstone, Lippa, Shiffrin, 2001; McMurray et al., in press; Roberson & Davidoff, 2000; Sloutsky & Fisher, 2004). However, if across broadly matched categorization conditions, we find that in some conditions there are corresponding similarity changes, but in other ones there are not, then one can make the additional step of inferring similarity changes over and above changes due to just the category label (see

also Roberson et al., 2007). It is an interesting issue to explore which part of similarity change is due to the linguistic label and which due to other aspects of stimulus representation, but one which we reserve for future work.

There is very little prior work on the factors which make it more or less likely to observe similarity changes (Folstein, Gauthier, & Palmeri, 2010; Freedman, Riesenhuber, Poggio, & Miller, 2003; Jiang, Bradley, Rini, Zeffiro, Vanmeter, & Riesenhuber, 2007; Livingston, Andrews, & Harnad, 1998). Folstein et al. (2010) wanted to understand why Jiang et al. (2007; see also Freedman et al., 2003) failed to observe any changes of similarity as a result of categorization, even after extensive successful training. They observed that the nature of the underlying stimulus space in the case of Jiang et al. (2007) was more complex than that of, e.g., Goldstone (1994) and they argued that it is this additional complexity which prevented the emergence of the anticipated similarity effects (complexity related to whether the stimulus space was equivalent to a standard two-dimensional coordinate space or, rather, had a more complex form). Folstein at al. (2010) supported their argument by creating matched experiments which differed only in terms of the nature of the underlying stimulus space and finding effects of categorization on similarity only when the stimulus space was simpler. The idea that complexity and difficulty, broadly defined, can impact on similarity change appears in Livingston et al. (1998) as well. These investigators looked at more and less well separated categories, and reported that similarity changes were equivalent.

Such research is indicative, though not conclusive. First, similarity change is clearly not a unitary concept, but rather it can be defined in different ways. Work on categorical perception (e.g., Goldstone, 1994; Harnad, 1987) explored similarity change in terms of compression (acquired equivalence) and expansion measures

(acquired distinctiveness; such measures have been popular in the corresponding animal learning literature as well, e.g., Delamater, 1998). But, categorical perception studies typically concern the impact of introducing a category boundary within a uniform stimulus space (i.e., in a stimulus space whereby the stimuli do not naturally cluster into categories). By contrast, we are presently more interested in putative similarity changes, when participants are taught to categorize stimuli which are well-clustered (to varying degrees) with respect to specific categories. Similarity change consistent with a taught classification could be reflected either in the stimuli within a category becoming more similar (henceforth within similarity change) or the stimuli between categories becoming less similar (henceforth between similarity change). Such measures are clearly analogous with those of compression and expansion. Moreover, they are consistent with the bulk of theoretical work in unsupervised categorization, which emphasizes the relevance of within and between category similarity as determinants of category structure (e.g., Love, Medin, & Gureckis, 2004; Pothos & Chater, 2002; Pothos & Bailey, 2009; Rosch & Mervis, 1975).

When considering the issue of similarity change for well-clustered category structures, we can recognize that some are more intuitive than others. Category intuitiveness characterizes a category structure and corresponds to the extent to which the category structure is natural, obvious, and likely to be spontaneously generated by participants (Pothos & Chater, 2002; Pothos et al., 2011; see also Feldman 2000, Shepard et al., 1961). Moreover, category structures which are more intuitive than others will be easier to learn as well, if one discounts the extra memory burden of keeping track of several category labels (Pothos et al., 2012). This consideration is important, as similarity change is typically (and also in the present study) studied as a

result of learning a categorization, rather than the spontaneous generation of a category.

The consideration of category intuitiveness/ difficulty can lead to the first hypothesis of when we are more likely to observe similarity change. Learning a more difficult category structure requires a greater cognitive effort. As the learner is faced with a harder task in identifying the intended categorizations of the relevant stimuli, so it perhaps becomes more likely that the stimulus representations may be elaborated in a way which supports the intended categorizations. But in what way? Some researchers have reported that different category learning tasks lead to the development of different category information. For example, classification learning appears to encourage emphasis on diagnostic features and training via feature inference tasks emphasis on prototypical features (Chin-Parker & Ross, 2004; Markman & Ross, 2003). The analogies between such research and the present empirical questions are somewhat tenuous. Nevertheless, we can perhaps motivate the idea that, where the category boundary is relatively easy to extract, any elaborative processes which contribute to category change will concern within category similarity, otherwise between category similarity would be expected. The simplest operational measure of category intuitiveness is learning difficulty, e.g., number of trials or errors to criterion, and it is this approach we adopt presently.

In sum, our first hypothesis is that greater category difficulty leads to greater similarity change (perhaps more so within category similarity change, for category structures for which a category boundary is easily extracted). Note that there are many models which make predictions of whether a category structure is more or less intuitive, but, in simple cases, this can be established by inspection. We employed category boundaries aligned to one of the dimensions of physical variation. Then,

clusters of stimuli intended for different categories could be closer or further away from each other, thus increasing or decreasing category difficulty (cf. Livingston et al., 1998). Corresponding easy and difficult category structures could thus be specified which are directly equivalent, in that the category boundary would have the same shape and be specified in terms of the same physical dimension. Specifically, consider the category structures labeled as Width Easy, Width Difficult, in Figure 1. These are two highly matched category structures, but such that the relevant categories are well-separated in one case (Width Easy; an easy category structure) and poorly separated in the other (Width Difficult; a difficult category structure). The Height Easy, Height Difficult category structures are exactly analogous and simply counterbalance the physical dimension along which the category boundary is specified.

