Pothos, E. M., Edwards, D. J. & Perlman, A. (2011). Supervised versus unsupervised categorization: Two sides of the same coin?. Quarterly Journal of Experimental Psychology, 64(9), pp. 1692-1713. doi: 10.1080/17470218.2011.554990

CITY UNIVERSITY LONDON

City Research Online

Supervised vs. unsupervised categorization: Two sides of the same coin?

Emmanuel M. Pothos[1], Darren J.Edwards[1], & Amotz Perlman[2]

in press QJEP

Running head: supervised vs. unsupervised categorization; **word count**: 10,026
Correspondence/ affiliations: 1: Department of psychology, Swansea University, Swansea SA2 8PP, UK; email: e.m.pothos@swansea.ac.uk, 225088@swansea.ac.uk. 2: Department of Psychology, Ben-Gurion University of the Negev, PO Box 653, Beer Sheva 84105, Israel; email: amotz@bgumail.bgu.ac.il.
Acknowledgements: This research was supported by ESRC grant R000222655 to EMP.

Abstract:

Supervised and unsupervised categorization have been studied in separate research traditions. A handful of studies have attempted to explore a possible convergence between the two. The present research builds on these studies, by comparing the unsupervised categorization results of Pothos et al. (submitted; 2008) with the results from two procedures of supervised categorization. In two experiments, we tested 375 participants with nine different stimulus sets, and examined the relation between ease of learning of a classification, memory for a classification, and spontaneous preference for a classification. After taking into account the role of the number of category labels (clusters) in supervised learning, we found the three variables to be closely associated with each other. Our results provide encouragement for researchers seeking unified theoretical explanations for supervised and unsupervised categorization, but raise a range of challenging theoretical questions.

How similar are supervised processes to unsupervised ones? This debate has been central to many themes in psychology, such as associative learning (e.g., Rescorla & Wagner, 1972; Zwickel & Wills, 2002, 2005), connectionism (e.g., Kohonen, 1982; Rumelhart & McClelland, 1986), and language learning (e.g., Chater & Manning, 2006; Plunkett et al., 1997). In categorization, it concerns the distinction between supervised and unsupervised categorization. The former is about the learning of pre-specified categories. In a laboratory setting, an experimenter may have decided that certain stimuli are in one category, while other stimuli are in a different one. The objective of a participant is to learn which stimuli go to which category, usually through a process of corrective feedback (that is, a participant sees a stimulus, guesses its category membership, and receives feedback as to whether his/her guess was correct or not). In real life, arguably many linguistic categories are taught through a process of supervised categorization. For example, a child can learn that certain objects are oranges and other objects are lemons, by guessing the category membership of a relevant novel exemplar and subsequently receiving corrective feedback from an adult. A key aspect of supervised categorization is that there are no (apparent) limits on the complexity of the classifications which can be taught (e.g., Ashby, Queller, & Berretty, 1999; Maddox et al., 2004; McKinley & Nosofsky, 1995).

Unsupervised categorization concerns the spontaneous impression we often have that a group of stimuli belong to the same category. Such an intuition is most obvious in perceptual grouping, whereby sometimes we have an immediate impression that there are clusters (e.g., see Figure 1; cf. Compton & Logan, 1999). With respect to real concepts, as with the perceptual grouping example of Figure 1, certain real life concepts are more coherent than others. For example, there is little ambiguity regarding membership into the

category of ‘chairs’. However, many naive observers will disagree as to what should be considered (a member of the concept) ‘literature’. In experimental studies of unsupervised categorization, participants are typically presented (sequentially or concurrently) with a set of stimuli and are asked to spontaneously classify them into either a fixed or unlimited number of groups.

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

An issue of central theoretical importance in categorization research is whether a distinction between supervised and unsupervised categorization processes is meaningful. In other words, should we seek to understand and model supervised and unsupervised categorization processes in similar ways, taking into account, of course, the differences between the corresponding tasks (cf. Wills & Pothos, submitted)?

Categorization researchers have mostly pursued the development of either supervised or unsupervised models of categorization (category acquisition in the former, but not the latter, is typically guided by corrective feedback to classification decisions). Hence, the implicit assumption is that supervised and unsupervised categorization processes ought to be understood in separate ways. For example, consider influential supervised categorization models, such as exemplar theory and prototype theory (Hampton, 2007; Minda & Smith, 2000; Nosofsky, 1988; see also, Kurtz, 2007; Vanpaemel & Storms, 2008), which assume that categorization of novel exemplars is driven by their similarity to either the members or the prototypes of the available categories. Similarity is typically computed as a function of distance in a putative psychological space. A key characteristic of such models is that they allow for the possibility that the process of category learning may transform the original psychological space, through the attentional weighting of different dimensions or overall stretching or compression of the space, so as to support the process

of category learning. For example, the attentional salience of a dimension would increase if it is highly diagnostic for a required classification (such transformations of the psychological space are plausibly driven by error correction mechanisms, not readily available in unsupervised categorization; cf. Goldstone, 1994).

Models of unsupervised categorization also often employ a principle of similarity. For example, Pothos and Chater's (2002) simplicity model is based on the idea of Rosch and Mervis (1975) that more obvious classifications should be ones for which within category similarity is maximum and between category similarity is minimum. Specifically, the model assumes that the similarities of all pairs of items that are in the same category should be greater than the similarities of all pairs of items that are between categories. The model predicts that if there are many and correct such 'constraints' then the resulting classification should be more intuitive. An alternative approach is to assume that category formation is driven by a prerogative to maximize the posterior probability of the particular feature combination of their members, given a particular category membership. For example, in the rational model classification of a novel instance depends on $P(k)P(F|k)$, where $P(k)$ is the prior probability of a category and $P(F|k)$ the likelihood of observing the particular combination of object features given the category (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006; cf. Corter & Gluck, 1992). But, Pothos (2007) compared the rational model and the simplicity model and found that the predictions of these models converged across a wide range of stimulus sets.

The above leads us to two intuitions regarding the relevant psychological principles in supervised and unsupervised categorization. First, both supervised and (some) unsupervised categorization models are based on some flavor of similarity. Second, however, in supervised categorization the similarity relations between the categorized

objects can be radically transformed, depending on the particular classification that is taught, but there has been no corresponding evidence in unsupervised classification. Regarding the latter, it only appears that in some cases the spontaneous classification of stimuli takes place on the basis of a single stimulus dimension (Ashby, Queller, & Berretty, 1999; Medin, Wattenmaker, & Hampson, 1987; Milton & Wills, 2004; Pothos & Close, 2008).

SUSTAIN (Love, Medin, & Gureckis, 2004; see also Gureckis & Love, 2003) was the first attempt to provide a single computational framework for both supervised and unsupervised categorization. In SUSTAIN there are separate, but interlinked, components responsible for each type of categorization. Regarding unsupervised categorization, categories emerge for groups of items which are similar to each other. Supervised categorization is supported by a learning mechanism similar to that embodied in current versions of the exemplar theory (e.g., Nosofsky, 1988). The supervised and unsupervised components of SUSTAIN can interact with each other so that, for example, the learning of a classification can be affected by prior perceptions of how intuitive the classification is. Therefore, in SUSTAIN both supervised and unsupervised categorization are supported by a principle of similarity, but (presumably) the exemplar-based learning mechanism in the supervised component allows for greater representational flexibility in supervised categorization. SUSTAIN embodies a particular hypothesis for the relation between supervised and unsupervised categorization: they are related (because they are both based on similarity), but correspond to separate processes. A complementary possibility is that the same basic model serves both supervised and unsupervised categorization. For example, Pothos and Bailey (2009) presented such a proposal for the Generalized Context Model (GCM; Nosofsky, 1988), which is a well-known model of supervised categorization.

In summary, more recent work in categorization modeling has addressed more directly the problem of the relation between supervised and unsupervised categorization. Such work appears to favor a convergence between supervised and unsupervised categorization. However, it is exactly here that is the heart of the problem with this research: despite the excitement associated with the possibility that supervised and unsupervised categorization might be two sides of the same (psychological process) coin, there is still a paucity of relevant experimental results. In an early study, Homa and Cultice (1984) examined whether a set of categories, whose members were all distortions of the corresponding prototypes, could be guessed without corrective feedback. Unsurprisingly, it was found that this was not the case, unless the category members were only minimally distorted from the prototype, but no detailed comparison was provided between performance in the unsupervised and supervised categorization tasks. In fact, there are only two studies directly comparing supervised and unsupervised categorization (Love, 2002; Colreavy and Lewandowsky, 2008). Both these studies are significant in many ways, though their overall conclusions diverge, thus illustrating the need for more empirical research.

