Instance Memorization and Category Influence: Challenging the Evidence for Multiple Systems in Category Learning

Mark K. Johansen*
JohansenM@cardiff.ac.uk
Cardiff University, School of Psychology
Tower Building, Park Place, Cardiff, United Kingdom CF10 3AT

Nathalie Fouquet
N.C.Fouquet@swansea.ac.uk
Swansea University, College of Human and Health Sciences
Glyndwr Building, Singleton Park
Swansea, United Kingdom SA2 8PP

Justin Savage
SavageJC@cardiff.ac.uk
Cardiff University, School of Psychology
Tower Building, Park Place, Cardiff, United Kingdom CF10 3AT

David R. Shanks
University College London, Division of Psychology and Language Sciences
26 Bedford Way, London, United KingdomWC1H 0AP

*Corresponding author

Abstract

A class of dual-system theories of categorization assumes a categorization system based on actively-formed prototypes in addition to a separate instance memory system. It has been suggested that, because they have used poorly differentiated category structures (such as the influential '5-4' structure), studies supporting the alternative exemplar theory reveal little about the properties of the categorization system. Dual-system theories assume that the instance memory system only influences categorization behaviour via similarity to single isolated instances, without generalization across instances. However, we present the results of two experiments employing the 5-4 structure to argue against this. Experiment 1 contrasted learning in the standard 5-4 structure with learning in an even more poorly differentiated 5-4 structure. In Experiment 2 participants memorized the 5-4 structure based on a 5 minute simultaneous presentation of all nine category instances. Both experiments revealed category influences as reflected by differences in instance learnability and generalization, at variance with the dual-system prediction. These results have implications for the exemplars versus prototypes debate and the nature of human categorization mechanisms.

Keywords: category learning, exemplars, prototypes, categorization, representation

Instance Memorization and Category Influence: Challenging the Evidence for Multiple Systems in Category Learning

The ability to categorize objects and events and to form conceptual representations of those categories is a core cognitive capacity. Research on categorization in both psychology and neuroscience has been heavily influenced in recent years by attempts to evaluate various 'dichotomy' frameworks such as the distinctions between implicit/explicit learning and rule-based/similarity-based category learning (Ashby & Maddox, 2005; Seger & Miller, 2010). Particularly prominent amongst these distinctions is that between exemplar versus prototype representation of categories, and debate over this distinction continues to be a core research issue in categorization and learning. While at least initially exemplar and prototype theories seem conceptually well differentiated, specifying what the debate is truly about has been more difficult, especially in the broadening context of multisystem models where for some the abstraction-based subsystem uses rules rather than prototypes (e.g., ATRIUM, Erickson & Kruschke, 1998).

Exemplar theory (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984; 1986) assumes that people represent categories during learning by storing the experienced instances of the categories in memory. This representation can still reflect some abstraction—for instance, selective attention can moderate similarity to make within-category instances more similar than between-category instances (Kruschke, 1992; Nosofsky, 1986)—but the representation is relatively unintegrated and unabstracted in that the individual exemplars are assumed to be separately encoded. The probability of a new instance being classified into a particular category is the sum of its similarities to the stored category instances relative to the corresponding summed similarities for other categories. So the greater the overall similarity to the instances of the category, the more likely a new case is to be classified into that

category. In brief, memory for instances is crucial and the formation of new abstractions is, at most, limited in the context of adaptive categorization.

In contrast, prototype theory (e.g., Blair & Homa, 2001; Homa, Rhoads & Chambliss, 1979; Homa, Sterling & Trepel, 1981; Posner & Keele, 1968; Smith & Minda, 1998; 2000) assumes that categories are represented during learning by an actively integrated average or central tendency of the observed instances, the category prototype. New instances are categorized at test based on their similarity to various category prototypes, where the probability of a particular categorization is proportional to the similarity of the instance to the category prototype relative to other category prototypes. Hence the greater the overall similarity to a category prototype, the more likely a new case is to be classified into that category. In brief, long-term memory for specific instances is not required while formation of new abstractions, that is actively integrated prototypes, is crucial for adaptive categorization.

Importantly, the exemplars versus prototypes debate is not about whether prototypicality effects occur. Virtually everyone now agrees that a fundamental property of most real world categories is that some instances are better, more typical, members than other instances. Moreover, exemplar models would not be nearly as successful as they have been if they could not provide some account of such effects (e.g., Nosofsky & Kruschke, 1992; but also see the recent debate about potential prediction differences in typicality gradients in the dot-distortion paradigm: Homa, Hout, Milliken, & Millikin, 2011; Smith, 2002, 2005; Zaki & Nosofsky, 2004, 2007). Nor does anyone dispute that participants can remember particular instances of categories or that these can influence categorization as demonstrated by the large body of evidence for exemplar effects, which most notably occur even in the presence of explicit and perfectly diagnostic rules (e.g., Allen & Brooks, 1991; Hahn, Prat-Sala, Pothos & Brumby, 2010). In particular, it is worth emphasizing in advance that prototype theory does not preclude the influence of instance memory on categorization or even categorization

behaviour based solely on instance memory in some cases, as we describe below. Needless to say, these complicating facts emphasize the importance of being clear about the key aspects of the debate.

To be useful, categories have to add something; category membership knowledge needs to improve adaptive functionality. Knowing that a particular animal is in the category dog is not very useful if it does not help predict unseen properties. This "category influence" can take many forms, for instance, it can influence the learnability of category instances as shown in the experiments reported below. And there are a variety of perspectives in the literature about what categories do; for example, the perspective embodied in Anderson's (1990; 1991) Rational Model is that categories are for optimized feature inference whereas the emphasis in Pothos and Chater's (2002) simplicity model is on representing information efficiently. But most fundamentally category influence involves integration of information from across the experienced instances of the categories in a way that improves prediction and control. The particular entity nearby is apparently a dog. Dogs tend to be territorial and protective as indicated by growling, in which case approach is inadvisable, but a wagging tail can indicate they are friendly and that approach is safe. Not only do exemplar and prototype theories make different predictions about how category integration occurs, but this difference in proposed mechanisms for category influence is the fundamental difference between them.

The crux of the difference between prototype and exemplar theories is whether or not category influence is the result of a separate, abstraction-generating, cognitive system distinct from instance memory (Blair & Homa, 2001, 2003; Nosofsky, 2000; Smith & Minda, 2000, 2002). Specifically, does category integration as measured particularly by prototypicality effects occur in a separate system from instance memory via an active process resulting in an abstracted category prototype? Or is category integration fundamentally a passive process

resulting from similarity to the category instances stored in a single memory system—possibly moderated by selective attention, forgetting, and interference?

Once prototypicality and exemplar effects in categorization are both acknowledged to occur, distinguishing prototype from exemplar representation can be conceptually, not to mention pragmatically, quite difficult. However, one useful difference between the theories follows from the conceptual distinction between a single system mediating categorization and instance memory (exemplar theory) versus separate systems for categorization and instance memory (prototype theory) and leads to the following crucial question: Does more than one category instance influence a given categorization decision? Or does memory for instances influence categorization only via a single nearest exemplar? Both Smith (2005) and Nosofsky (2000) have specified this as a key issue while arguing from opposite theoretical positions. In particular, Smith (2005, p. 47) specifies the theoretical distinction in terms of "...whether exemplar generalization in memory and categorization is broad and collective – extending to many related exemplars stored in memory – or whether it is focused and singular – extending only to highly similar (nearly identical) exemplars."