The issue of category structure is multifaceted and, plausibly, cannot be resolved just by considering category difficulty. Several researchers have presented arguments for why certain kinds of category structures may be processed in qualitatively different ways, even if overall category difficulty can be in principle equated. For example, according to the COVIS model of categorization (Ashby & Ell, 2002; Ashby et al., 1999; Ashby, Queller, & Berretty, 1999), category boundaries aligned to a dimension of variation elicit learning processes distinct from ones which are not (e.g., Height or Width category structures vs. diagonal ones, as in Figure 1). In the former case, a hypothesis-testing learning process is likely to lead fairly quickly to explicit knowledge of the appropriate category boundary, which can then be applied to classify all stimuli, while in the latter case a passive mode of learning is more likely to be adopted. It is possible that such differences in the learning process may impact on the degree of similarity change.

Another relevant consideration concerns whether it is possible to specify a linear category boundary for a category structure or not. The latter category structures are called non-linearly separable (NLS) and are significant in categorization theory, because they are consistent with one influential approach to categorization (exemplar theory) but not another (prototype theory; Pothos, Chater, & Stewart, 2004; Ruts, Storms, & Hampton, 2004). It is currently still an issue of controversy whether NLS category structures are harder to learn than matched linearly separable ones (e.g., Blair & Homa, 2001; Pothos & Bailey, 2009; Smith, Murray, & Minda, 1997;Yamauchi, Love, & Markman, 2002). Moreover, linear separability is an important constraint in connectionist modeling as well, as NLS problems have to be transformed into linearly separable ones at their hidden layer, otherwise learning is not possible (indeed, the inability of perceptrons to learn NLS category structures has been at the heart of the famous critique of Minsky & Papert, 1969; see also Rumelhart & McClelland, 1986). If the cognitive system shares processing constraints with connectionist systems, maybe it would try to re-represent a NLS classification in an LS way, so that there would be more similarity change in learning an NLS classification, compared to an LS one.

It should now be clear that, over and above category difficulty, there are other considerations which may impact on the cognitive processes for learning of a categorization and, *possibly*, corresponding similarity changes. Therefore, a second (more exploratory) hypothesis corresponds to whether similarity change will be different for category structures, broadly matched for difficulty, but differing in terms of whether the category boundary is aligned with a dimension of physical variation, is diagonal, or the categorization is NLS (Figure 1). Note that, regarding the diagonal category structure, we tested two category structures, one with the boundary sloping

upwards (Diagonal A, as shown in Figure 1) and another with the boundary sloping downwards (Diagonal B).

----------FIGURE 1----------

Experimental investigation

Participants and Design

We recruited 219 experimentally naïve participants, all Swansea University students, mostly in the Psychology Department. Participants were predominantly (approximately 80%) females of university age. The experiment lasted approximately 50 minutes for the experimental groups and 30 minutes for the control groups. All participants received course credit for their participation.

There were 139 participants experimental group participants, with approximately 20 in each of the seven (between-participant) experimental conditions. Each condition corresponded to asking participants to learn one of the category structures in Figure 1 (accordingly, the conditions are labeled as Width Easy, Width Difficult, Height Easy, Height Difficult, and NLS; the diagonal category structure in Figure 1 will be called Diagonal A and the matched one with a sloping downwards category boundary Diagonal B).

For all experimental conditions there was a control group providing similarity ratings for the stimuli, but without having gone through the categorization task first. The control groups were shared between some of the conditions (since the stimuli were the same). Accordingly, for the Width Easy, Height Easy, and NLS conditions there was a control group of 20 participants, for the Width Difficult and the Height Difficult conditions a different control of 20 participants, for the Diagonal A group a

different control group of 20 participants, and finally a different group of 20 control participants for the Diagonal B condition.

Materials

We used stimuli that varied along two separable dimensions of variation. Figure 2 shows an example of the stimuli for the Width Easy, Height Easy, and NLS conditions (note that, even though the stimuli in these conditions are identical, the classifications participants were asked to learn differed). We employed yellow surface-rendered arrow-like shapes that varied in terms of the width of the arrowhead (horizontal dimension) and the length of the arrow (vertical). The smallest arrow's trunk measured 4.5 centimeters (cm) in height and its head measured 3.0 cm wide. Twenty-four more stimuli were created by incrementing trunk height and head width by 12%. The stimuli employed in the experimental conditions were subsets of this original set of stimuli. The shortest arrow trunk in all six conditions was 4.5cm high and the narrowest arrow head 3.0cm wide. The tallest arrow trunk was 12.5cm in the Width Easy, Height Easy, Diagonal, and NLS conditions and 7.1cm in the Width Difficult and Height Difficult conditions. The widest arrow head was 8.3cm in the Width Easy, Height Easy, and NLS conditions, 4.7cm in the Width Difficult and Height Difficult conditions and 5.3cm in the Diagonal A and Diagonal B conditions.

----------FIGURE 2----------

Procedure

A standard supervised categorization task was employed. A stimulus was presented at the center of a computer screen against a white background, until the participant decided whether it belonged to category A or B, at which point he/she

received corrective feedback. Participants continued to categorize stimuli until no mistakes were made for 32 consecutive trials (i.e., all stimuli shown twice) or for a maximum of 256 trials (note that, given enough training, there is evidence that participants can learn very hard category structures; McKinley & Nosofsky, 1995). Five participants failed this criterion (three in the NLS condition and two in the Diagonal A condition) and these participants were not asked to complete the similarity part of the study. It is consistent with expectation that, if any participants were to fail the learning task, this would happen in the more difficult category structures.