Love (2002) employed the classic stimulus sets and classifications from Shepard, Hovland, and Jenkins (1961) and compared performance in a standard supervised categorization task with speed of discovering the underlying, intended classifications without corrective feedback (the latter task is, of course, an unsupervised categorization one; in fact, two kinds of unsupervised tasks were employed). In both cases, the training stimuli were presented repeatedly over several blocks. The test phase was the same for both the supervised and unsupervised categorization tasks and it involved presenting each training stimulus either with its correct category label (which was a stimulus feature over and above the features specified in the Shepard et al., 1961, study) or an incorrect category

label; participants had to decide which version of the stimulus they had encountered in the training phase. Love (2002) identified differences between the supervised and unsupervised tasks, especially in relation to a non-linearly separable classification (a XOR problem): In the unsupervised categorization setting, the linearly separable classifications were acquired more quickly compared to the non-linearly separable one, while under supervised categorization conditions there seems to be no difference (Medin and Schwanenflugel, 1981; note that this conclusion has been challenged more recently, e.g., Blair & Homa, 2001; Ruts, Storms, & Hampton, 2004[1]).

Colreavy and Lewandowsky (2007) employed an unsupervised categorization procedure analogous to that of Love (2002), in that the training stimuli were presented repeatedly and participants were asked to classify them, without receiving any feedback. The main difference between the two studies was in the stimulus sets and corresponding classifications which were employed. While Love (2002) studied a single classification for each of the stimulus sets[2] of Shepard et al. (1961), Colreavy and Lewandowsky (2007) studied a single stimulus set (actually, two stimulus sets, which were meant to be equivalent; results were collapsed across the two stimulus sets), and examined primarily two classifications participants could develop in an unsupervised way from these stimuli. These two classifications corresponded to dividing the available stimuli along one or the other stimulus dimension (the stimuli were two-dimensional). Note that all the classifications Colreavy and Lewandowsky (2007) studied were linearly separable. The learning rate (i.e., the rate of convergence to the classification eventually adopted) was generally slightly faster in the unsupervised case, than in the supervised case. However, the learning rate curves for the participants in the supervised conditions were approximately parallel to the ones for the participants in the unsupervised conditions. Thus, these

investigators concluded that (p.762) “unsupervised categorization …shares many properties of supervised category learning.”

Overall, it is clear that previous empirical research leads to somewhat conflicting intuitions regarding the putative equivalence between supervised and unsupervised categorization. With the present research, we wish to collect additional empirical data and so help address this important issue. Moreover, both Love (2002) and Colreavy and Lewandowsky (2007) primarily intended to study the emergence of knowledge about particular intended classifications, under unsupervised conditions. This objective was achieved by limiting the range of stimulus sets/ classifications and also requiring participants (in all cases) to divide the stimuli into only two categories. While such a procedure was well suited for addressing the particular research objectives of Love (2002) and Colreavy and Lewandowsky (2007), it does raise the question of whether there would be any equivalence between supervised categorization performance and unsupervised categorization performance, under entirely unconstrained grouping conditions for the latter. Indeed, there has been a long tradition of unsupervised categorization work whereby participants are presented (concurrently or sequentially) with a set of stimuli and are asked to divide them into any categories they think are natural or intuitive (such tasks are sometimes called free sorting tasks; Handel & Presser, 1970; Handel & Imai, 1972; Imai & Garner, 1965). It is important to take into account entirely unconstrained unsupervised tasks in the study of the putative equivalence between supervised and unsupervised categorization, as such tasks have been prominent in discussions of key notions in unsupervised categorization (such as category intuitiveness; Pothos & Chater, 2002).

Our research has been organized in two manuscripts. In Pothos et al. (submitted; for an early conference presentation see Pothos et al., 2008) we discuss in detail the

unsupervised categorization task we employed (and attempt to describe the results with various models of unsupervised categorization). The task was a standard free sorting one, that is, a completely unconstrained spontaneous classification task: Participants were concurrently presented with the stimuli in a particular set and were asked to divide them into whichever clusters they thought were natural and intuitive. There were no constraints on the number of clusters participants could employ (or any other constraint) and participants were free to change their classification decisions as many times as they wanted, before settling onto their final classification. A problem with this unsupervised categorization task is that it leads to a large amount of variability in participant responses. Its major advantage is that it closely corresponds to our intuition of spontaneous grouping processes (Pothos and Chater, 2002; see also Compton & Logan, 1999). To our knowledge, this is currently the most extensive study of unsupervised categorization and, therefore, it provides a rich dataset against which to examine possible relations with supervised categorization. A particular advantage of this dataset is that it includes stimulus sets for which the empirically preferred classification does not have two clusters—for some stimulus sets the preferred classification has as many as five clusters.

The present paper describes the results from two carefully matched supervised categorization tasks. For each of the nine stimulus sets employed in Pothos et al. (submitted; 2008), we noted the classification which was produced most frequently by participants—call this classification the 'preferred classification'. Then, in the matched supervised categorization tasks, participants were required to learn these preferred classifications. The supervised learning procedure involved presenting each stimulus in a set to participants one by one, asking them to make a decision regarding how it should be categorized, and providing corrective feedback. To reinforce learning, participants saw the

stimuli with their correct category labels at various intervals throughout the task. The stimuli in the unsupervised categorization task were the same as the ones in the supervised task. The procedure for the unsupervised categorization task involved printing each stimulus individually on a card and presenting all the stimuli in a set concurrently to participants. This procedure has been adopted in other unsupervised categorization research (e.g., Handel & Presser, 1970; Handel & Imai, 1972; Imai & Garner, 1965; more recently: Pothos & Chater, 2002, 2005; Pothos & Close, 2008) and has the advantage that it allows participants to flexibly handle the stimuli and indicate their classifications. Supervised learning requires computer-based presentation of the stimuli, so as to implement the corrective feedback. The appearance, and in particular the size, of the stimuli when shown on cards (in the unsupervised categorization conditions) and on the computer screen (in the supervised categorization conditions reported here) were as carefully equated as possible.

The remaining issue we have to address is how to compare participant performance in the unsupervised and supervised categorization tasks. One way to approach this problem is this: psychologically, in an unsupervised categorization task a researcher can ask whether a particular classification is more intuitive than another one. For example, consider the left panel in Figure 1: in this case, there is an immediate impression that the dots in the diagram can be organized into two clusters. We would expect most naïve observers to agree that this is the most appropriate classification for the dots. By contrast, there is more ambiguity about how the dots in the right panel should be classified. In such a case, different observers will probably classify the stimuli in different (but obviously related) ways. As Pothos et al. (submitted; 2008) have argued, if there is more agreement between participants on how a stimulus set should be classified, then we can consider the corresponding classification as more intuitive. Therefore, one can measure for each stimulus set the frequency with which

the *preferred* classification is produced, with higher frequencies implying that the corresponding classification is more intuitive. Note that Pothos et al. (submitted; 2008) checked that the distribution of classification frequencies was sharply unimodal in the case of all the structured stimulus sets and that various metrics regarding the dispersion of classification frequencies (e.g., entropy) correlated very highly with the frequency of the preferred classification. In other words, the frequency of the preferred classification is a suitable dependent variable to extract from unsupervised categorization results, a conclusion supported by Pothos and Chater (2002, 2005) as well.

The situation in supervised categorization is more straightforward. Researchers consider a taught classification as psychologically more natural if it can be learned quickly (for early studies see Shepard et al., 1961. or Nosofsky, 1984). Despite the manifest intuition of this assumption, a subtlety arises. As noted, the psychological processes involved in learning a categorization are typically assumed to involve some process of transforming the initial stimulus representation into one which is more compatible with the taught classification. Such an assumption is common across a wide range of models, from models specified in terms of psychological spaces (e.g., Minda & Smith, 2000; Nosofsky, 1988) to connectionist models (e.g., Kruschke, 1992; Kurtz, 2007). This raises the question of whether the assumed changes in stimulus representation are short lived or not (cf. Harnad, 1987). Psychologically, in a supervised setting, we would like to consider as more intuitive a classification which is easier to learn *and* one which is more resistant to forgetting (but see e.g., Bjork & Bjork, in press, for a different perspective in an educational setting). We could not find any studies of the latter issue and, so, ease of learning and resistance to forgetting have to be assumed as potentially independent. Thus, in seeking to examine the relation

between unsupervised categorization and supervised categorization, in the present work we considered both ease of learning and memory for a classification.

We can now formulate a particular test of the putative equivalence between supervised and unsupervised categorization: Is it the case that classifications which appear more intuitive in an unsupervised setting are more easy to learn (or better remembered) in a supervised setting? In other words, we are asking whether the psychological process which allows us to appreciate one classification as more obvious than another (cf. Figure 1) is coupled (in the sense that its outcome is consistent) with the psychological process which underwrites our ability to learn how a set of stimuli ought to be mapped to specific category labels. This is a novel research question in categorization, which cannot be answered by the previous related work of Love (2002) and Colreavy and Lewandowsky (2008), as argued above. It has the potential to inform progress with computational models of categorization, where there is currently uncertainly regarding whether supervised and unsupervised categorization should be modeled in a unitary or separable way (Kurtz, 2007; Love et al., 2004; Pothos & Bailey, 2009).