The view that the true (i.e., abstraction-generating) categorization system is mentally separate from the instance memory system suggests the prediction that there should be a lack of generalization between category instances when a new instance that needs to be categorized queries the instance memory store. Obviously, specific category instances can be memorized, and these instance-category pairings could be stored in an instance memory system quite distinct from the prototype formed in a separate categorization system. So this suggests a conceptually precise way in which exemplar and prototype theory can be systematically specified and differentiated, especially in the historically popular binary featured category structures: on the prototype theory, a response determined by the memory

system, as distinct from the categorization system, should be the result of similarity to only a single memorized category instance.

As a preview, the purpose of the present research was to evaluate this key issue of lack of exemplar generalization in instance memory as distinct from generalization in a putative categorization system. While the exemplars versus prototypes debate has usually involved contrasting exemplar and prototype model accounts of a given data set and picking a winner, such an approach does not directly address the key generalization issue, which is fundamentally about the behaviour of the exemplar model. So rather than contrasting exemplar and prototype models (as has been done numerous times in the past), we have operationally evaluated exemplar generalization, exemplar "crosstalk", by unpacking the exemplar model's behaviour. To emphasize, our focus is not on whether the exemplar model can account for the data we present but rather on how it accounts for that data. When the exemplar model determines the probability that a test case belongs in a particular category, it calculates the overall similarity of the test case to the category exemplars (in contrast to exemplars of other categories; the Appendix presents a formalization of this account). So the issue of exemplar generalization/crosstalk reduces to looking at the exemplar similarity components of the overall category similarity, specifically the proportion of overall category similarity that is *not* due to the single nearest instance as the exemplar model is at least conceptually compatible with wide or narrow generalization. If test item category assignment is determined by similarity to a single nearest exemplar, then overall category similarity should be only negligibly greater than the similarity to the single exemplar. On the other hand, if multiple exemplars are contributing to overall category similarity when determining test item categorization, then this exemplar crosstalk should be apparent in terms of more than one exemplar contributing nontrivially to the overall category similarity. Before discussing the critical problem of how to evaluate instance memory generalization *in*

isolation from generalization in a categorization system, we need to make clear that the hypothesis of nongeneralizing instance memory is not simply a straw man.

This position is most closely implied by the conclusions of Blair and Homa (2003) but is also related to the narrow exemplar generalization argued for by Smith (2005) and others: "Researchers have raised the possibility that participants learning categories based on binary-valued dimensions (BVD) may simply memorize each member, rather than generalize across members (Blair & Homa, 2001; Smith & Minda, 2000)" (Blair & Homa, 2003, p. 1293). If a response is based on more than one instance in the memory system, then this system would be likely to generate prototypicality effects, typicality gradients, etc., and the theoretical value of a separate prototype-based categorization system would be much reduced both on the grounds of parsimony and on the grounds of pragmatically differentiating the theories. In effect, prototype theory's separate memory store for instances would be generating categorization behaviour in a similar way to exemplar theory's single system thus calling into question the conceptual and practical utility of a separate categorization system based on prototypes.

Additionally, the recent debate about predicted differences in typicality gradients for exemplars versus prototypes in the dot-distortion category learning paradigm has done little to reduce the conceptual importance of non-generalizing exemplars to differentiating the theories (Homa, Hout, Milliken, & Milliken, 2011; Smith, 2002, 2005; Zaki & Nosofsky, 2004, 2007). In this paradigm, participants make category endorsement judgments for prototypes and for low, medium, and high distortions of a prototype. Smith (2002) argued that exemplar theory necessarily predicts a flatter typicality gradient around the prototype than prototype theory and that prototypes better account for the observed pattern of responding. The intuitive argument for this (Smith, 2002) is to imagine a ring of exemplars with an untrained prototype at the centre. Consider test items which approach the prototype

along a straight line. As test items get progressively closer to the ring of exemplars, there should be a sharp rise in category endorsement, but this typicality gradient should flatten out for items within the exemplar ring because as the test item is getting closer to items on the far side of the ring it is getting farther from exemplars on the near side of the ring. Zaki and Nosofsky (2004) agree that this prediction is to some degree correct but argue that the data in this paradigm are misleading due to methodological problems, for example test phase stimuli, especially prototypes, being incorporated into the representation. Zaki and Nosofsky (2007) further support this argument by providing evidence of higher rates of endorsement for high distortions than for prototypes by simply including many high distortions in the test phase. In response, Homa et al. (2011) argue that the conclusions of Zaki and Nosofsky (2004, 2007) are in turn based on an artifact deriving from use of the single category endorsement paradigm, and that false prototype enhancement effects largely disappear in a multi-category paradigm. However, examination of the typicality gradients obtained by Homa et al. (2011) suggests that they look rather more like the flat gradients that Smith (2002) ascribes to the exemplar model than the steep gradients typical of the prototype, and in addition Homa et al. did not report exemplar and prototype model fits to their results.

As is not uncommon in the exemplar versus prototype debate, the arguments are becoming more and more complex, and there does not (at least yet) seem to be a clear winner in the dot-distortion paradigm. Nor, in our opinion, is there likely to be one in the near future as some of the key things that make the dot distortion paradigm interesting, such as the complexity and apparent ecological plausibility of the stimuli, also make clear assessment of the precise similarity relationships between the instances in individual participants difficult. In addition, the complex nature of the stimuli seems likely to induce highly idiosyncratic representations in different participants in part from large differences in selective attention as

well as complex interactions with prior representations along the lines of the different patterns people see when looking at the stars.

A priori, establishing differences between typicality gradients seems likely to correspond to a considerably weaker theoretical contrast than to argue for the presence versus absence of generalization across instances. Hence, the present research focuses on contrasting exemplar theory with a dual-system prototype theory: In addition to specifying the behaviour of the separate, prototype-based categorization system, this dual-system theory makes a clear prediction for how the instance memory system should behave in a category learning task; namely it computes similarity to only a single category instance without generalization across category instances. This is the position strongly espoused by Blair and Homa (2003) and at least partly espoused by Smith (2005). In addition, it makes clear, falsifiable predictions about whether multiple instances influence categorization, as will be detailed below. To evaluate these predictions behaviourally, there must be some way of separately detecting the influence of both of these two systems on responding. In particular, there needs to be a way of testing the hypothesized instance memory system, preferably uncontaminated by the prototype-based categorization system, to see if only a single stored instance influences categorization responding without generalization across instances.

Applying a classic neuropsychological methodology to a cognitive problem, Blair and Homa (2003; also see Smith, 2005; Smith & Minda, 2000, 2002) not only proposed that the prototype-based categorization system can be effectively ablated but argued that this is in fact what has happened in many categorization studies supporting exemplar theory because they have used ecologically implausible category structures. Specifically, a lot of support for exemplar theory has come from studies using two categories composed of a small number of poorly differentiated instances with features that are only binary-valued across instances (e.g., the 5-4 structure from Medin & Schaffer, 1978). In this context, “poorly differentiated”

means that the prototypical features of one category occur quite frequently in instances of the other category and vice versa. These structures are argued to be ecologically invalid because most real word categories putatively have well-differentiated prototypes generated from many instances composed of numerous features and are learned in contrast to far more than a single other category. So binary-valued category structures with few instances do not represent a reasonable simplification of the vast majority of categories which occur in the real world, and as such, categorization performance on these categories tells us little if anything about how the mind actually generates adaptive category-based behaviour.

From this dual system perspective, requiring people to memorize poorly differentiated categories composed of binary-valued stimuli should have the useful consequence of resulting in a failure of the categorization system to abstract category prototypes. This leaves responding to be based solely on single memorized instances retrieved from the separate memory system and so allows the influence of that system on responding to be directly evaluated, uncontaminated by the prototype-based categorization system which has been ablated and rendered irrelevant. Effectively, the argument is that the mind does not treat this as a categorization task at all and so just deals with it as an instance memorization task where some of the instances happen to share common labels.