Participants, who completed the categorization task successfully, subsequently received the similarity ratings task. In that task, each trial started with a 'Ready?' prompt at the center of the screen. Two stimuli appeared at the screen center for 500ms each, one after the other, with an inter-stimulus interval of 500ms. All possible 16x16=256 stimulus pairs were presented and participants were asked to rate their similarity on a 1-9 scale, such that 1 corresponded to 'very dissimilar' and 9 to 'very similar'. Participants were encouraged to use the entire scale. Participants in the control groups went through the similarity ratings, without having done the categorization task first.

Results

Inclusion criteria

We employed two simple checks that the participants were sufficiently attentive during the similarity ratings task. Participants were excluded if they did not use the whole 1-9 similarity rating scale (specifically, those who used five or fewer rating values in total) and if they failed to rate two identical stimuli as most similar

(by assigning to them a rating of 7 or higher) more than three times. This procedure led to the elimination of 3 participants from the Width Easy group, 3 from the Height Easy one, 2 from the Width Difficult one, 1 from the NLS group, 1 from the Diagonal B one, and 5 participants from the control groups. Note that the similarity ratings for pairs of identical stimuli were not employed for any purpose, other than to check that participants were attending to the task.

Learning Results

We examined the number of trials required to learn the different category structures (trials to criterion) and the number of errors until perfect classification had been achieved (Table 1). Both the trials to criterion and the errors varied across category structures ($F(6,123)=13.83$, $p<.0005$ and $F(6,122)=11.13$, $p<.0005$, respectively). Note that trials to criterion and errors correlated highly with each other ($r=.86$, $p<.0005$), so, henceforth, we will just consider trials to criterion. Note also that a preliminary assessment revealed no significant difference in learning between the Diagonal A, Diagonal B conditions, so we pooled results across the two conditions (henceforth, the combined condition will be referred to as the Diagonal condition).

We adopted a planned contrasts approach, regarding the comparisons for learning, between the conditions relevant to the two hypotheses of interest. The first hypothesis concerns whether more difficult category structures lead to greater similarity change and the category structures created for this hypothesis were the Width Easy, Difficult ones and the Height Easy, Difficult ones. As expected, a contrast comparing learning in the Width Easy and Height Easy category structures

against the Width Difficult and Height Difficult ones was significant ($t(70)=2.53$, $p=.014$; here and elsewhere, all pairwise comparisons were two-tailed).

The second hypothesis concerned differences in similarity change, between category structures broadly matched in difficulty, but differing in the type of category boundary involved. The category structures we intended for this hypothesis were the NLS, Diagonal and the hardest one between Width Difficult and Height Difficult (as Table 1 shows, the Height Difficult condition was associated with the greatest number of trials, though note that the difference with Width Difficult was not statistically significant). We compared trials to criterion in the Diagonal condition vs. the Height Difficult one ($t(57)=3.278$, $p=.002$) and in the Diagonal condition vs. the NLS one ($t(56)=3.320$, $p=.002$). The first comparison is relevant as it allows us to explore similarity change for category structures with a linear boundary, but with (Height Difficult) and without (Diagonal) alignment with a dimension of physical variation. The second comparison concerns a category structure with a linear boundary (Diagonal) vs. one for which a linear boundary is not possible (NLS). The only implication for the similarity change analyses from these results is that we need to consider learning trials to criterion as a covariate, in the corresponding statistical tests for similarity change, for the second hypothesis.

Similarity measures

We sought to quantify similarity change with two dependent variables, motivated both from work on similarity change and theoretical approaches to category structure. Within (category) similarity referred to the extent to which the stimuli within the different categories in a category structure were perceived to be similar to each other.

The definition for between (category) similarity was analogous, but for the fact that this concerned stimuli in different categories.

Within and between similarity variables were computed directly from the similarity ratings participants provided, by taking into account all the relevant stimulus pairs. For a particular participant, for a particular category structure, within similarity would be the average of his/ her similarity ratings for all pairs of stimuli, such that both stimuli were in the same category. Between similarity was computed as the average of all pairs of stimuli in different categories. Also, within similarity *change* refers to the change in within similarity, as a result of category learning, and likewise for between similarity *change*.

Control groups of participants provided similarity ratings for the stimuli in each category structure condition and this information was used to compute within, between similarity values, for each category structure, prior to learning. These within, between similarity values were then compared with similarity values from participants trained with the corresponding category structures. If the similarity structure of the stimuli changes so that it becomes more consistent with the learned categorization, we expect within similarity after learning to be increased, relative to the value without learning and between similarity after learning to be decreased, relative to the value prior to learning (but, in principle, between similarity change is independent of within similarity change).

Similarity results

In this section we employ families of pairwise comparisons. With such families, there is a risk of Type 1 error and so we applied the Bonferroni Holm correction for significance levels (e.g., Abdi, 2010), as a suitable compromise between the need to

control for family-wise Type 1 error, while not inflating Type 2 error and reducing power (e.g., see Nakagawa, 2004, and Perneger, 1998, who argued against the use of Bonferroni corrections in multiple t-tests). Our reporting strategy for multiple t-tests is to report the *uncorrected* p-value and then state whether this is significant or not, according to the Bonferroni Holm correction.