It is possible that the dependent variable in unsupervised categorization task (frequency of the preferred classification) will directly predict the dependent variables in the supervised tasks (speed of learning and memory of these preferred classification). It is also possible that these variables will be related, but only after taking into account other possible variables, which might characterize differences between supervised and unsupervised categorization processes. For example, perhaps structural aspects of the stimulus sets, such as the average within and between category similarity of clusters, may differentially influence performance in supervised and unsupervised categorization tasks. If supervised categorization of linearly separable categories involves identifying optimal

category boundaries, then possibly the discovery of such boundaries is affected primarily by between category similarity, but less so by within category similarity (e.g., Ashby & Maddox, 2005). By contrast, there is strong indication that unsupervised categorization is affected by both within and between category similarity (Pothos & Chater, 2002; Rosch & Mervis, 1975). Another possibility, already alluded to, is that the supervised categorization process makes extensive use of attentional selection of stimulus dimensions, while attentional selection is more limited in unsupervised categorization (e.g., Medin et al., 1987, vs. Nosofsky, 1988). A third possibility concerns the number of clusters (=the number of category labels), since in unsupervised categorization there is no (obvious) reason why classifications with more clusters would be more or less intuitive, while in supervised categorization it may be more difficult to keep track of classifications with more clusters. Some of the above possibilities are easier to translate into an analytical procedure than others and the details are reserved for the results sections of the two experiments in this paper.

Experiment 1

Participants

Participants were 180 undergraduate students at a UK university, who had not taken part in any related experiments. They participated in the study for course credit or a small payment. Experimental design was between participants, so that each participant was tested with only one stimulus set (exactly 20 participants were tested with each stimulus set). Note that the unsupervised categorization results were collected from participants who did not take part in either Experiment 1 or 2 of this study.

Materials

Stimuli were created so as to broadly resemble spiders; the two relevant dimensions of variation were the length of the 'legs' (after the joints) and the length of the central body. We adopted lengths as the relevant dimensions of variations, since this makes it relatively straightforward to assume a Weber fraction (in both cases 8%; Morgan, 2005). For both dimensions, the actual lengths were between 40mm and 80mm. An example of the stimuli is shown in Figure 2. The stimuli were intentionally created to resemble some real-life creature, as a manipulation to increase the coherence of the two dimensions. It was important that the two stimulus dimensions could be perceived together without analytic effort (cf. Milton & Wills, 2004; Pothos & Close, 2008). If analytic effort were required to perceive the two stimulus dimensions together, then it would be less meaningful to talk about the processing of individual stimuli. The stimuli employed in this study were nearly identical (apart from possible minor overall scaling) to those in the unsupervised conditions reported in Pothos et al. (submitted; 2008). As noted, in the unsupervised conditions the stimuli were individually printed and presented to participants as cards. In the presently reported supervised categorization tasks, the stimuli appeared on a computer screen. We took care to ensure that the appearance (and in particular the overall size) of the stimuli in the unsupervised and corresponding supervised tasks was as similar as possible.

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

The key design aspect of this research concerns the range of stimulus sets employed. In this study and in Pothos et al. (submitted; 2008), we employed the same nine different stimulus sets, each having 16 stimuli, which were meant to capture a range of intuitions regarding unsupervised categorization. First, we created several stimulus sets which were variations of a basic two-cluster structure. For example, there was a stimulus set in which there were two well-separated equally-sized clusters, a variation in which the clusters were

closer to each other, another one in which the clusters were of unequal size, and one in which there were two (fairly) well-separated clusters but there were also some ambiguous items in between the two clusters. This emphasis on two-cluster classifications follows the tradition of related work (Colreavy & Lewandowsky, 2007; Love, 2002). Second, we included some stimulus sets which were intended to be consistent with a classification having more than two clusters. For example, we created a stimulus set in which the classification we anticipated would be preferred had three clusters and another which had five clusters. Finally, in some stimulus sets we intended there to be no salient classification at all. Such stimulus sets were included so as to provide a contrast with the more structured stimulus sets. The nine stimulus sets can be referred to as 'two clusters', 'unequal clusters', 'spread out clusters', 'three clusters', 'ambiguous points', 'poor two clusters', 'five clusters', 'random', and 'embedded'. All stimulus sets are shown in Figure 3.

Procedure

We adopted a supervised categorization procedure. The experiment was organized in units, such that each unit consisted of one presentation of all the stimuli with their correct category labels (each stimulus was presented one by one with its correct category label; e.g., "This is a Chomp"), and two presentations of the stimuli without the labels—in the latter case, the participant had to guess the correct label and corrective feedback was provided after each response (as is standard in experiments of supervised categorization). Regarding the presentation of the stimuli with their correct category label prior to the 'guessing' trials with corrective feedback, we thought that if participants had a chance to occasionally review the intended classifications, this might facilitate the learning process. When participants were not required to make a response each stimulus was presented for

1000ms, when participants were required to respond, a stimulus would be shown until a response was made. The learning criterion was to go through all the stimuli in a learning unit without making any errors (the experimenter was able to determine when this happened, because a sound indicated an incorrect response). When a participant managed to do this, the experiment stopped. Otherwise, the participant would be presented again with the stimuli in a unit. A different randomized order of stimulus presentation was employed each time.

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

The classifications taught to participants for each stimulus set are shown in Figure 3. Note that the number of categories varies from two to five. In the cases of the stimulus sets 'two clusters', 'unequal clusters', 'spread out clusters', 'three clusters', 'poor two clusters', and 'five clusters' the taught classifications were the ones preferred by participants in the unsupervised categorization tasks of Pothos et al. (submitted; 2008). Regarding the stimulus sets 'random', 'embedded', and 'ambiguous points', the frequency of the preferred classifications was very low: for each of the three stimulus sets, 3, 2, and 3 respectively. The number of distinct classifications for these stimulus sets were 158, 149, and 160 respectively (the same number of participants was assigned to each stimulus set, the design of the experiment was within participants). Given such very high response variability, it is highly arguable as to whether we should have confidence that there was something special about the classifications which were preferred for these stimulus sets. Rather, it is possible that one classification was simply produced a little bit more often (with a frequency of 3 or 2, instead of a frequency of 1) by chance. Therefore, for the stimulus sets 'random', 'embedded', and 'ambiguous points' the taught classifications in the supervised categorization tasks were *not* the preferred classifications in the corresponding

unsupervised tasks, but rather the classifications predicted as optimal by the simplicity model of unsupervised categorization (Pothos & Chater, 2002). The simplicity model has been shown to accurately predict the preferred classification for a set of stimuli in several studies (Hines, Pothos, & Chater, 2007; Pothos & Chater, 2002, 2005; Pothos & Close, 2008) and hence it seemed a reasonable model for deriving an appropriate classification for use with the supervised categorization tasks for the stimulus sets 'random', 'embedded', and 'ambiguous points'. Note that the classifications predicted as optimal by the simplicity model for the stimulus sets 'random' and 'embedded' were very similar to the preferred ones. Also, we confirmed that the simplicity model correctly predicted the preferred classifications in the cases of the 'two clusters', 'unequal clusters', 'spread out clusters', 'three clusters', 'poor two clusters', and 'five clusters'.

Results

We recorded two dependent variables, the number of learning units required to achieve criterion and the total number of errors before criterion had been achieved (note that each learning unit consisted of a presentation of all the stimuli with their labels and two presentations of the stimuli without the labels—in the second case participants had to guess the correct classification of each stimulus and received corrective feedback). There was a highly significant correlation between the two variables (r=.64, p<.0005). Accordingly, we will restrict the analyses to only one of the variables, the number of learning units required to reach criterion.

Table 1 shows how the number of units differed for the nine stimulus sets we employed. Also, it summarizes the key dependent variable from the unsupervised categorization results of Pothos et al. (submitted; 2008; this is the frequency of the

preferred classification; the unsupervised categorization experiment involved asking 169 naïve participants to spontaneously classify each stimulus set in a within-subjects design). Note, first, that there are differences between the ease of learning of different datasets: $F(8,171)=35.22$, $p<.0005$. This result confirms the expectation from Table 1, that it was much easier to learn the required classification for certain stimulus sets, compared to others.

----------TABLE 1, 2----------

The critical research question concerns a possible relation between the unsupervised and supervised categorization results. From an unsupervised categorization perspective, the higher the frequency of the preferred classification, the more psychologically intuitive this classification should be. From a supervised categorization perspective, the lower the number of units required to reach the learning criterion, the easier (and hence more intuitive) the taught classification should be (cf. Pothos & Bailey, 2009). The objective in the analyses below is to examine whether these two measures of category intuitiveness, from an unsupervised and supervised categorization task, are related or not.