Of all these problematic binary-valued category structures, arguably the most influential is the widely employed 5-4 category structure from Medin and Schaffer (1978). Blair and Homa (2003) proposed that this particular category structure, shown on the left in Table 1, unduly favours exemplar representation because it encourages instance memorization due to its small number of poorly differentiated instances (see also Smith & Minda, 2000, for a similar line of reasoning). In this category structure there are five instances of category A and four instances of category B, designated A1-A5 and B1-B4 at the top left of Table 1. In addition, there are seven generalization test items, designated T1-T7 at

the bottom of the Table. Each category instance is represented in a row of the Table and is composed of one of two possible features on each of four feature dimensions.

In this context, Blair and Homa (2003) proposed a formal measure of a category influence on category learning which they called “category advantage”. They had participants learn the instances of the 5-4 category structure as an identification (rather than a categorization) task over a series of trials with feedback where each of the nine instances was assigned a unique label. Blair and Homa then compared performance in this identification learning task with performance in a standard category learning task, where participants learned to assign the instances to two categories. Category advantage is defined as higher accuracy at a given point in learning for the categorization task compared to the identification task once differences in guessing are controlled for, which is crucial because of the difference in the number of responses in the two tasks (2 versus 9). So a significant category advantage at any point in learning was argued to reflect a well-differentiated category structure having a category influence on the speed of learning due to generalization between instances. In brief, similarity among category instances influences their individual learnability. On the other hand, the absence of a category advantage was argued to indicate little if any category influence on learning due to generalization between instances. In that case, categorization behaviour would simply be due to instance memorization just as in the identification learning task.

Across several experiments with different stimulus sets, Blair and Homa (2003) found no systematic category advantage for the 5-4 category structure as measured by accuracy in the categorization task relative to the identification task. That is, participants were not able to learn the 5-4 category structure in a categorization task any faster than they were able to learn to assign the nine instances to nine unique outcomes, each with a different label, in an identification task.

Blair and Homa used this lack of a category advantage to argue that "the widely used and influential 5-4 categories are learned chiefly by memorization, with no generalization across category members" (p. 1300). More generally: "If one assumes that there is little or no relationship between memorization and categorization, then the 5-4 category learning task has little to do with categorization . . . ." (p. 1299). It follows that the enormous amount of support that exemplar theory has received from binary-valued category structures is called into question. Also, as we have argued above, this "generalization across category members" is, indeed, important to systematically differentiate prototype and exemplar theory both pragmatically and theoretically.

It is worth emphasizing that this is not just a minor methodological dispute over a specific category structure, however influential that structure has been. Paradoxically the very attributes of this structure which have been argued to be ecologically invalid, maybe even eliminating much of the evidence for exemplar theory all at once, make it an ideal candidate for evaluating the dual-system prototype theory described above: Categorization responding in the absence of well-differentiated prototypes should exhibit no category influence but rather be based solely on a single memorized exemplar retrieved from instance memory with no generalization across category members. Experiments 1 and 2 were designed to operationally evaluate this claim.

Experiment 1

The purpose of this experiment was to assess category influences on instance "memorization" (Blair & Homa, 2003) using the poorly differentiated 5-4 category structure, shown on the left in Table 1 (Medin & Schaffer, 1978). However, unlike Blair and Homa, we do not contrast category and identification learning because we accept they have demonstrated an absence of "category advantage" for this category structure (though we reconsider the interpretation of this evidence in the General Discussion). It is worth noting that if we have

been too quick to accept Blair and Homa's evidence for the absence of a category advantage for the 5-4 structure, then their argument is immediately invalidated. For instance, identification learning of this structure may simply be harder than classification learning and there may be a true category learning advantage which they did not detect. However, our approach does not call into question their key empirical starting assumption, and it is also worth noting that our evaluation of category influence and crosstalk similarity is relevant even if this assumption is incorrect.

Rather than comparing categorization with identification learning, our approach was to contrast instance memorization in the standard 5-4 category structure with instance memorization in an even more poorly differentiated category structure. We generated this low-differentiation category structure by swapping two of the instances in the standard category A with two of the instances in the standard category B, as shown in the right-most column of Table 1. It is not material that this low-differentiation structure is linearly inseparable and accurate performance cannot be based on prototype representation: The prototype-based system has been argued to already be uninvolved in learning the standard 5-4 structure anyway. Both conditions were otherwise identical.

(Table 1 about here)

A variety of different theoretical perspectives provide a practical rationale for thinking that poorly differentiated category structures should be harder to learn than better differentiated ones, including Pothos and Bailey's (2009) concept of category intuitiveness (see also Pothos et al., 2011, and Pothos, Edwards, & Perlman, 2011). Pothos and Bailey used category intuitiveness framed in the context of unsupervised categorization and an exemplar model to account for testing trial differences from standard feedback learning of the 5-4 structure. However, our specification of the low-differentiation condition was most

closely dictated by the discussion of the relatively poor differentiation of the 5-4 category structure itself as described in Smith and Minda's (2000) consideration of thirty replications of this category structure. Specifically, they evaluated the 5-4 categories in terms of structural ratio as a measure of category differentiation (related to a similar concept from Homa, Rhoads, & Chambliss, 1979), that is, the ratio of within category similarity (including self-similarity) to between category similarity in terms of average numbers of features shared between instances. The ratio for these categories is 2.4/1.6=1.5 where 1 represents a complete lack of differentiation with identical instances in both categories. Hence these instances are quite poorly differentiated[1]. However, our swapping the category assignment of two pairs of instances from the standard to the low-differentiation condition (Table 1) resulted in even poorer differentiation as measured by a structural ratio of 2.2/1.9=1.2, that is fewer shared features within the same category and more shared features between categories. Before moving on to why we were interested in learnability differences, it is worth emphasizing that, while a variety of theoretical perspective predict these differences, our research does not presuppose the truth of any of these perspectives. Rather our conclusions are based on the learnability differences we observed as reported below.

The theoretical rationale for this instance swapping manipulation was simply that if participants learn the standard 5-4 category task as an instance identification task, *with no category influence from other category members on responding*, then it should not matter how the instances are assigned to categories. The low-differentiation category structure should be as easy to learn as the standard one and all the instances in both conditions should be equally easy to learn because the participant is simply learning to assign a label to each unique instance. On the other hand, if there is a category influence on learning, then the low-differentiation 5-4 structure should be even harder to learn than the standard 5-4 structure,

and differences in instance learnability should correspond to differences in similarity across instances.

Specifically, the dual-system perspective of prototypes plus memory for single instances predicts no category influence for either learning condition because these poorly differentiated structures do not give the prototype-based categorization system anything to work with, leaving responding to be based solely on memorized single instances without generalization. This implies that test accuracies on the different training items (A1-B4 in Table 1) should not be systematically different from each other and the resolution of ambiguities in terms of nearest single category exemplars for the generalization test items, T1-T7, should be arbitrary (e.g., for the standard condition on the left in Table 1, T5 2121 shares an equal number of features with A4 1121 and with B3 2221). Test items should not reflect the influence of multiple instances or the number of prototypical features as the prototype based categorization system should have been effectively ablated. Generalization test items with many features typical of the prototype of one category (e.g., T3 1111 for the standard condition in Table 1) which match many exemplar features of that category should be no more likely to be categorized as members of that category than instances with fewer features (e.g., T6 2211 for the standard condition in Table 1) because neither matches a training exemplar exactly and both have multiple ambiguous matches based on subsets of features.