The category structures were designed so that some were meant to be easier to learn than others. A concern was that for intuitive category structures, within similarity prior to learning may have been so high that no further changes would be possible after learning (and, likewise, the prior between similarity may have been so low that further decreases due to learning may have been impossible). We therefore conducted single sample t-tests of within similarity values computed from control participants, against the highest possible value for within similarity (nine); likewise, we conducted single sample t-tests of control between similarity values against the lowest possible value for between similarity (one). All the above t-tests were (Bonferroni Holm corrected) significant. Regarding within category similarity values, the magnitude of all t-tests was greater than 15.60 (degrees of freedom varied between 18 and 36). Regarding between category similarity values, the magnitude of all t-tests was greater than 12.28 (degrees of freedom between 18 and 35). So, we can conclude that prior to learning there was 'room for improvement', so to say, with respect to both within and between similarity values, prior to learning.

We next consider the two hypotheses for the conditions which may impact on similarity change. The first hypothesis was that, in otherwise matched category structures, the more difficult ones will lead to greater similarity change. We therefore sought to compare similarity change in the Height Easy vs. Height Difficult classifications and likewise for the Width Easy vs. the Width Difficult ones, by

comparing the similarity values computed from the control participants, with the similarity values from corresponding experimental participants. Since some experimental conditions shared the same control (as the underlying stimuli were identical and it was only the category structure which differed; e.g., Width Easy and Width Difficult), the most appropriate statistical approach to the first hypothesis was a family of independent samples t-tests between the similarity values computed from the control participants and those from the experimental participants (Figure 3; it is interesting to explore similarity change for all category structures, not just the ones relevant to the first hypothesis, and so the Bonferroni Holm correction was applied to a family of comparisons including all category structures).

Regarding within similarity change, we observed significant differences for both the Width Difficult ($t(36)=2.72$, $p=.01$, as just noted, all statements of significance in multiple comparisons are based on the Bonferroni Holm procedure) and Height Difficult conditions ($t(38)=2.86$, $p=.007$), but not the Width Easy or the Height Easy ones. We take this result as consistent with our first hypothesis, insofar that more difficult category structures were more likely to lead to similarity change. More generally, within similarity change was observed for poorly separated category structures, whether category boundaries align with a dimension of physical variation (Height Difficult, Width Difficult) or not (Diagonal; $t(72)=2.79$, $p=.007$) and, finally, even in cases when there is no linear category boundary at all (NLS; $t(36)=2.73$, $p=.01$). Note that within similarity change in all cases was an increase in within similarity, as a result of learning, which shows that stimuli within the same categories became more similar to each other.

Regarding between similarity change, there were no reliable differences in between similarity ratings with and without learning (all p-values greater than .068,

degrees of freedom between 34 and 72). It is clear that the particular set of category structures and measures we employed appears better suited for the study of within similarity change.

We next considered the second hypothesis, that is, whether for, categorizations broadly matched in overall difficulty, the nature of the category boundary impacted on similarity change. We ran a 3x2 between participants ANCOVA, with two independent variables. The three-level variable concerned category structure. We were interested in comparing similarity change as a result of learning broadly equally difficult category structures, with a linear simple boundary (Height Difficult) vs. a linear boundary not aligned to one of the dimensions of variations (Diagonal) vs. a non-linear category boundary (NLS). Recall, there were significant differences in the trials to criterion for some of these category structures, so we employed trial to criterion as a covariate (the covariate does not apply to control participants, so to control participants we assigned the mean from the corresponding experimental conditions). The two-level variable concerned the distinction between the control group and the experimental group.

Regarding within similarity, as expected, the control group vs. experimental group factor was highly significant ($F(1,144)=20.653$, $p<.0005$), but not the category structure factor ($F(2,144)=.987$, $p=.375$) or the crucial interaction ($F(2,144)=.142$, $p=.867$; without the covariate, we obtained $p=.865$). Regarding between similarity change, running the same 3x2 ANCOVA, the category structure factor was significant ($F(2,144)=14.619$, $p<.0005$), but not the control group vs. experimental group factor ($F(1,144)=2.914$, $p=.090$) or the interaction ($F(2,144)=0.267$, $p=.766$; without the covariate, $p=.764$). Thus, there was no evidence that the nature of the category

boundary mattered on these measures, for the Height Difficult, NLS, and Diagonal category structures.

----------TABLE 1----------

----------FIGURE 3----------

Discussion

There has been considerable interest in changes in similarity (or perception) induced as a result of categorization, though few studies have attempted a systematic study of the factors which make such changes likely (for exceptions see Folstein et al., 2003 or Livingston et al., 1998). The overarching question in this research was whether category structure is a relevant factor in trying to understand changes in similarity as a result of categorization. One hypothesis is that representation (and so presumably changes in representation) must be partly driven by cognitive demands on which categories are easier vs. more difficult to learn. A second hypothesis concerned the nature of the category boundary (aligned to a dimension of physical variation vs. diagonal vs. NLS), for category structures broadly equally difficult.

Our approach was to specify a range of category structures, which varied in potentially relevant ways. Two category structures were defined in terms of a category boundary along one dimension of variation, but one was expected to be easier to learn than the other (Height Easy, Height Difficult). Two further category structures were defined in the same way, but with a category boundary defined along the other dimension (Width Easy, Width Difficult). Note that any expectations regarding category difficulty were (partly) confirmed experimentally. We also created an NLS category structure, since in such a case there are interesting, conflicting

expectations regarding the extent to which similarity changes as a result of categorization might take place (Blair & Homa, 2003; Rumelhart & McClelland, 1986). Finally, we included a category structure with a diagonal category boundary, to contrast any findings regarding similarity change with those from category structures in which the category boundary was aligned with dimensions of physical variation. Note that 'category structure' is not a methods variable which can be counterbalanced in the way methods variables are manipulated in e.g. learning or attention experiments. The inclusion of each additional category structure requires an additional between participants' conditions with several participants (20 participants per condition, in the present study). Indeed, other research involving a range of category structures involved restrictions analogous to those in the present study (Pothos et al., 2011; Shepard et al., 1961).