A simple test of a putative association between the measures of category intuitiveness from the unsupervised categorization results of Pothos et al. (submitted; 2008) and the supervised categorization results from the present experiment is a correlation, for each stimulus set, between the frequency of the preferred classification and the number of learning units required to reach criterion. This correlation was low and not significant, although in the right direction ($r=-.47$, $p=.21$). To appreciate the disparity between the supervised and unsupervised categorization results consider, for example, the 'two clusters' and 'five clusters' stimulus sets. In the unsupervised setting, the frequency of the preferred classifications for the two stimulus sets was 31 and 55 respectively. Accordingly, we conclude that the preferred classification in the 'five clusters' stimulus set was more

intuitive and obvious to participants than the one in the 'two clusters' stimulus set. By contrast, the supervised categorization results of this experiment reveal an opposite pattern, so that participants required 4.10 learning units to learn the required classification in the 'two clusters' stimulus set, but 13.45 learning units were required to teach the required classification in the case of the 'five clusters' stimulus set.

The above result highlights a possible sharp difference between supervised and unsupervised categorization. However, as noted in the introduction, the analysis does not take into account a range of factors which may inform the difference between supervised and unsupervised categorization. We therefore computed a number of characteristics for each stimulus set, as a way to converge the results from the unsupervised and supervised categorization tasks. All these characteristics were computed with respect to the taught classifications (as described in the Procedure section; Figure 3). First, we computed an index for the average within category similarity of all clusters for the taught classification for each stimulus set, as the average Euclidean distance of all distances between unique pairs of points in each cluster. Second, in a similar manner we computed an index for the average between category similarity, by taking into account the distance between all unique pairs of points, such that each point was in a different cluster. Third, we noted the number of clusters in each of the taught classifications. This is a major way in which the present study diverges from those of Love (2002) and Colreavy and Lewandowsky (2008), as in these studies participants were restricted into producing (or learning) two cluster classifications. Specifically, an increased number of category labels is likely to affect executive function and working memory resources, both of which might disrupt a process of supervised learning (Maddox et al., 2004).

Finally, we wanted a measure of how participants allocated their attention to the two stimulus dimensions in the supervised categorization task. This is a more involved issue. Typically, in supervised categorization experiments, allocation of attention is inferred by examining the classification of novel stimuli with computational models, such as the GCM (e.g., Nosofsky, 1988). Such models employ attentional parameters, which can inform as to which dimensions were weighted more heavily in the classification of novel stimuli. But in our case, there were no novel stimuli, just learning of the same set of training stimuli. Colreavy and Lewandowsky (2008) could examine attentional allocation directly because the two most common participant classification strategies involved dividing up the available stimuli into two categories either along one dimension of stimulus variation or the other--thus, participants' classifications directly indicated which stimulus dimension they were attending to. This was a strength of the Colreavy and Lewandowsky (2008) study, but it came at the expense of restricting the procedure to only two-cluster classifications and also employing effectively the same stimulus structure (actually, two stimulus structures were employed, but they were equivalent). We employed a much more diverse range of stimulus sets than Colreavy and Lewandowsky (2008), but this came at the expense of being unable to directly infer attentional allocation from empirical results.

Regarding attentional allocation, we adopted a modeling approach. Pothos and Bailey (2009) adapted the GCM for examining a classification as a whole. In brief, the model evaluates the combination of parameter values which best predicts the classification of each stimulus in a set, against all the others. Thus, and as Pothos and Bailey suggested, this leads to an overall measure of how intuitive a particular classification is, according to the GCM. Relevant to the present study, applying the GCM in this way basically leads to an estimate for the attentional weight parameters, that is optimized with respect to the classification of

all the stimuli in a set, relative to their intended classification. We therefore applied Pothos and Bailey's modification of the GCM to each stimulus set in this study. The only input to the model was the coordinates of the stimuli in each set. Parameter optimization was carried out 300 times for each stimulus set, to ensure that the best fit was identified (suitable random starting values for the parameters were employed in each run). Finally, we employed the same parameters as in Pothos and Bailey (2009; these are the standard GCM parameters) and parameter range was as in standard GCM applications, with the exception of the sensitivity parameter, whose upper limit was restricted to 0.2 (see Pothos & Bailey, 2009, for an extensive discussion of why it is important to restrict the sensitivity parameter when applying the GCM in this way).

An alternative approach to the issue of dimensional allocation is to adopt the method of Colreavy and Lewandowsky (2008; see also Pothos & Close, 2008). These investigators employed the simplicity model of Pothos and Chater (2002) by considering which combination of attentional weights (in 10% increments) led to the least codelength for a given classification and a given stimulus set. According to the simplicity model, least codelength means that the corresponding classification should be most intuitive to naïve observers. Therefore, this procedure provides us with a measure of which attentional allocation is spontaneously most intuitive to participants. However, note also that empirical research in attentional allocation in unsupervised categorization has not found evidence for the *fine* attentional changes Colreavy and Lewandowsky (2008) assumed were possible (Medin et al., 1987, is the main study). Therefore, we followed the procedure of Pothos and Close (2008), who suggested that when two-dimensional stimuli are spontaneously categorized, it is either the case that both dimensions are taken into account, or one dimension is entirely ignored. Pothos and Close (2008) discriminated between these two

possibilities by examining the codelength for a particular classification either on the basis of both stimulus dimensions (*xy* configuration) or on the basis of either dimension individually (*x* or *y* configuration). We adopted the same procedure in the present investigation; the input to the simplicity model is the stimulus coordinates of each stimulus set and the output is a codelength value which reflects how intuitive each classification is. We subsequently compared the codelength for the *xy* configuration with the codelength for the *x* and *y* ones. In cases where the *xy* codelength was lower than both the *x* one and *y* one, this scheme predicts a preference for equal attentional allocation to both dimensions and vice versa. In cases where the *xy* codelength was equal to the least codelength between the *x* one and the *y* one, we assumed there might be a preference for attentional allocation to a single dimension (cf. Medin et al., 1987).

Overall, there are five separate variables against which we can assess the putative link between frequency of preferred classification and speed of learning, as shown in Table 2. A condition for any variable to be a mediator in the relationship between a dependent variable (here, assumed to be the frequency of the preferred classification in each stimulus set) and an independent variable (the number of learning units to criterion in the supervised tasks) is that there is a significant association between the putative mediator and the independent variable (Baron & Kenny, 1986). We thus computed the correlations between the five variables in Table 2 and the number of learning units to criterion. Significant correlations were identified only in the case of the number of labels (r=.72, p=.03; positive correlation means that the more the clusters the more difficult the learning) and the simplicity-predicted preference for unidimensional sorting vs. classification based on both dimensions (r=-.84, p=.004; negative correlation means a preference for unidimensional classification was associated with easier learning). For each of these two variables, we next

regressed the number of learning units on the variable and recorded the unstandardized residuals—these residuals provide us with an estimate of the variance in the number of learning units which cannot be accounted for by differences in the variable. Correlating, next, each of these two sets of unstandardized residuals with the frequency of the preferred classification for each stimulus set, significance was attained only for the residuals from the regression with the number of labels ($r=-.81$, $p=.008$). Thus, the number of labels was the only variable which had a mediating role in the association between the speed of learning classifications in supervised tasks and the spontaneous preference for the same classifications in an unsupervised task.

Discussion

The literature in categorization has, to a large extent, been organized around the distinction between supervised and unsupervised categorization. For example, most categorization models are specifically proposed as either models of supervised categorization (e.g., Minda & Smith, 2000; Nosofsky, 1988) or models of unsupervised categorization (e.g., Anderson, 1991; Pothos & Chater, 2002). There is no doubt that the distinction between supervised and unsupervised categorization is a highly intuitive one. However, the present empirical results have failed to support it.

In brief, Experiment 1 was a standard supervised categorization learning paradigm. We asked different participants to learn a particular classification for nine different stimulus sets. A natural dependent variable in this context is the difficulty with which different classifications are learned (cf. Nosofsky, 1984; Shepard et al., 1961). Certain classifications were easier to learn than others. Are these the same classifications which are spontaneously

produced more frequently by participants? We utilized the unsupervised categorization results of Pothos et al. (submitted; 2008) for the same stimulus sets.

The analyses clearly failed to reveal a direct equivalence between supervised and unsupervised categorization results. The question then becomes to examine which particular difference between supervised and unsupervised categorization can account for the corresponding differences in performance. We considered a range of hypotheses, relating to whether supervised and unsupervised categorization might depend differentially on within (or between) category similarity, the number of clusters, the attentional allocation to the two stimulus dimensions as predicted by the GCM (Nosofsky, 1988), and the predicted preference for uni- vs. two-dimensional classification, as predicted by the simplicity model (Pothos & Chater, 2002). We examined the association between frequency of the preferred classification (unsupervised categorization) and speed of learning (supervised categorization), by partialling out variance due to each of these variables in turn (cf. Baron & Kenny, 1986). A significant result was obtained only for the factor enumerating the number of clusters in each of the taught classifications. Specifically, if one eliminates variance due to the number of clusters in the supervised categorization results, classifications which were easier to learn were indeed the ones more likely to be produced spontaneously. Our results therefore show that the aspects of category structure which make a classification easy to learn are the same as the ones which make a classification 'stand out' in a spontaneous categorization setting, as long as one takes into account the differential role of the number of clusters in supervised and unsupervised categorization. Broadly speaking, this observation is consistent with Colreavy and Lewandowsky's (2008) conclusion of compatibility between supervised and unsupervised categorization.