The single-system exemplar theory, on the other hand, is consistent with differences in accuracy for the training items corresponding to a category influence in terms of differences in similarity to multiple category instances. That is, category instances that are similar to many other category instances and correspondingly dissimilar to instances in the other category should be easier to learn than category instances that are not similar to multiple instances in the same category and that are rather similar to instances in a contrast

category. Likewise since categorization and instance memory are assumed to be part of a single system, the influence of multiple instances on responding is consistent with prototypicality effects in the generalization test items: items with many typical features of a given category should tend to have a higher proportion of responses for that category than items with fewer typical features because they will tend to be similar to many instances of the category. Lastly, exemplar theory is consistent with a category influence in terms of differences in learning and generalization test performance between the standard and low-differentiation conditions based on differences in similarity relations across instances for the two conditions.

Method

Participants

There were 40 participants in the standard and 41 participants in the low-differentiation conditions, all undergraduate psychology students at Cardiff University.

Materials

The abstract category structure for the standard condition on the left in Table 1 is the 5-4 structure from Medin and Schaffer (1978). There are five instances of category A, four instances of category B (at the top left of the Table), as well as seven generalization test items (at the bottom of the Table). Each category instance is represented in a row and was composed of features from four binary-valued stimulus dimensions as indicated by 1's and 2's in the Table. The low-differentiation condition on the right of Table 1 was generated from the standard condition by swapping the category assignment for two pairs of instances: A1 1112 was switched with B3 2221 in the standard structure to be A1 2221 and B3 1112 in the low-differentiation condition, and A3 1211 was switched with B2 2112 to become A3 2112 and B2 1211. In Table 1, the switched items are in bold in the low-differentiation condition. All other training and test items were the same in both conditions.

The stimuli were the alien insects from Johansen and Kruschke (2005). Four binary-valued stimulus dimensions were randomly assigned to the four abstract category dimensions for each participant in the standard and low-differentiation conditions. These four stimulus dimensions were randomly chosen from a set of five possible dimensions—shape of the head (round or square), orientation of the nose (up or down), length of the tail (short or long), shape of the antennae (curved or straight) and leg number (eight or four)—and the fifth dimension had a fixed value across all the stimuli for a given participant. In addition, the polarity of the assignment of stimuli within each stimulus dimension was also randomly assigned between participants as was the assignment of the category labels LORK and THAB to the abstract categories A and B in Table 1. Each stimulus was displayed on a computer monitor, and participants indicated their response by mouse clicking in a box containing one of the category labels.

Procedure

The procedure for both the standard and low-differentiation conditions was the same as the classification condition in Johansen and Kruschke (2005). Participants from both conditions were run simultaneously, half in one condition and half in the other based on even or odd participant numbers, and the participant instructions were identical throughout. The instructions told participants that they should learn to assign the instances to categories and would receive feedback. They were also warned that there would be a test phase at the end of the experiment to evaluate how much they had learned and that they would not receive corrective feedback during this phase.

Each training trial displayed one of the nine category instances from either the standard or low-differentiation conditions shown at the top of Table 1. After making their response, participants received feedback, either CORRECT! or WRONG!, the correct category label was displayed above the stimulus, and participants were given up to 30

seconds to study the correct answer. The message FASTER was displayed if they did not make their initial response within 20 seconds of the start of the trial. They started the next trial by clicking the mouse on a box which had the message "After you have studied this case (up to 30 seconds), click here to see the next one." If they did not start the next trial within 30 seconds of their initial response, the message FASTER appeared and then the next trial started automatically.

In the training phase, the nine training instances in each of the two conditions were presented in 25 randomly ordered blocks for a total of 225 training trials. After the training phase, participants were instructed that they would be tested without feedback but that they should base their responses on what they learned when feedback was being given. During the test block, the nine training items and seven test items for the standard and low-differentiation conditions (Table 1) were presented in random order. Participants were given no feedback on each trial other than that their response had been recorded.

Results and Discussion

The averaged learning curves for all participants by condition, standard or low-differentiation, are in the top panel of Figure 1 which shows a systematic difference between conditions throughout the course of learning: the low-differentiation structure was considerably harder to learn than the standard structure as measured by average accuracy across sets of 5 blocks ($F(1,79) = 35.053$, MSE = 0.028, $p < 0.001$, in an ANOVA with training condition and training block set as factors). In addition, the proportions of good versus poor learners in reference to a learning criterion (≥ 0.75 in the last two blocks of training, as in Johansen and Palmeri, 2002), was much higher in the standard condition (17/40 good learners) than in the low-differentiation condition (3/41) ($\chi^2(1) = 13.48$, $p < 0.001$). (Figure 1 about here) In fact, the differences in performance for the two learning conditions occurred across the course of learning even in the relatively poor learners who did

not ultimately achieve the ≥ 0.75 criterion as shown in the bottom panel of Figure 1. (We do not report the results for the best learners in this way as there were only three good learners in the low-differentiation condition). Overall the standard category structure was easier to learn throughout and resulted in higher performance at the end of learning compared to the low-differentiation structure.

However, overall average accuracy does not tell the complete story. Importantly, we need to ask whether there were response differences between category instances, and we need to examine test generalization to new instances. In their study, Blair and Homa (2003) reported neither of these analyses, but they are crucial for differentiating the dual-system theory (a prototype-based categorization system plus a nongeneralizing memory for instances system) from the single-system exemplar theory, particularly in the context of assessing category influence.

Figure 2 shows the category A response proportions from the test phase for the nine training and seven generalization test items in each training condition (Table 1) based on all participants in both conditions. The dashed lines are the approximate 95% binomial confidence interval around 0.5 and provide a useful reference for the amount of learning at the end of training: All but one of the training items fell outside this interval in the standard condition whereas all but two of the training items fell *within* this interval in the low-differentiation condition. Thus the test phase results also show that accuracy on the individual test items was generally higher in the standard condition than the low-differentiation condition. (Figure 2 about here)

Further, there were systematic and stable differences in test trial performance within the standard condition. Figure 3 shows performance by condition for good versus poor learners, i.e., high versus low end of training accuracy, as specified in reference to the learning criterion. For the standard condition, similar qualitative patterns of responding

occurred in the good versus poor learners, e.g., A1, A2, and A3 (labels from the standard condition in Table 1) were the best learned instances from category A and B3 and B4 from category B. Also the generalization test item corresponding to the prototype of A, T3 1111, was strongly assigned to category A while T7 2212 was quite strongly assigned to category B in both. Importantly, these patterns of responding also correlate closely with the data from Medin and Schaffer (1978), shown at the bottom of Figure 3, $r = 0.94$ for the good learners. The contrast between good and poor learners in the low-differentiation condition is not informative because there were so few good learners, and there was only weak evidence for differences between the test items in the low-differentiation condition, though by no means completely absent. (Figure 3 about here)

In addition, the differences in training trial performance in the standard condition's test block were quite stable because they correspond to similar differences throughout the course of learning. Figure 4 shows learning curves for different training trial types by condition (Table 1) in terms of proportions of category A responses averaged into sets of 5 training blocks and based on all participants. As in the test data in Figure 2, exemplars A1, A2, and A3 as well as B3 and B4 were the easiest to learn in the standard condition and this pattern persisted throughout learning. On the other hand, the training items were only weakly different from each other in the low-differentiation condition (on the right in Figure 4) as emphasized by the fact that some category A items were not even systematically differentiated from the category B items. (Figure 4 about here)

The most compelling evidence for category influences on learning in the standard condition comes from a comparison of individual category instances that had the same category assignment in both the standard and low-differentiation conditions and were learned well in the standard task. Consider A2 1212 (shown by squares connected with solid lines in both conditions in Figure 4). There were systematic differences between conditions

throughout learning for this item with strong assignment to category A in the standard condition and weak assignment to category B in the low-differentiation condition (despite the feedback to the contrary in the low-differentiation condition). Likewise, B4 2222 (shown by circles connected by dashed lines in both conditions in Figure 4) was quite strongly assigned to category B in the standard condition throughout learning, but weakly assigned to A throughout most of learning in the low-differentiation condition. The remaining category instances were not as clearly differentiated between the conditions, but these instances were not learned particularly well in the standard condition in the first place.