The first hypothesis was that more difficult category structures would lead to greater similarity change. For the matched Height Easy, Difficult and the matched Width Easy, Difficult category structures, we did observe that within similarity change was significantly greater (i.e., stimuli in the same categories becoming more similar) for more difficult, compared to easier, category structures. Interestingly, this finding resonates with evidence from the animal cognition literature, that similarity change may be greater with harder learning tasks (cf. Delamater, 1998, if one assumes that learning to distinguish stimuli from the same modality is harder than distinguishing stimuli from different modalities). We also tried to motivate the idea that the separateness of the clusters in a category structure may impact on whether within or between similarity change was more pronounced (cf. Chin-Parker & Ross, 2004; Markman & Ross, 2003), but there was no support for these suggestions.

The second hypothesis we examined concerned whether, for category structures

of broadly equivalent difficulty, other category structure characteristics might impact on the degree of similarity change. We compared similarity change for difficult category structures with a linear category boundary aligned to a dimension of variation (Height Difficult condition) vs. a more complex linear category boundary (Diagonal condition) vs. a non-linear category boundary (NLS condition). However, there was no evidence that the nature of the category boundary impacted on similarity change, either for within similarity change or between similarity change.

We highlight our inability to detect between similarity change, in apparent contrast with previous work, such as Goldstone's (1994; cf. Harnad, 1987, but even work with non-humans, e.g., Delamater, 1998). The main difference of the present work with previous such work concerns our use of measures of similarity change, which were directly defined against the relevant category structure. This meant that similarity change was measured as an average across several stimulus pairs, while in previous research similarity change was typically examined at the level of individual stimulus pairs (e.g., Goldstone, 1994). Moreover, we employed readily distinguishable stimuli, with the view to detect similarity change as changes in similarity ratings, rather than changes in confusability. Finally, Goldstone (1994) and related studies employed stimuli that uniformly spanned the relevant region of similarity space, so that there were no naturally occurring clusters (and so there was perhaps more room for similarity to change, in a way consistent with a learned categorization). We had to implement other, less major, changes, so as to shift the research focus away from similarity measures on individual pairs and towards similarity measures directly defined on the relevant category structures. We think such a shift is theoretically important and justifies this initial difficulty in understanding the empirical impact of the various methods changes.

In addressing the current research questions, we opted against using a within-participant design, whereby similarity ratings would be collected prior and post category learning. There are two reasons for doing so. First, the repetitive nature of similarity ratings makes it tricky to expect participants to provide too many such ratings (e.g., both prior and post classification) in the same experimental session. Second, extended prior exposure to the stimuli (e.g., if similarity ratings are provided prior to category learning) may lead to confounds, which would be difficult to control. For example, extended exposure may lead to unsupervised categorization of the stimuli (which could be associated with similarity change; Gureckis & Goldstone, 2008) or other changes in representation.

Relatedly, a reasonable alternative approach for collecting control similarity information would be to pre-expose the stimuli to control participants, for a number of trials equivalent to the learning trials for the corresponding categorization. Arguably, such a control would exactly tell us whether exposure to the stimuli vs. learning as such is the critical factor in determining similarity change. The problem with such a control is that it would be extremely difficult to dissociate exposure from any kind of learning, which may occur with prolonged engagement with the stimuli. For example, as just noted, with increased exposure, unsupervised categorization of the stimuli might occur, which could in turn produce some similarity change. How could we prevent unsupervised categorization of the stimuli? Existing theoretical insights do not allow any prescriptions. Overall, it is fair to say that the merits and demerits of different methods for collecting similarity information are fairly well-balanced and no one procedure is obviously better than another.

Such issues could be clarified with greater confidence, if we were able to specify with more precision the mechanisms which lead to similarity change.

Following from the relevant discussion in the introduction, one could speculate that increased category difficulty forces participants to look harder for commonalities between the members of each category. It is currently difficult to offer additional insight, but we can preclude two possibilities regarding the underlying mechanisms. First, the kind of attentional mechanisms postulated in models of categorizations, such as exemplar or prototype theory (e.g., Nosofsky, 1984) are unlikely to be the whole story, since they involve a uniform attentional change, whereas we observed within similarity change, without concomitant between change (though it is possible that a combination of changes in the attentional parameters and the sensitivity parameter may work). Second, it is unlikely that *all* our results could be explained by some process involving explicit hypothesis testing (cf. Ashby et al., 1998), since some of the category structures for which we did observe within similarity change were complex enough to preclude explicit hypotheses (e.g., the Diagonal or the NLS category structures). Overall, models of categorization have emphasized computational principles, rather than process, and clearly further work is needed in this direction.

To sum up, our results enable some specific insights regarding the circumstances which are more likely to lead to similarity change. We hope this work will further incentivize research on this novel and important aspect of representation and category learning.

Acknowledgments

A preliminary report of this work was made at the 2013 meeting of the Cognitive Science Society. EMP was supported by Leverhulme Trust grant RPG-2013-004 and Air Force Office of Scientific Research (AFOSR), Air Force Material Command, USAF, grant FA 8655-13-1-3044. The U.S Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright notation thereon. We would like to thank Lucy Kift and Thom Wilcockson for help with data collection, and Rob Goldstone for his helpful comments.

References

Abdi, H. (2010). Holm's sequential Bonferroni procedure. In N. Salkind (Ed.) "Encyclopedia of Research Design", Thousand Oaks, CA: Sage.