In Experiment 1 we considered one possible hypothesis of how we can decide whether a categorization taught to participants is intuitive or not: if a categorization is easier to learn, then it should be more intuitive. There is an alternative perspective: we can ask whether a particular association between category labels and stimuli is more resistant to forgetting (e.g., Brown, Neath, & Chater, 2008). If a classification for a set of stimuli is better remembered several days after it has been taught, then we should conclude that this classification is more intuitive. Accordingly, we can examine whether category intuitiveness in terms of remembering a taught classification correlates with category intuitiveness in terms of preference in a spontaneous categorization task. Experiment 2 addresses this issue.

Experiment 2

Participants

Participants were 195 undergraduate students at a UK university, who had not taken part in Experiment 1 or any other related experiments. They participated in the study for course credit or a small payment. Experimental design was between participants. Participants were divided between the nine stimulus sets as shown in Table 2.

Materials and Procedure

The materials were identical to those employed in Experiment 1. Experiment 2 consisted of two parts. First, there was a part in which participants had to learn the given classification. This part proceeded in a way analogous to that of Experiment 1, although some modifications were introduced. The learning part was organized in units consisting of a block of 16 trials such that each stimulus appeared with its correct label; in each of these trials, the stimulus and label appeared on the screen until the participant pressed the key with the

corresponding label (this was done so as to reinforce the stimulus—label associations). These 16 trials were followed by five blocks of 16 trials each, such that each stimulus appeared without its correct label, participants had to guess the correct label, and corrective feedback was provided for each response. The learning criterion was analogous to the one employed in Experiment 1: participants had to respond to all 16 stimuli consecutively without making any errors. The training part would stop as soon as participants achieved the learning criterion, otherwise the learning unit (i.e., the 16 presentations of the stimuli with their correct labels followed by the five blocks of 'guessing' trials with corrective feedback) would keep repeating itself.

With the above procedure it is clearly the case that participants would experience a different number of trials, depending on how easy it would be to learn different classifications. As in this case we were interested in the recall of stimulus, category label associations, we included a manipulation which would somewhat equate exposure to the classifications for different stimulus sets, once correct knowledge for these classifications had been attained (arguably, if while learning a participant thought a stimulus was an *A*, but it turned out to be a *B*, then this would not count as an instance of correctly being exposed to the stimulus and its appropriate category label, so that it seemed desirable to somewhat equate for exposure after learning has taken place). Accordingly, after the learning criterion had been achieved, participants saw all the stimuli three more times, in a way that each stimulus with its correct label appeared on the screen, and participants had to press the key with the corresponding label before proceeding to the next stimulus.

Participants were invited to come again to the laboratory seven days later (a deviance of one day was tolerated). To encourage participants to do so, they would not receive any compensation until they came for the second time. Nearly all participants did

attend both experimental sessions. The second experimental session was identical to the learning unit described above (five blocks such that each block consisted of a single presentation of each of the 16 stimuli), but without the presentation of the correct stimulus—category label associations at the beginning. In other words, this was a recall test for the correct label for each stimulus, except for the fact that participants received corrective feedback for their responses. This last manipulation was essential so as not to excessively penalize participants who broadly remembered the classification, but could not remember the particular correspondence between clusters and category labels.

Results

We first consider the dependent variables which are analogous to those in Experiment 1, the number of blocks required to achieve the learning criterion and the errors made before criterion could be achieved (note that we define a learning block in Experiment 2 to correspond to one presentation of the 16 stimuli, so that it differs from the learning unit as defined in Experiment 1). Table 3 shows these results. As before, there was a highly significant correlation between number of blocks and errors (r=.92, p<.0005). It is also interesting to check whether the supervised learning results in Experiment 2 were equivalent to those in Experiment 1, which turned out to be the case (r=.87, p=.002). This result is reassuring, since there were only superficial differences between the training procedure in Experiment 1 and that of Experiment 2.

----------TABLE 3----------

In Experiment 2 there was a novel dependent variable, the number of memory errors in recalling the category label—stimulus associations a week after training (Table 3). We focus the analysis on this variable. Correlating the number of learning blocks with the

number of memory errors shows that classifications which were easier to learn were also the ones which were better remembered a week after learning (r=.97, p<.0005). Regarding the relation between memory retention of a classification and preference in spontaneous classification, we proceeded in the same way as in Experiment 1. First, we correlated the frequency of the preferred classification with the number of memory errors, to find (as before) a non-significant result: r=-.30, p=.43. Second, we considered the five variables in Table 2, as hypotheses regarding the locus of difference between the number of memory errors and the frequency of the preferred classification. Correlating the number of memory errors with these variables, significant correlations were observed only for the variables between category similarity (r=-.68, p=.04; a negative correlation means a lower between category similarity is associated with fewer memory errors), number of labels (r=.82, p=.007), and the simplicity-predicted preference for unidimensional sorting vs. classification based on both dimensions (r=-.70, p=.04). For each of these three variables, we next regressed the number of memory errors on the variable and recorded the unstandardized residuals, which were subsequently correlated with the frequency of the preferred classification for each stimulus set. A significant result was observed only in the case of the residuals from the regression with the number of labels (r=-.74, p=.024), a result which echoes that of Experiment 1. The two panels of Figure 4 provide a graphical illustration of the relation between the results of Experiment 1 and Experiment 2 and the mediating role of category labels.

----------FIGURE 4----------

Discussion

The memory for a particular classification is a dependent variable which has not featured prominently in categorization research. However, it is an important empirical variable, since it informs our insight of what kinds of classifications might be more resistant to forgetting. Presumably, as categorization researchers, we would like to conclude that classifications which are remembered better are ones which are cognitively 'special', in some sense. A classification which is easy to learn is not necessarily the same as a classification which is resistant to forgetting. For example, clusters which are closer to each other may be more prone to forgetting from interference, even if they are straightforward to learn in the first place (cf. Brown et al., 2007).

Equally, learning a categorization sometimes appears to involve particular transformations of the psychological space for the corresponding stimuli. In fact, most models of supervised categorization postulate some mechanism which alters the initial representation of the stimuli into one which is most consistent with the taught categorization (e.g., Kruschke, 1992; Kurtz, 2007; Medin & Schaffer, 1978; Minda & Smith, 2000; Nosofsky, 1988; Rehder & Murphy, 2003). Such an assumption seems to be supported by work on categorical perception (e.g., Goldstone, 1994; Harnad, 1987; Schyns, Goldstone, & Thibaut, 1997), although note there is some controversy as to the exact nature of categorical perception effects (e.g., Goldstone, Lippa, & Shiffrin, 2001; Roberson & Davidoff, 2000). The key issue is that there has been no research as to how long-lived such transformations are. For example, a particular classification may be easy to learn after a fairly radical transformation of psychological space (e.g., involving the projection of all stimuli along a single dimension). However, if this transformation is short-lived, then one would expect that memory for the corresponding classification would likewise decay

quickly. Thus, the ease of learning a classification is in principle independent of the memory for a particular classification.

Despite the above considerations, the present results showed that the memory for a particular taught classification correlated highly with the ease of learning the classification in the first place. The key research question is whether the memory for a particular taught classification could be associated with its salience in an unsupervised categorization task. Our corresponding results closely mirrored the results of Experiment 1. While there was no direct association between memory errors (from Experiment 2) and the frequencies of the preferred classifications, a highly significant correlation was revealed after partialling out variance due to the number of clusters in each of the taught classifications.

General discussion

We have examined two measures of supervised categorization, with nine different stimulus sets, and related the results to spontaneous preference for the taught classifications in an unsupervised categorization task. Each of the different categorization tasks can be seen as providing a different measure of category intuitiveness. The standard supervised categorization task in Experiment 1 can discriminate between classifications which are easy to learn and ones which are more difficult to learn, and it seems uncontroversial to suggest that the former would be psychologically more intuitive compared to the latter (e.g., Kurtz, 2007; Shepard et al., 1961). The supervised categorization task augmented with a recall task (Experiment 2) allowed us to identify the classifications which are more resistant to memory decay and forgetting. Classifications which are better remembered must also be more obvious and intuitive. Finally, the unsupervised categorization procedure employed by Pothos et al. (submitted; 2008) provided a measure of spontaneous preference for a

categorization. More intuitive categorizations would be the ones that are spontaneously produced more frequently.