In summary, category membership in the standard 5-4 structure has a pronounced but differential influence on the learnability of exemplars both in reference to each other and to members of a more poorly differentiated 5-4 structure. Further these differences are conceptually not compatible with participants treating the standard task pragmatically as an identification learning task with no generalization across instances and no category influences on learning. Overall, these data indicate that there are category influences on learning of the standard Medin and Schaffer (1978) 5-4 category structure despite a lack of category advantage for that structure (Blair & Homa, 2003) and despite the difficulty of learning it. In addition, exemplar modelling supports this category influence conclusion.

Exemplar Modelling of the Results

Results from the 5-4 category structure have been modelled and discussed many times in the past (Medin & Schaffer, 1978; Nosofsky, 2000; Smith & Minda, 2000; etc.), so it will come as no surprise that the exemplar model provides a good account of the results from the present experiment as detailed below. However the deeper purpose of this modelling analysis is to take a closer look at how the exemplar model accounts for the results in the context of two contrasting conclusions: firstly, Blair and Homa's (2003) no-category-advantage conclusion that this category structure induces instance memorization with no generalization

across instances and, secondly, the fact that there are clear category influences on learning for this structure.

Category learning research has commonly instituted a learning criterion such that only those participants who have learned the categories sufficiently well are included in the key data set, which is then used to address primary research questions such as how participants represent categories, use them, etc. Moreover the assessment of the participant proportions who reach the various criteria for the 5-4 structure have lead Blair and Homa (2003) as well as Smith and Minda (2002) to emphasize the relative difficulty of learning this structure, with substantial proportions of participants not learning to criterion. However, we have focused on modelling the results for all participants as conceptually a learning criterion has the potential to compress or mask differences in instance learnability against a ceiling of perfect performance.

Figure 5 shows the exemplar model's best fit predictions plotted against the data from the standard and low-differentiation conditions respectively. (The details of the model and the maximum likelihood modelling procedure can be found in the Appendix.) The model does a reasonable job of accounting for the data in both conditions with an overall fit of G^2 = 10.270 corresponding to a Root Mean Squared Deviation (RMSD) of 0.053 for the standard condition[2] (with dimensional attention parameters of 0.369, 0.108, 0.297, and 0.227, for the four dimensions respectively, and a similarity scaling parameter of c = 4.040) and with an overall fit of G^2 = 15.146 corresponding to RMSD = 0.073 for the low-differentiation condition[3] (with dimensional attention parameters of 0.406, 0.100, 0.339, and 0.155, respectively, and a similarity scaling parameter of c = 2.746). (Figure 5 about here)

Tables 2 and 3 also show the exemplar model's predictions for the standard and low-differentiation conditions by trial type, where the data and model prediction values for each trial type are the data points in the scatter plots shown in Figures 5. In particular, the column

in Tables 2 and 3 labelled “SumSim A” has the sum of each test item’s similarities to all the exemplars in category A, and “SumSim B” is the sum of the item’s similarities to all the exemplars in category B. If a test item perfectly matches an exemplar in the category representation, then its similarity to that exemplar is 1.0. Since the training items, A1-B4, all exactly match one item, the extent to which the summed category similarities for these items are greater than 1.0 partly indexes a category influence on responding, that is, the influence of exemplars other than the perfectly matching memorized instance. As none of the training items A1-B4 were members of both categories, a training item never perfectly matched an exemplar in the opposite category, so the summed similarity to exemplars of the opposite can be less than 1.0, but it doesn’t have to be as the cumulative similarity to multiple exemplars has the potential to be greater than 1.0.

The training items from the standard condition can be divided into high accuracy items—A1, A2, A3, B3, and B4—and low accuracy items--A4, A5, B1, and B2—in Figure 2. It is important to emphasize that this is not just an arbitrary division, but one both strongly suggested by the standard condition learning curves (Figure 4) and widely replicated. For example, this pattern of testing trial accuracies occurred in Medin and Schaffer’s (1978) Experiment 3 shown at the bottom of Figure 3, in the various additional data sets summarized in Figure 7, and in the average data from 30 replications of this category structure reported by Smith and Minda (2000).

For the high accuracy training items from the standard condition, the average of the summed similarity to members of their own category in Table 2 is 1.92. If self-similarity is removed from this by subtracting 1, then the proportion of category similarity due to other exemplars is 0.92/1.93 = 0.48, or roughly half of the category similarity. In contrast, the average summed similarity for low-accuracy training items was 1.36, so the proportion due to other category exemplars is 0.36/1.36 = 0.26, or only about one quarter of the category

similarity. Correspondingly, the average summed similarities of the high accuracy training items (Table 2) to members of the opposite category was quite low, 0.40, while twice as high for the low accuracy items, 0.80.

(Table 2 about here)

The division of the low-differentiation condition training items (Table 3) into a high accuracy set—A1, A3, A5, B1, and B3—and a low accuracy set—A2, A4, B2, and B4—was somewhat more arbitrary because this structure was much harder to learn and the specific item learning curves in Figure 4 were more poorly differentiated. However, even in this very poorly differentiated category structure, the average summed similarity for high accuracy items to members of their own category in Table 3 was 2.02 versus 1.64 for the low accuracy items. With self-similarity removed, the proportion of category similarity due to other exemplars was 0.51 for high accuracy items and 0.39 for low-accuracy items. So even for this poorly differentiated structure the amount of category influence was somewhat larger for high than low accuracy items, though the difference was not as large as for the standard condition data. But the key difference from this account of the standard condition was that for the low-differentiation condition, average summed similarity to members of the contrast category for high accuracy items was 1.13 and was 1.54 for low accuracy items. Hence contrast category similarity was substantially higher in the model's account of the low-differentiation condition as would be expected from the poorer differentiation compared to the standard structure.

(Table 3 about here)

In summary, the purpose of this assessment is not to argue that prior modelling analysis of data from this category structure contrasting exemplar and prototype theory is wrong. Rather the purpose has been to look at how the exemplar model accounts for the data in the context of the claim that learning is based on instance memorization without generalization between instances. The exemplar model explains the differences in training item accuracy observed in the data in terms of differential category influences on learning from both the member and contrast categories, that is differential generalization across multiple instances. In fact, we propose it is the relatively poor differentiation of these category structures that enhances differential item learnability and makes arguing for nongeneralizing exemplars so implausible. While selective attention helps people and the exemplar model to differentiate the categories and so achieve high training item accuracy in the first place, differential within and between category similarity compellingly explains differential learnability. The latter is very hard to explain, in contrast, if this categorization task is equivalent to an identification learning task with no exemplar crosstalk.