Ashby, G. F. & Ell, S. W. (2002). Single versus multiple systems of category learning: Reply to Nosofsky and Kruschke (2002). Psychonomic Bulletin & Review, 9, 175-180.

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

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, 105,442-481.

Blair, M. & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. Memory & Cognition, 29, 1153-1164.

Blair, M. & Homa, D. (2003). As easy to memorize as they are to classify: The 5-4 categories and the category advantage. Memory & Cognition, 31, 1293-1301.

Chin-Parker, S. & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 216-226.

Delamater, A. R. (1998). Associative mediational processes in the acquired equivalence and distinctiveness of cues. Journal of Experimental Psychology: Animal Behavior Processes, 24, 467-482.

Delamater, A. R. (2012). On the nature of CS and US representations in Pavlovian learning. Learning & Behavior, 40, 1-23.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. Nature, 407, 630-633.

Folstein, J.R., Gauthier, I., & Palmeri, T. J. (2010). Not all spaces stretch alike: How the structure of morphspaces constrains the effect of category learning on shape perception. Vision Sciences Society, 10th annual meeting.

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

Goldstone, R. L. (1994). Influences of Categorization on Perceptual Discrimination. Journal of Experimental Psychology: General, 123, 178-200.

Goldstone, R. L. (1995). Effects of categorization on color perception. Psychological Science, 6, 298-304.

Goldstone, R. L. & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. Journal of Experimental Psychology: General, 1, 116-139.

Goldstone, R. L., Lippa, Y., Shiffrin, R. M. (2001). Altering object representations thought category learning. Cognition, 78, 27-43.

Gureckis, T. M. & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 843). Austin, TX: Cognitive Science Society.

Harnad, S. (Ed.) (1987). Categorical Perception. Cambridge: Cambridge University Press.

Jiang, X., Bradley, E., Rini, R.A., Zeffiro, T., Vanmeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. Neuron 53, 891-903.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical Perception Effects Induced by Category Learning. Journal of Experimental Psychology: Learning, Memory, & Cognition, 24 (3), 732-753.

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

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label feedback hypothesis. Frontiers in Cognition, 3(54).

Markman, A. B. & Ross, B. H. (2003). Category use and category learning. Psychological Bulletin, 129, 592-613.

Mathy, F., Haladjian, H. H., Laurent, E., & Goldstone, R. L. (under review). Similarity-dissimilarity competition in disjunctive classification tasks.

McKinley, S.C., & Nosofsky, R.M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. Journal of Experimental Psychology: Human Perception and Performance, 21,128-148.

McMurray, B., Aslin, R. N., Tanenhaus, M.K., Spivey, M. J., & Subik, D. (in press). Gradient sensitivity to within-category variation in words and syllables. Journal of Experimental Psychology: Human Perception and Performance.

Medin, D. L. & Schwanenflugel, P. J. (1981). Linear Separability in Classification Learning. Journal of Experimental Psychology: Human Learning and Memory, 7, 355-368.

Minda, J.P., & Smith, J.D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. Journal of Experimental Psychology: Learning, Memory & Cognition, 27,775-799.

Minsky, M. L. & Papert, S. A. (1988). Perceptrons. 3rd Edition. MIT Press: Cambridge, MA.

Murphy, G. L. (1991). Parts in object concepts: Experiments with artificial categories. Memory & Cognition, 19, 423-438.

Nakagawa S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology, 15, 1044-1045.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 104-114.

Ozgen, E. & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. Journal of Experimental Psychology: General, 131, 477-493.

Pothos, E. M. & Wills, A. J. (2011). Formal approaches in categorization. Cambridge University Press.

Pothos, E. M. & Chater, N. (2002). A simplicity principle in unsupervised human categorization. Cognitive Science, 26, 303-343.

Pothos, E. M. & Chater, N. (2005). Unsupervised categorization and category learning. Quarterly Journal of Experimental Psychology, 58A, 733-752.

Pothos, E. M. & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the Generalized Context Model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 1062-1080.

Pothos, E. M., Edwards, D. J., & Perlman, A. (2011). Supervised vs. unsupervised categorization: Two sides of the same coin? Quarterly Journal of Experimental Psychology, 64, 1692-1713.

Pothos, E. M., Chater, N., & Stewart, A. J. (2004). Information about the logical structure of a category affects generalization. British Journal of Psychology, 95, 371-386.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. Cognition, 121, 83-100.

Roberson, D. & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. Memory & Cognition, 28, 977-986.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. Journal of Experimental Psychology: General, 3, 369-398.

Roberson, D., Damjanovic, L., & Pilling, M. (2007). Categorical perception of facial expressions: evidence for a "category adjustment" model. Memory & Cognition, 35, 1814-1829.

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. Cognitive Psychology, 50, 378-411.

Rosch, E. & Mervis, C. B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. Cognitive Psychology, 7, 573-605.

Rumelhart, D. E. & McClelland, J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition. Cambridge: MIT Press.

Ruts, W., Storms, G., & Hampton, J. (2004). Linear separability in superordinate natural language concepts. Memory & Cognition, 32, 83-95.

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

Schyns, P. G. & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. Cognition, 69, 243-265.

Schyns, P., Goldstone, R. L., & Thibaut, J. (1997). The development of Features in object concepts. Behavioral and Brain Sciences, 21, 1-54.

Sloutsky, V. M. & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. Journal of E xperimental Psychology: General, 133, 166-188.

Smith, J. D., Murray, M. J. Jr.,& Minda, J. P. (1997). Straight talk about linear separability. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 659-680.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. British Journal of Psychology, 89, 39-57.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 585-593.