In comparing the dependent variables from the supervised categorization tasks to the one from the unsupervised task, the first conclusion was that there is *not* a direct association. We therefore next considered a range of factors which could mediate the association between the supervised task dependent variables and unsupervised task one. Each of these factors can be seen as a hypothesis for what is the difference between supervised and unsupervised categorization (with respect to the particular stimulus sets and classifications we employed). Thus, we examined structural characteristics of the stimuli, such as average within cluster similarity and average between cluster similarity, the number of clusters for the classifications in different stimulus sets, and attentional allocation; the latter was computed either on the basis of Pothos and Bailey's (2009) modification to the GCM or Pothos and Chater's (2002) simplicity model. It turned out that excluding variance due to the number of clusters in the supervised categorization performance led to a very close association between the supervised and unsupervised task results. We interpreted this result as showing that, for our particular stimulus sets and classifications, the main difference between the process for unsupervised and supervised categorization relates to the additional difficulty of keeping track of category labels in supervised categorization.

Note, first, that this conclusion goes beyond previous related work, since both Love (2002) and Colreavy and Lewandowsky (2007) employed designs where participants were asked to divide the available stimuli into two categories—thus, it was not possible to examine the potential role of category labels/clusters in the relation between supervised and unsupervised categorization.

To further understand the relation between our conclusion and those of Love (2002) and Colreavy and Lewandowsky (2007) it is worth considering in detail the key differences in design. Both these studies employed an unsupervised categorization procedure which was effectively one of category discovery without feedback: Participants were presented with the same stimuli over several blocks, so that eventually their classifications converged to a fairly stable pattern. In Love's (2002) case, this pattern was assessed against an underlying target classification (which was either discovered or not) and in Colreavy and Lewandowsky's (2007) case, participant response patterns were examined in relation to two main classification strategies (each strategy was characterized in terms of the stimulus dimension along which a category boundary could be defined). However, in both studies, the unsupervised categorization procedure imposed restrictions to the categorizations participants could produce. In Love's case, for each stimulus set there was only a single intended classification which was either discovered or not. In Colreavy and Lewandowsky's case, participants' performance was examined in terms of primarily two possible classifications. But, in both cases, these procedures fall short in relation to the typical variability in classification performance under entirely unsupervised conditions (e.g., Pothos & Chater, 2002). For example, Pothos et al. (2008, submitted) recorded over 1000 distinct classification when 169 participants each spontaneously classified the stimuli in Figure 3.

So, we think the main strength of the present study is that the dependent variable from the unsupervised categorization tasks is more immediately related to unsupervised categorization performance and category intuitiveness. Of course, the main strength of Love's (2002) and Colreavy and Lewandowsky's (2007) studies is that it was possible to examine in more detail the development of classification strategies and, also, to have slightly more control over the studied classifications. For example, Love (2002) was able to

examine non-linearly separable classifications with his unsupervised procedure, something which we do not believe is possible under entirely unconstrained classification procedures. Our research and the research of Love (2002) and Colreavy and Lewandowsky (2007) have complementary strengths.

Moreover, the above discussion immediately allows us to understand the differences in the overall conclusions from our research and that of Love (2002). Love concluded that it is not possible to equate supervised and unsupervised categorization performance, contrary to our own conclusion (though note that in later work he sought a more integrated approach to understanding supervised and unsupervised categorization; Love et al., 2004). However, some of the differences he identified between unsupervised and supervised categorization related to non-linearly separable category structures, while it was not possible to examine such category structures with our unsupervised categorization procedure. Also, our overall conclusion strongly resonates with that of Colreavy and Lewandowsky (2007; cf. Zwickel &Wills, 2005), which is as expected since in both cases the classifications employed were linearly separable.

Our results indicate that the psychological process which allows us to consider certain classifications as more obvious than others must be intimately related, or be partly equivalent, to the psychological process which enables the learning of a required classification. If such a conclusion proves to be general, it would have important implications for the development of models of categorization. Currently, nearly all categorization models are specifically proposed either as models of supervised categorization (e.g., Minda & Smith, 2000; Nosofsky, 1988) or models of unsupervised categorization (e.g., Anderson, 1991; Pothos & Chater, 2002). Some researchers have sought to modify models of supervised categorization so that they can function as models of unsupervised categorization (e.g.,

Kurtz, 2007, Pothos & Bailey, 2002; Zwickel & Wills, 2005). Also, there have been attempts to integrate a component for supervised categorization and one for unsupervised categorization within the same formalism (e.g., Gureckis & Love, 2003; Love et al., 2004). The results in this paper inform our understanding of these approaches.

At the same time, it seems clear that supervised categorization processes must go beyond unsupervised categorization processes, at least under some circumstances. Supervised learning can allow a naïve observer to learn classifications which would never be produced spontaneously (e.g., McKinley & Nosofsky, 1995; Maddox et al., 2004). The learning of such complex classifications appears to involve radical transformations of psychological space, so that the similarity structure of the stimuli evolves to become more consistent with the taught classifications. Such transformations can include fine attentional modulation (Nosofsky, 1984, 1988), changes in the *grain* of the similarity space (Nosofsky, 1984, 1988), or even the creation of novel features (Schyns et al., 1997; Goldstone, 2000). By contrast, in unsupervised categorization, there has been evidence only for a possible 'crude' attentional selection process, whereby a stimulus dimension may be spontaneously ignored if it does not appear to add to the overall intuitiveness of a classification (Ashby, Queller, & Berretty, 1999; Pothos & Chater, 2005; Pothos & Close, 2008). This issue relates to the well-known issue of unidimensional biases in early studies of spontaneous grouping (Milton & Wills, 2004; Regehr & Brooks, 1995; Medin, Wattenmaker, & Hampson, 1987; cf. Murphy, 2004).

So, it seems reasonable to assume that supervised categorization can involve more complex processes, relating to the transformations of representations, compared to unsupervised categorization. One possibility is that there is a categorization system subserving both unsupervised and supervised (under some circumstances) problems.

However, in the case of complex categorization problems, supervised categorization can draw on augmented mechanisms for representational flexibility, not available to unsupervised categorization. In other words, simply put, supervised categorization is all unsupervised categorization is and a little more. This would be an ambitious and exciting proposal for understanding the complete range of human categorization abilities, but its full exploration would require considerable additional work.

More specific further empirical questions concern the precise nature of the interaction between biases from supervised categorization and unsupervised categorization. Our results show that, in the case of learning naturalistic (linearly separable, fairly intuitive) classifications, it is generally the case that more intuitive classifications are easier to learn (if variance due to the number of category labels is excluded). In other words, the unsupervised categorization biases have a major influence on supervised categorization. But is it the case that unsupervised categorization biases influence learning regardless of the complexity of a learned classification? Or is there a point at which, if the taught classification is too complex, the unsupervised component is simply suppressed? It is worth pointing out here that even though there have been several demonstrations of naïve observers learning complex classifications (e.g., McKinley & Nosofsky, 1995; Minda & Smith, 2000; Nosofsky, 1988) some researchers have questioned whether performance in such tasks reflects categorization behavior as such, as opposed, for example, to memorization of which category labels go with which stimuli (Blair & Homa, 2003).

In sum, categorization researchers have made impressive progress in understanding supervised categorization (Hampton, 2007; Kurtz, 2007; Minda & Smith, 2000; Nosofsky, 1988; Vanpaemel & Storms, 2008) and, more recently, unsupervised categorization (Anderson, 1991; Pothos & Chater, 2002; Sanborn et al., 2006). However, there has been

very limited both theoretical (Love et al., 2004; Pothos & Bailey, 2009) and empirical (Colreavy & Lewandowsky, 2008; Love, 2002) work on the relation between supervised and unsupervised categorization. This is an important obstacle before a more complete understanding of human categorization processes can be achieved. Our results extend the research of Colreavy and Lewandowsky (2008) and Love (2002) on the putative equivalence between supervised and unsupervised categorization and illustrate the range of the corresponding theoretical challenges.

Footnotes

Footnote 1. The presence of the stimulus characteristic corresponding to the intended category label in Love's (2002) stimuli would have led to a stimulus dimension which would enable a perfectly linearly separable classification even in the XOR example of Shepard et al. (1961). It is not clear whether the presence of such a dimension affected performance with the XOR classification in Love's experiments.

Footnote 2. Shepard et al. (1961) employed a single stimulus set and six different classifications for this stimulus set. But, as Love (2002) augmented the stimuli with an additional feature indicating their intended classification, it is simpler to just talk about six separate stimulus sets in the case of that study.

References

Anderson, J. R. (1991). The Adaptive Nature of Human Categorization. *Psychological Review, 98*, 409-429.

Ashby, F. G & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149-178.

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

Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

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.

Bjork, E. L., & Bjork, R. A. (in press). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society*. New York: Worth Publishers.

Brown, G.D.A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576.

Chater, N. & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*, 335-344.

Colreavy, E., & Lewandowsky, S. (2008). Strategy development and learning differences in supervised and unsupervised categorization. *Memory & Cognition, 36*, 762-775.

Compton, B. J. & Logan, G. D. (1999). Judgments of perceptual groups: Reliability and sensitivity to stimulus transformation. *Perception Psychophysics, 61*, 1320-1335.

Corter, J. E. & Gluck, M. A. (1992). Explaining Basic Categories: Feature Predictability and Information. *Psychological Bulletin, 2*, 291-303.

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

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 86-112.