Experiment 2

The basic claim we have critiqued in Experiment 1 is that learning of the 5-4 category structure just invokes simple instance memorization with no generalization across instances. The purpose of Experiment 2 was to provide a methodological contrast to the standard learning with feedback paradigm used in Experiment 1 and substantially emphasize memorization by explicitly telling participants to memorize the category instances from a simultaneously presented summary (Figure 6). The basic intent of this paradigm was to induce participants to encode the category instances from the standard condition (on the left in Table 1) into memory in as methodologically simple a way as possible. In the standard category learning paradigm, participants learn a category structure by categorizing single instances over a protracted series of trials with feedback where the resulting error drives

selective attention and gradually improves associative performance. Conceptually, selective attention driven by explicit feedback might operate to make cross-instance similarity more differential than it might otherwise have been. In contrast, participants in this experiment were simultaneously presented with the instances from the standard 5-4 structure (on the left in Table 1) grouped with category labels (Figure 6) and explicitly instructed to memorize them in a short interval. Thus participants received no feedback because there were no responses and no trial structure. (Figure 6 about here)

Method

Participants

There were 30 participants, who were undergraduate students at University College London or Cardiff University.

Materials and Procedure

Participants were instructed to memorize the instances from two categories on a category summary sheet as shown in Figure 6. The category instances were described as being rocket ships from Planets A and B. After being given 5 minutes to memorize these instances, the summary sheet was removed, and participants were asked to assign each of the 16 cases from the standard condition in Table 1 to one of the categories by circling one of the two possible category labels below the instance (Planet A or Planet B). For methodological simplicity, the presentation order of the test trials was the same for all participants—T1, T4, T2, T3, T5, T6, T7, A1, B4, B1, A5, A3, A4, B3, A2, and B2—as was the assignment of abstract to physical dimensions—Dimension 1 = wing width (1=narrow/2=wide), Dimension 2 = cone shape (1=curved/2=pointed) dimension 3 = booster number(1/2) and Dimension 4 = portal orientation (1=down/2=up).

Results and Discussion

Overall categorization performance on the 9 category instances resulted in an average accuracy of 0.73 based on all participants. Comparison with various published learning results from the standard paradigm and different numbers of training blocks, on the right in Figure 7, shows that this is similar to that observed in the learning results from after a full 16 blocks of training, 0.72 (Nosofsky, Palmeri & McKinley, 1994), where each block had all 9 category instances for a total of 144 trials. Thus the overall performance level obtained with the present explicit memorization task is highly consistent with one aspect of Blair and Homa's (2003) argument, that participants in the standard paradigm memorize the instances of the 5-4 category structure. (Figure 7 about here)

The left panel of Figure 7 shows that categorization performance on the 9 category instances, A1-B4, was highly varied, ranging from almost chance accuracy (e.g., instance B1), to more than 90% accuracy (instance B4). Importantly, this variation was not random but rather was systematic and highly correlated with the published learning results based on different amounts of training in the standard trial-by-trial learning with feedback paradigm (the right panels of Figure 7), e.g. $r = 0.92$ with results from after 32 blocks of training (Johansen & Palmeri, 2002; Nosofsky, Palmeri and McKinley, 1994; Palmeri & Nosofsky, 1995). In particular, these results duplicated the same differences in training item accuracy observed in the standard condition of Experiment 1 both at the end of learning (Figure 2) and over the course of learning (Figure 4), that is, higher accuracy on A1-A3, B3, and B4 than on A4, A5, B1, and B2. And lastly, the generalization test results for T1-T7 showed a moderate level of correspondence to those from the learning reference data, most notably in terms of a large prototype effect for T3, the prototype of category B.

The comparability of these to prior results demonstrates that not only were there differences in accuracy for the category members but that this methodologically simple paradigm produced results that are surprisingly similar to those from the far more elaborate

multi-trial learning with feedback paradigm. In the context of the dual-system prototype theory being considered here, there are two contrasting interpretations that can be drawn from these results depending on whether they are considered to represent the complete ablation or the active facilitation of the mind's categorization system, as distinct from the nongeneralizing-exemplar-memorization system.

One perspective is to use the minimalist instance-memorization paradigm to argue that the explicit memorization instructions should even more strongly invoke instance memorization and hence even more completely ablate the categorization system. In addition, the lack of feedback might further minimize the effects of selective attention and similarity, also arguably key components of the categorization system, and leave responding to be firmly based on the nongeneralizing instance memorization system. From this perspective, the fact that the differences in training item accuracy in this experiment so closely replicate those from the standard paradigm strongly argues against nongeneralizing exemplars in the dual-system theory. So even when instances are encoded into memory in as simple a way as possible there are still clear category influences, and it is not clear from the perspective of parsimony what the dual-system account adds here.

However, an alternative perspective is that our instance-memorization paradigm had the exact opposite of the desired effect; far from completely ablating the categorization system, maybe this paradigm actively facilitated the categorization system by allowing participants to more easily observe the commonalities and differences between instances, both within and between categories. That is, the summary presentation methodology effectively enhanced the perceived differentiation of the categories and thus actually activated rather than ablated the categorization system. Probably the most compelling evidence for this perspective is that the summary memorization paradigm produced a large prototype effect during the testing phase (T3), as large as that in the extensive training

condition from Johansen and Palmeri (2002). But if the summary memorization results represent enhanced performance of the activated categorization system then that system produced results that look remarkably similar to those from the standard task in Experiment 1 which are so well accounted for by the exemplar model in terms of generalizing across multiple category instances. Thus from the perspective of parsimony, it is unclear what dual-system prototype theory with nongeneralizing exemplar memory is adding in this alternative interpretation.

General Discussion

The two experiments presented here evaluated category influences on instance learnability in the 5-4 category structure (Medin & Schaffer, 1978) from two different methodological perspectives. Experiment 1 contrasted feedback learning in the standard 5-4 structure with learning in an even more poorly differentiated category structure generated by switching the category assignment of two pairs of instances. Blair and Homa (2003) compared the standard category learning task with an identification learning task and argued that the comparable learnability of the two tasks indicated a lack of “category advantage”, which they took as evidence that responding was based on instance memorization with no generalization across instances. However, the present results indicate that a more poorly differentiated category structure was even harder to learn, a difficult result to explain if multiple category instances are not influencing responding, particularly as there were systematic differences in performance across category instances. Further, as would be expected from the history of this category structure, the exemplar model provided a good account of these data. In particular, it accounted for the differential learnability of the training items in terms of differential similarity to multiple instances both within and between categories. In essence, Experiment 1 uses a variation of Blair and Homa’s category advantage methodology to show an advantage in learning some instances in the standard 5-4 structure

compared to a less differentiated category structure. Pure instance memorization without generalization cannot explain this observation.

Experiment 2 was a minimalist memorization task where participants were presented with a simultaneous display of all the category instances (Figure 6) and given 5 minutes to memorize them. The results of this task look strikingly like those from the standard category learning paradigm (e.g., Johansen & Palmeri, 2002; Medin & Schaffer, 1978; Nosofsky, Palmeri and McKinley, 1994; Palmeri & Nosofsky, 1995), despite the large differences in the methodology. While this suggests that instance memorization is a fundamental part of that paradigm–consistent with the claims of Blair and Homa (2003) and Smith and Minda (2000; also Smith, 2005)–these findings also indicate that even this simple way of encoding instances into memory resulted in a category influence on learning and generalization. This category influence is based on comparison to more than one instance and yields clear prototypicality effects, even though this poorly differentiated category structure putatively does not lend itself to anything other than the memorization of specific instances with no generalization across multiple instances: "The lack of a category advantage . . . strongly suggests that participants who learn these categories learn them by memorization and receive no benefit from generalizing among members of the same category" (Blair & Homa, 2003, p. 1298).