Tables

Table 1. Trials to criterion and errors to criterion.

Category structure	Trials to criterion (mean, SD)	Errors (mean, SD)
Width Easy	45.41, 32.69	7.25, 14.68
Width Diff	71.00, 49.66	5.61, 4.92
Height Easy	71.47, 32.55	15.00, 13.40
Height Diff	101.35, 60.26	21.10, 21.07
NLS	102.58, 50.42	34.58, 22.89
Diagonal	154.59, 58.44	33.97, 14.93

Figure captions

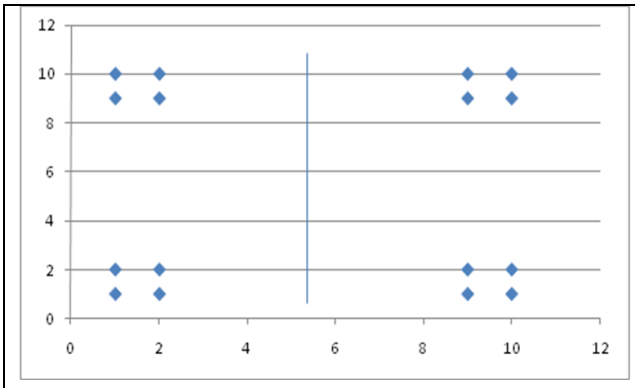
Figure 1. The six category structures employed in the study.

Figure 2. Illustration of an example of the stimuli employed in the current study, reduced in size.

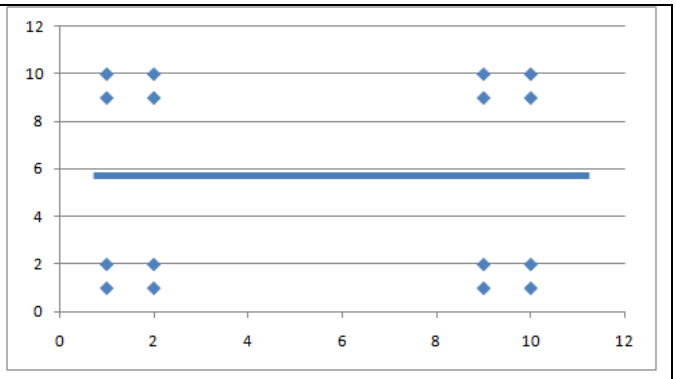
Figure 3. The comparison concerning within and between similarity, prior and post learning, for the different category structures. Error bars represent one standard error of the mean. Stars mark significant results.