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

Gureckis, T.M. and Love, B.C. (2003). Towards a Unified Account of Supervised and Unsupervised Learning. *Journal of Experimental and Theoretical Artificial Intelligence*, *15*, 1-24.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*, 355-384.

Handel, S. & Preusser, D. (1970). The free classification of hierarchically and categorically related stimuli. *Journal of Verbal Learning and Verbal Behavior, 9*, 222-231.

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

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

Hines, P., Pothos, E. M., & Chater, N. (2007). A non-parametric approach to simplicity clustering. *Applied Artificial Intelligence*, *21*, 729-752.

Homa, D. & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 83-94.

Imai, S. & Garner, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology, 69*, 596-608.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59-69.

Kruschke, J. K. (1992) ACLOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review, 14*, 560-576.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829-835.

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

Maddox, W. T., Filoteo, J. V., Hejl, K. D., Ing, A. D. (2004) Category number impacts rule-based but not information-integration category learning: further evidence for dissociable category learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 227-235.

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.

Medin, D. L., & Shaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Medin, D. L. & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 75*, 355-368.

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

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.

Minda, J. P., & Smith, J. D. (2000). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 775-799.

Morgan, M. J. (2005). The visual computation of 2-D area by human observers. *Vision Research, 45*, 2564-2570.

Murphy, G. L. (2004). *The big book of concepts*. MIT Press: Cambridge, USA.

Murphy, G. L. & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92*, 289-316.

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 54-65.

Plunkett, K., Karmiloff-Smith, A., Bates, E., & Elman, J. L. (1997) Connectionism and developmental psychology. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 38, 53-80.

Pothos, E. M. (2007). Occam and Bayes in predicting category intuitiveness. *Artificial Intelligence Review, 28*, 257-274.

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

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., Perlman, A., Edwards, D. J., Gureckis, T. M., Hines, P. M., & Chater, N. (2008). Modeling category intuitiveness. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, LEA: Mahwah, NJ.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., & Hines, P. (submitted). Measuring category intuitiveness in unconstrained sorting tasks.

Regehr, G. & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 347-363.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin and Review , 10*, 759–784.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research* (pp. 64 - 99). New York: Appleton-Century-Crofts.

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

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. & the PDP Research Group (Eds.) (1986). *Parallel Distributed Processing: Vol. 1, Foundations*. Cambridge, Mass: M.I.T. Press

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

Sanborn, A. N., Griffiths, T. L., & Navarro, D. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Schyns, P. G. (1991). A Modular Neural Network Model of Concept Acquisition. *Cognitive Science, 15,* 461-508.

Schyns, P., Goldstone, R. L., & Thibaut, J. (1997). The Development of Features in Object Concepts. *Behavioral and Brain Sciences, 21*, 1-54.

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

Vanpaemel, W. & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15*, 732-749.

Wills, A. J. & Pothos, E. M. (submitted). How can we make progress in the formal modeling of categorization.

Zwickel, J. & Wills, A. J. (2002). Is competitive learning an adequate account of free classification? *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, pp. 982-987.

Zwickel, J. & Wills, A.J. (2005). Integrating associative models of supervised and unsupervised categorization. In A. J. Wills (Ed). *New Directions in Human Associative Learning*. London: LEA.

Tables

Table 1. A summary of the unsupervised categorization results of Pothos et al. (2008) and the supervised categorization results obtained in Experiment 1.

Stimulus set	Frequency of most preferred[1]	Mean number of units[2]	Range[3]	Standard deviation[4]
Two clusters	32	4.10	2—10	2.22
Unequal clusters	33	4.15	2—11	2.28
Spread out clusters	8	7.40	2—17	4.14
Three clusters	55	9.30	3—21	5.29
Ambiguous two clusters	3	14.45	3—27	8.17
Poor two clusters	17	9.65	3—24	5.76
Five clusters	60	13.45	4—28	7.42
Random	3	25.40	12—33	5.14
Embedded	2	22	9—35	6.91

Notes: [1]The frequency with which the preferred classification was produced, in a sample of 169 participants. [2]The mean number of learning units required to reach the learning criterion. [3]The lowest and highest number of learning units required to reach criterion. [4]The standard deviation associated with the number of learning units required to reach criterion.

Table 2. Variables corresponding to hypotheses about the difference between supervised and unsupervised categorization processes.

	Within category similarity[1]	Between category similarity[2]	number of clusters	GCM attentional weight for dominant dimension[3]	simplicity codelength *x*[4]	simplicity codelength *y*	simplicity codelength *xy*	simplicity-predicted preference for unidimensional sorting (=1)
Two clusters	1.58	10.33	2	0.5	50.2	50.2	50.2	1
Unequal clusters	1.72	10.39	2	0.515	50	50	50	1
Spread out clusters	2.65	8.89	2	0.5	50.2	50.2	50.2	1
Three clusters	1.37	8.58	3	0.57	58.9	76.3	58.9	1
Ambiguous points	2.85	8.04	2	0.812	59.2	63.7	58.7	0
Poor two clusters	1.84	4.75	2	0.781	52	71.1	55.9	1
Five clusters	1.14	7.83	5	0.5	83.8	83.8	74.9	0
Random	2.7	6.66	4	0.527	85.5	85.9	71.2	0
Embedded	2.03	6.39	5	0.518	85.1	81.9	72.1	0

Notes: [1]The average Euclidean distance between all pairs of stimuli in the same cluster. [2]The average Euclidean distance between all pairs of stimuli in different clusters. [3]We fitted the GCM as in Pothos and Bailey (2009). [4]The simplicity model codelengths were computed as in Pothos and Chater (2002) for the taught classification for each stimulus set, on the basis of one stimulus dimension (*x*), the other (*y*), or both (*xy*).

Table 3. The supervised categorization results obtained in Experiment 2.

Stimulus set	Participants	Mean number of blocks[1]	Range[2]	Standard deviation[3]	Memory errors[4]
Two clusters	25	1.36	1—3	0.64	1.21
Unequal clusters	27	2.04	1—8	1.58	1.28
Spread out clusters	32	2.22	1—11	1.93	2.67
Three clusters	13	9.23	2—37	9.33	5.33
Ambiguous two clusters	21	3.57	1—18	3.98	3.59
Poor two clusters	18	6.39	1—17	4.25	5.00
Five clusters	19	10.42	3—31	7.42	6.47
Random	20	18.15	3—47	10.99	11.33
Embedded	20	24.95	6—60	15.05	11.65

Notes: [1]The mean number of learning blocks required to reach the learning criterion. [2]The lowest and highest number of learning units required to reach criterion. [3]The associated standard deviation. [4]The number of errors in reproducing the category label—stimulus associations a week later.

Figure captions

Figure 1. Assume that the diagrams correspond to some putative psychological space and that each dot corresponds to an instance in our experience. There is an immediate impression that there are two clusters on the left panel, but this is not so for the right panel.

Figure2. An example of the stimuli used. The stimuli varied in terms of the length of the legs after the joint and the length of the central body.

Figure 3. A schematic representation of the nine stimulus sets employed in this research. Each point in each stimulus set is indexed by a number from 0 to 15. The curves show the classifications taught to participants in each case.

Figure 4. The top panel shows frequency of preferred classification (from Pothos et al., submitted; 2008), number of learning units (Experiment 1), and number of memory errors (Experiment 2) for the nine stimulus sets. Regarding the bottom panel, we first computed the residuals when regressing learning units on category labels (Experiment 1) and memory errors on category labels (Experiment 2). We then scaled these residuals to correspond as closely as possible to the frequency of preferred classifications, with linear regressions between the respective pairs of variables.

Figures

Figure 1.