At minimum, these results suggest that the conclusions from Blair and Homa's (2003) category advantage methodology should be viewed with caution. Their argument is that lack of category advantage for category learning compared to identification learning indicates a poorly differentiated category structure unrepresentative of real world categories which does not invoke the brain's categorization system and which is dealt with only by instance memory. Our counterargument is that even for this poorly differentiated category structure, there are still clear category influences.

We suggest that the problem with Blair and Homa's (2003) category advantage argument is as follows: While categorization solely in reference to single memorized instances certainly predicts a lack of a category advantage, a lack of category advantage does not necessarily imply that categorization performance is based on memorized single instances without a category influence of multiple instances. The lack of category advantage can arise for other reasons in the presence of a category influence. For example, the crosstalk between individual instances may be such that learning is facilitated for some and harmed for others and thus average out to being comparable to identification learning, but this category influence at the level of specific instances may not be apparent from looking at an overall learning rate.

More generally, it thus seems that even an elaborated dual-system prototype theory which has been paradoxically motivated to make predictions about instance memory fails. If category decisions are based on more than one instance in the memorization system, as our results suggest, then this system should generate prototype effects, typicality gradients, and so on and the added theoretical value of a separate prototype-based 'true' categorization system becomes unclear. In effect, prototype theory's separate memory store for instances would be generating categorization behaviour in a similar way to exemplar theory's single system, at which point it is reasonable to ask: what is the prototype-based categorization system serving to explain?

There are several possible ripostes to this conclusion. One is to argue that the dual system prototype plus nongeneralizing exemplar theory we have considered is effectively just a straw man. However, given past criticisms of the 5-4 structure (Blair & Homa, 2003; Smith & Minda, 2000; etc.) and the absence of a "category advantage" with this structure, the theory is anything but a straw man. It is distinctly non-trivial, explains a range of findings, and the experiments reported here could, in principle, have yielded further support. The

following reasonably summarizes this dual-system position (Blair & Homa, 2003, p. 1299): "Data from previous research using the 5-4 categories has been difficult for prototype models to fit, and these data are generally seen as supporting exemplar theories. If one assumes a strong relationship between memorization and categorization, then the failure of a prototype model can be seen as a critical weakness of the theory. If one assumes that there is little or no relationship between memorization and categorization, then the 5-4 category learning task has little to do with categorization and is therefore an inappropriate test of prototype theory."

Another possible riposte to our conclusions is that the theoretical debate has moved on from the coarse distinction between exemplar generalization versus nongeneralization to the more subtle distinctions between typicality gradients (e.g., Homa, et al. 2011). We have summarized our reactions to the dot-distortion paradigm in the introduction. In addition, our differences in instance and category learnability provide some converging support for recent developments in theories of category intuitiveness involving assessments of what makes categories hard or easy to learn (Pothos & Bailey, 2009; Pothos et al., 2011; Pothos, Edwards, & Perlman. 2011). We would like to emphasize the overall logic and minimal conclusions of our research. We take our results as evidence for exemplar theory and thus still consider it a viable candidate to explain broad areas of categorization behaviour, perhaps even all categorization behaviour (with suitable elaboration). However, we acknowledge that some may consider research on the 5-4 category structure (Medin & Schaffer, 1978) to have reached a state of theoretical sterility, perhaps because they believe that such binary-valued category structures are unrepresentative of the real world or because they believe that prototypes clearly provide a better account of typicality gradients, even when multiple exemplars are allowed to generalize between each other. Nevertheless, even if both of these are completely true, we believe that Blair and Homa's (2003) antiexemplar category-

advantage argument is not viable and at minimum this evidence against exemplar theory should be discounted and future prototype theories suitably constrained.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120*, 3-19.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

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*, 149-78.

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.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107-140.

Hahn, U, Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition, 114*, 1-18.

Homa, D., Hout, M. C., Milliken, L., & Milliken, A. M. (2011). Bogus concerns about the false prototype enhancement effect. *Journal of Experimental Psychology: Learning Memory and Cognition, 37,* 368-377.

Homa, D., Rhoads, D., & Chamblish, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 11-23.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 418-439.

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning Memory and Cognition, 31*, 1433-1458.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482-553.

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

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

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

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. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) "Thirty categorization results in search of a model". *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 1735-1743.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. *The Psychology of Learning and Motivation, 28,* 207-250.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53-79.

Palmeri, T. J., & Nosofsky, R.M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21,* 548-568.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract *ideas. Journal of Experimental Psychology, 77,* 353-363.

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

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,* 1692-1713.

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.

Seger, C A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience, 33,* 203-219.

Smith, J. D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science, 13*, 437-442.

Smith, J.D. (2005). Wanted: A new psychology of exemplars. *Canadian Journal of Experimental Psychology, 59*, 47-53.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 411-1430.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3-27.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 800-811.

Zaki, S. R., & Nosofsky, R. M. (2004). False prototype enhancement effects in pattern categorization. *Memory & Cognition, 32*, 390-398.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition, 35*, 2088-2096.

Acknowledgements

We thank Lewis Bott, Marc Buehner, Stephan Lewandowsky, and several anonymous reviewers for detailed comments and suggestions on an earlier version of this manuscript.

Footnote

1. Structural ratio as a measurement does not have a fixed upper bound corresponding to perfect differentiation as, among other reasons, sharing zero features with instances of a contrast category leaves the ratio undefined to avoid division by 0. But some indication of a relevant upper bound can be had by making all of the instances of category A identical to the category prototype, A 1111, and likewise the B instances to B 2222 except for one B instance which shares a single feature with A, i.e. B2221. In this case the structural ratio is 3.8/0.3 = 12.7, clearly a very long way from the 1.5 for the standard 5-4 structure.
2. Fitting the model with RMSD directly rather than G^2 resulted in a virtually identical fit and set of parameters as well as predicted values: RMSD = 0.053, with dimensional attention parameters of 0.371, 0.102, 0.301, and 0.226, respectively, and a similarity scaling parameter of c = 4.031. The (unreported) scatterplot of the model's predictions against the data was virtually identical to the left panel of Figure 5.
3. Fitting the model with RMSD directly rather than G^2 also resulted in extremely similar parameters, fit, and predictions for these data: with dimensional attention parameters of 0.403, 0.126, 0.321, and 0.150, respectively, and a similarity scaling parameter of c = 2.959). The (unreported) model predictions were also very similar to those in the right panel of Figure 5.

Appendix

Exemplar Model Specification and Modelling Procedure

The exemplar model applied to the results of Experiment 1 was the Generalized Context Model (Nosofsky, 1986). In overview, the model calculates the total similarity of a test item to all the instances in a category and then specifies a category response probability by contrasting that category's summed similarity with the total similarity to all categories. The predictions from the model for these data were derived using the following two equations, four free parameters, and a best fitting parameter procedure.

In detail, each category is represented in the model by the instances composing it, the exemplars participants were told in the training phase were in that category (for example A1-A5 for category A in Table 1), each composed of features on four binary-valued feature dimensions. The model specifies the similarity of a given test item, *i*, to a given category exemplar, *j*, using equation 1.

$$\eta_{ij} = \exp\left(-c \sum_{\dim k} w_k \mid x_{ik} - x_{jk} \mid \right) \qquad (1)$$

In this equation, x_{ik} is the feature value of test item *i* on feature dimension *k*, and correspondingly x_{jk} is the feature value of exemplar *j* on the same feature dimension *k*. The absolute value of the difference between these two features, $| x_{ik} - x_{jk} |$, is then multiplied by the dimensional attention parameter for feature dimension *k*, w_k ,which specifies how much a feature difference on that dimension should matter in the similarity calculation. These weighted differences are calculated for all four feature dimensions, summed and multiplied by -*c* to yield the content of the parentheses in equation 1, where *c* is an overall scaling parameter for similarity. This scaled sum of weighted feature difference values is then exponentiated to give the similarity of test item *i* to exemplar *j*, η_{ij} .