Figures

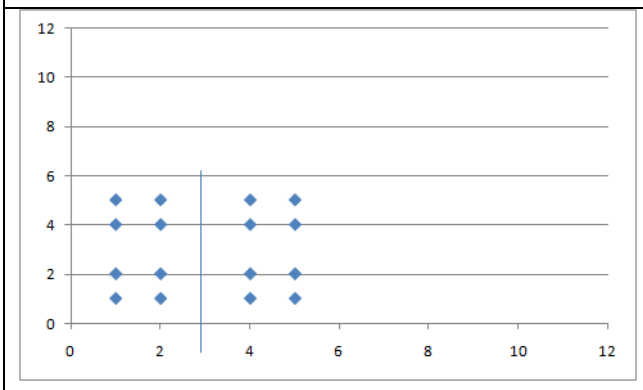
Figure 1



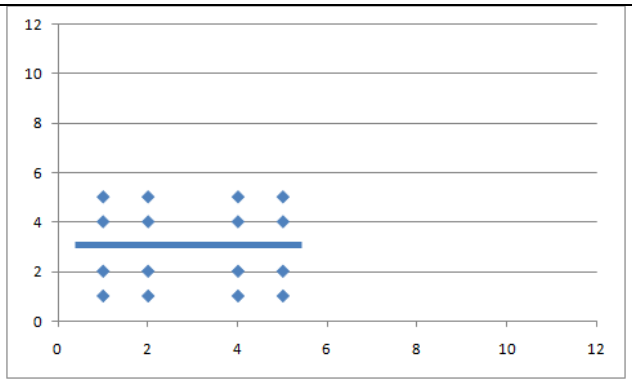
Width Easy



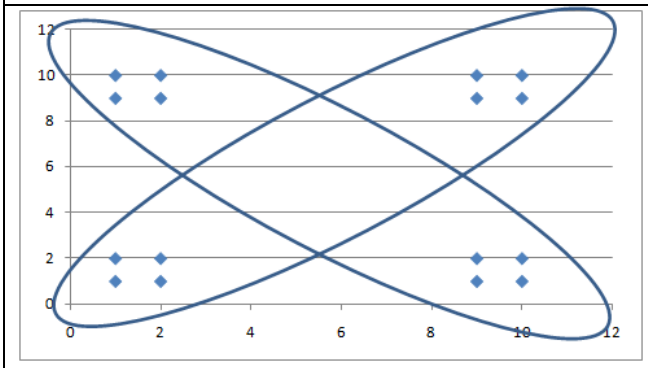
Height Easy



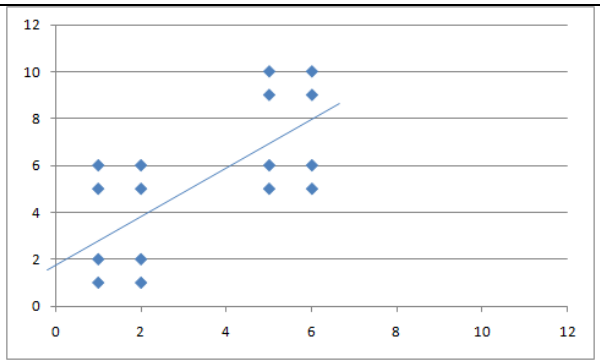
Width Difficult



Height Difficult



NLS



Diagonal

Figure 2

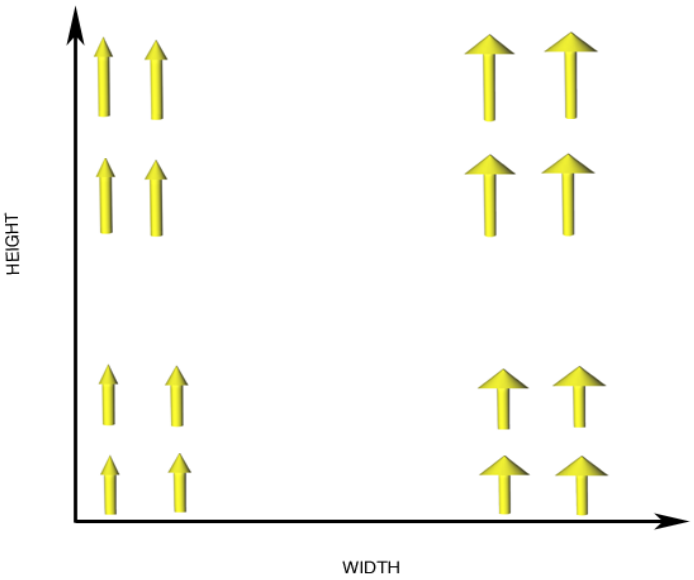


Figure 3.

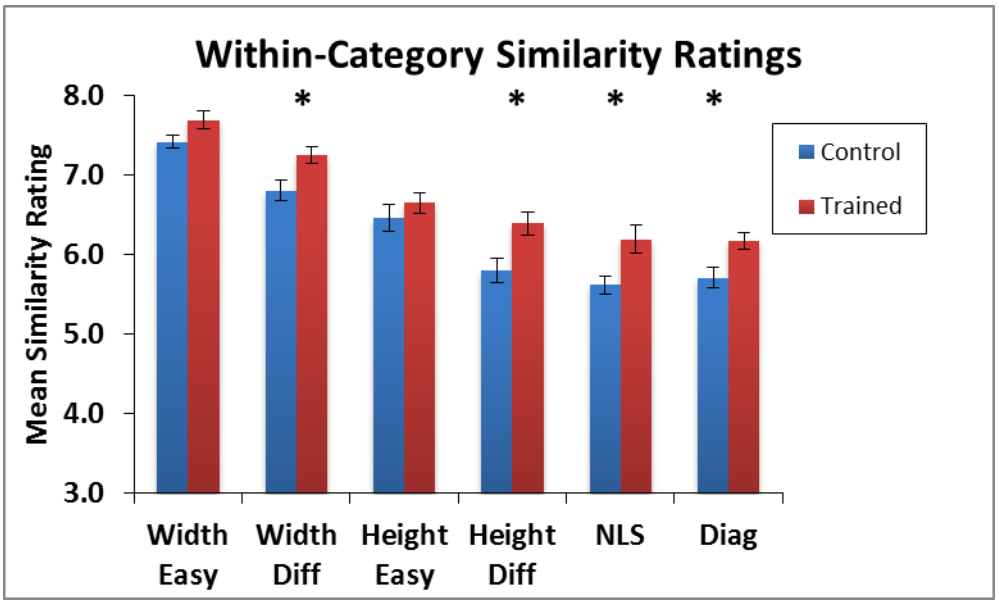


Figure 3a

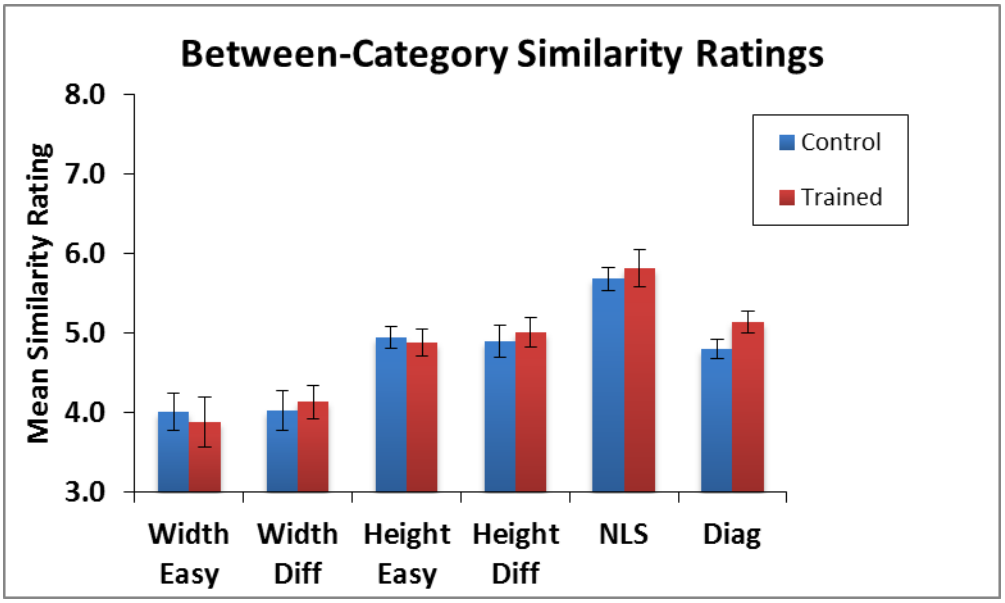


Figure 3b