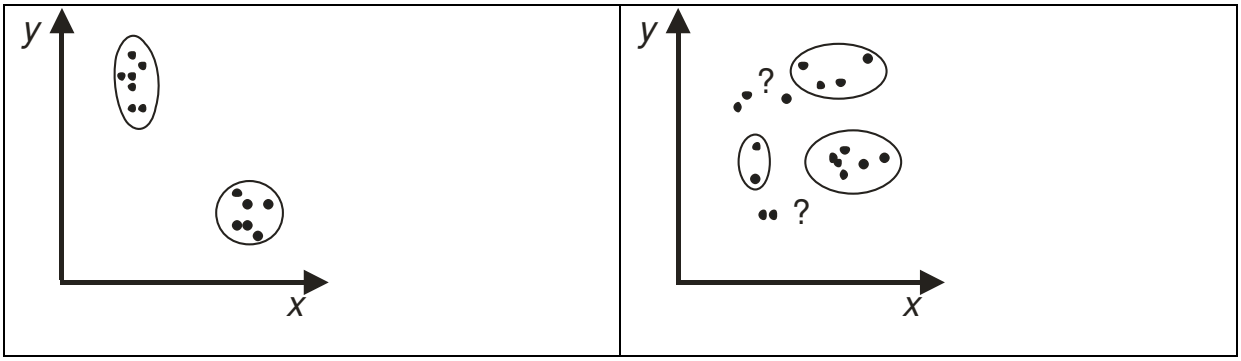


Figure 2

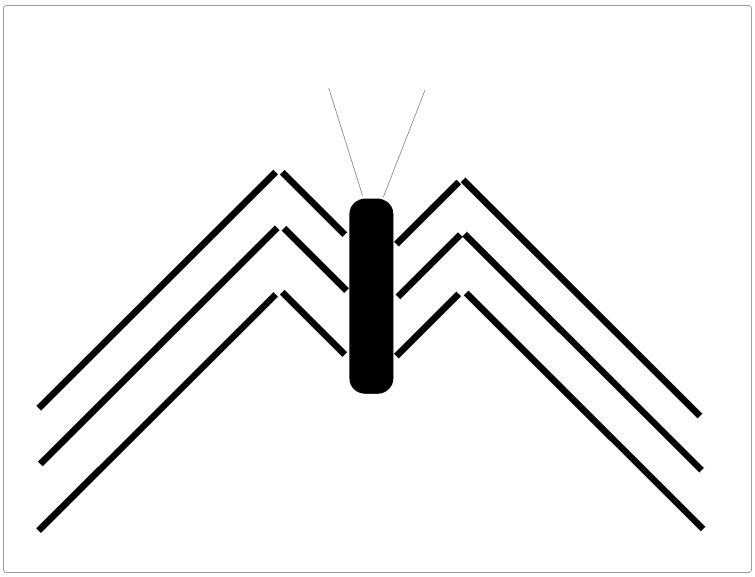


Figure 3

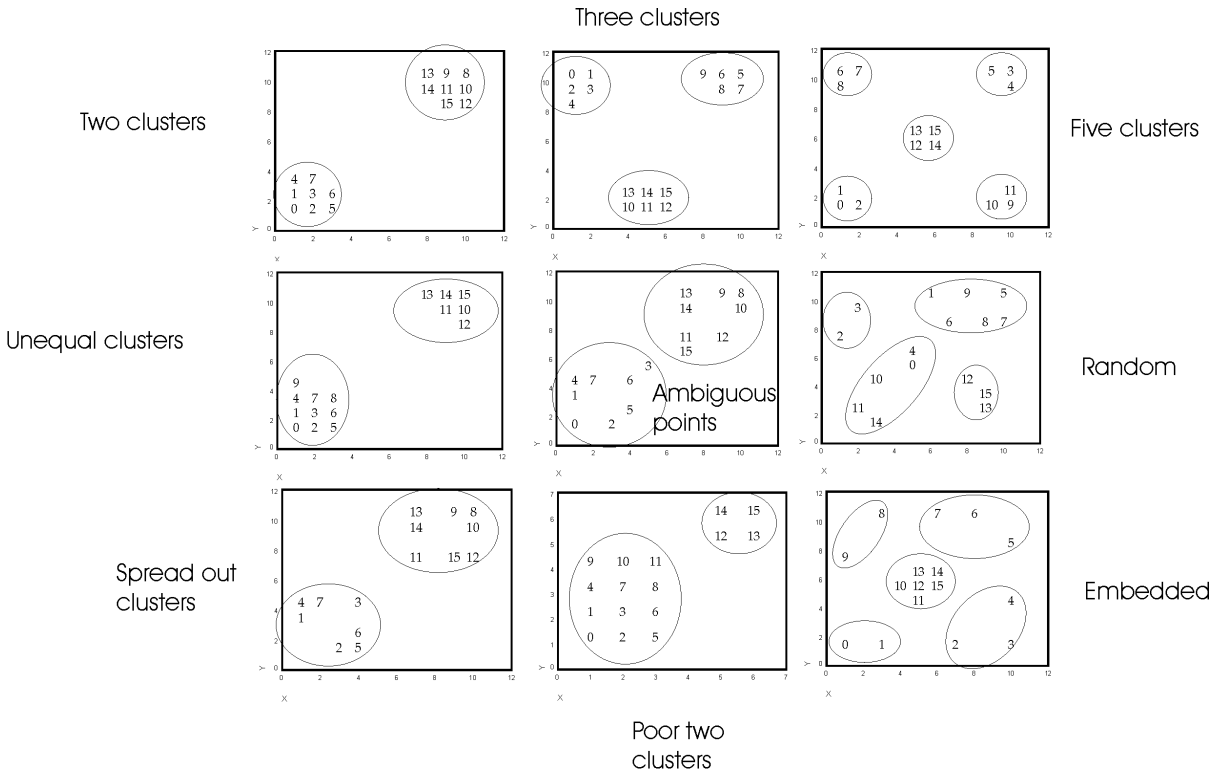
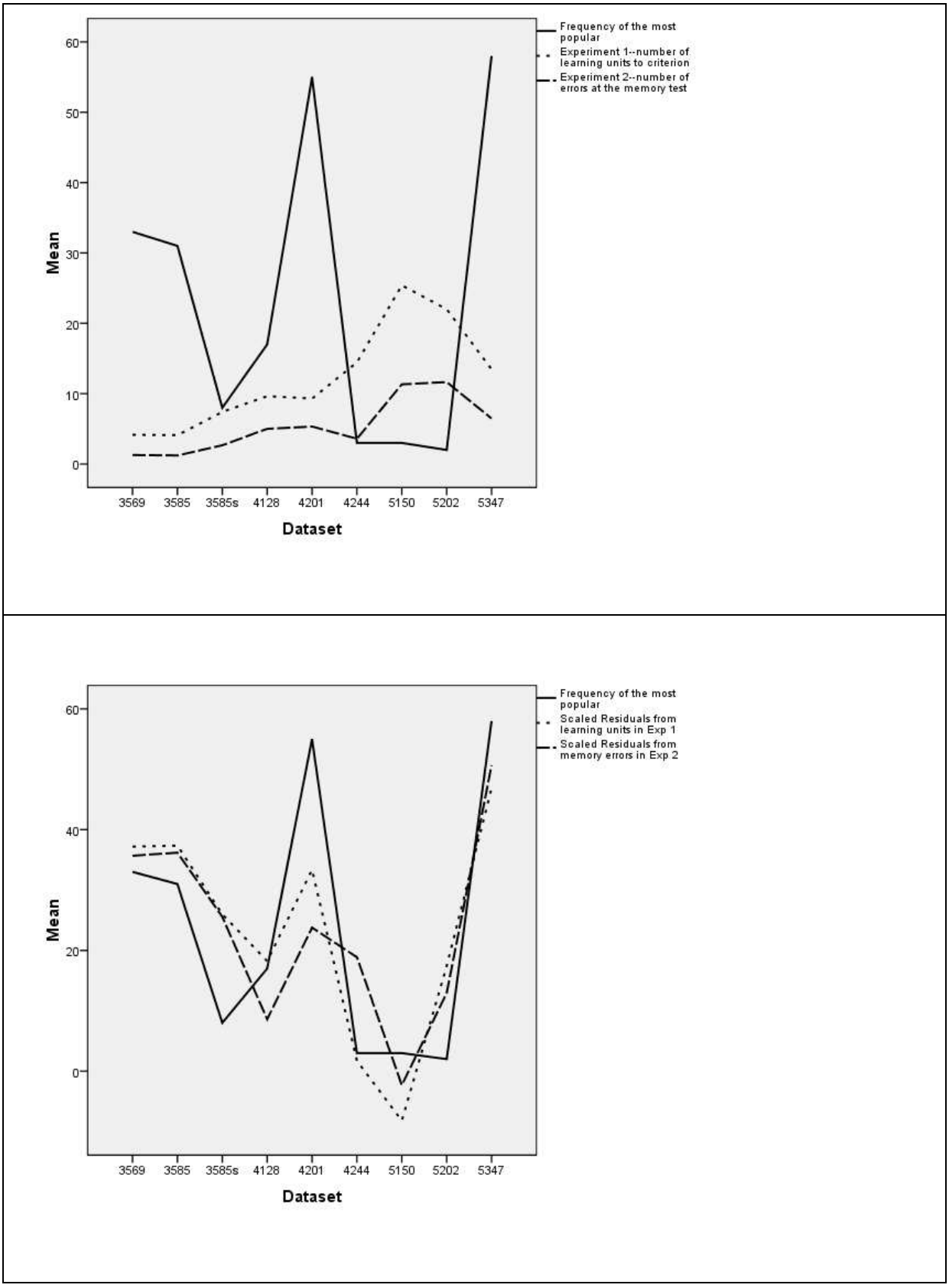

Three clusters
Two clusters
Five clusters
Unequal clusters
Random
Ambiguous points
Spread out clusters
Embedded
Poor two clusters

Figure 4


Frequency of the most popular
Experiment 1--number of learning units to criterion
Experiment 2--number of errors at the memory test
Mean
Dataset
3569
3585
3585s
4128
4201
4244
5150
5202
5347
Frequency of the most popular
Scaled Residuals from learning units in Exp 1
Scaled Residuals from memory errors in Exp 2
Mean
Dataset