Using the similarities generated by equation 1, the category response probability that a test item *i* will be from category A is determined by equation 2. Specifically, the model calculates

$$P(catA \mid i) = \frac{\sum_{j \in catA} \eta_{ij}}{\sum_{j \in catA} \eta_{ij} + \sum_{j \in catB} \eta_{ij}} \qquad (2)$$

the total similarity of a test item *i* to all the instances in category A by summing across its similarity to each of the specific instances in that category as in the numerator of equation 2, $\sum_{j \in catA} \eta_{ij}$. This summed similarity to the representation for category A is divided by the summed similarity to all the instances in both categories, category A on the left and category B on the right in the denominator.

The model's best fitting predictions for the data in a given condition of Experiment 1 were generated by adjusting the free parameters to minimize the maximum likelihood statistic G^2. We have not drawn any statistical conclusions about the goodness of fit of this model because the data substantially violate independence. In practice, using G^2 resulted in very similar predictions to the more traditional procedure of minimizing root mean squared deviation (RMSD), the discrepancy between the model's predictions and the data. So even though the fits used G^2, we also report RMSD to be comparable to previously reported fits of this model. Of the model's four dimensional attention parameters, w_k , in equation 1, three are free parameters and the fourth is constrained such that the sum of the four attention parameters is 1. The overall similarity scaling parameter *c* is a free parameter rather than simply being redundant with the four attention parameters by the distributive rule. Various sets of initial free parameter values were chosen and adjusted via a hill-climbing procedure. The model's best fitting parameter values and predictions for the two conditions of Table 1 are reported in the main text.

Table 1. Standard 5-4 Category Structure (Medin & Schaffer, 1978), Low-differentiation 5-4 structure, and Test Cases Used in Experiment 1's Standard and Low-differentiation Conditions.

trial types	standard	low-differentiation
A1	A 1 1 1 2	**A 2 2 2 1**
A2	A 1 2 1 2	A 1 2 1 2
A3	A 1 2 1 1	**A 2 1 1 2**
A4	A 1 1 2 1	A 1 1 2 1
A5	A 2 1 1 1	A 2 1 1 1
B1	B 1 1 2 2	B 1 1 2 2
B2	B 2 1 1 2	**B 1 2 1 1**
B3	B 2 2 2 1	**B 1 1 1 2**
B4	B 2 2 2 2	B 2 2 2 2
T1	? 1 2 2 1	? 1 2 2 1
T2	? 1 2 2 2	? 1 2 2 2
T3	? 1 1 1 1	? 1 1 1 1
T4	? 2 2 1 2	? 2 2 1 2
T5	? 2 1 2 1	? 2 1 2 1
T6	? 2 2 1 1	? 2 2 1 1
T7	? 2 1 2 2	? 2 1 2 2

Note. In the test cases, ?'s indicate that there was no correct category as participants never received feedback for these items. The category instance trial type labels A1-B4 are for reference purposes only as are the generalization test trial labels T1-T7. The two bold instances in each category of the low-differentiation condition were members of the other category in the standard category structure in the middle column. Specifically, A1 and A3 in the low-differentiation condition are B3 and B2, respectively, in the standard condition, and B2 and B3 in the low-differentiation condition are A3 and A1, respectively, in the standard condition.

Table 2. Exemplar Model Predictions for the Standard Condition from Experiment 1.

trial types	standard	data	model	SumSim A	SumSim B
A1	A 1 1 1 2	0.78	0.78	2.12	0.59
A2	A 1 2 1 2	0.83	0.83	2.18	0.44
A3	A 1 2 1 1	0.88	0.90	2.00	0.23
A4	A 1 1 2 1	0.68	0.70	1.46	0.63
A5	A 2 1 1 1	0.68	0.66	1.36	0.70
B1	B 1 1 2 2	0.33	0.44	1.00	1.27
B2	B 2 1 1 2	0.40	0.39	0.86	1.34
B3	B 2 2 2 1	0.30	0.23	0.45	1.54
B4	B 2 2 2 2	0.05	0.14	0.28	1.74
T1	? 1 2 2 1	0.70	0.67	1.19	0.59
T2	? 1 2 2 2	0.48	0.47	0.89	1.01
T3	? 1 1 1 1	0.83	0.87	1.83	0.27
T4	? 2 2 1 2	0.43	0.40	0.74	1.11
T5	? 2 1 2 1	0.30	0.36	0.62	1.12
T6	? 2 2 1 1	0.50	0.60	1.07	0.70
T7	? 2 1 2 2	0.20	0.19	0.34	1.43

Note. Both the data and model predictions are in terms of category A response proportions. SumSim A indicates each test item's summed similarity to all the exemplars of category A, and SumSim B to all the exemplars of category B.

Table 3. Exemplar Model Predictions for the Low-differentiation Condition from Experiment 1.

trial types	low-differentiation	Data	model	SumSim A	SumSim B
A1	**A 2 2 2 1**	0.61	0.64	1.83	1.01
A2	A 1 2 1 2	0.46	0.48	1.69	1.84
A3	**A 2 1 1 2**	0.63	0.70	2.18	0.92
A4	A 1 1 2 1	0.54	0.55	1.66	1.37
A5	A 2 1 1 1	0.63	0.75	2.25	0.74
B1	B 1 1 2 2	0.34	0.42	1.33	1.84
B2	**B 1 2 1 1**	0.37	0.46	1.49	1.78
B3	**B 1 1 1 2**	0.32	0.45	1.62	1.99
B4	B 2 2 2 2	0.39	0.50	1.44	1.43
T1	? 1 2 2 1	0.51	0.54	1.51	1.30
T2	? 1 2 2 2	0.37	0.44	1.27	1.64
T3	? 1 1 1 1	0.39	0.47	1.53	1.73
T4	? 2 2 1 2	0.68	0.67	1.91	0.96
T5	? 2 1 2 1	0.61	0.67	1.80	0.89
T6	? 2 2 1 1	0.68	0.71	1.96	0.81
T7	? 2 1 2 2	0.61	0.53	1.46	1.28

Note. Both the data and model predictions are in terms of category A response proportions. SumSim A indicates each test item's summed similarity to all the exemplars of category A, and SumSim B to all the exemplars of category B.

Figure Captions

Figure 1. Experiment 1 training condition accuracy averaged across sets of five training blocks for all participants (top panel) and for poorer learners (bottom panel) who did not ultimately achieve the learning criterion of ≥ 0.75 in the last two blocks of training (error bars are standard errors).

Figure 2. Experiment 1 test trial response proportions by training condition with an approximate 95% binomial confidence interval on the population proportion of 0.5 (based on $n = 40$ for simplicity).

Figure 3. Experiment 1 test trial response proportions by training condition separated for high and low accuracy participants in reference to a learning criterion, ≥ 0.75 average accuracy in the last two training blocks. Data from Medin and Schaffer's (1978) Experiment 3 are shown for reference.

Figure 4. Experiment 1 category A response proportions by condition and with mean block accuracy averaged across sets of five training blocks for all participants (errors bar are standard errors).

Figure 5. Exemplar model best fit predictions for the standard and low-differentiation conditions of Experiment 1.

Figure 6. Category summary sheet from Experiment 2 with instances from two categories corresponding to the abstract category structure for the standard condition in Table 1.

Figure 7. Average data for the explicit memorization task from Experiment 2 in terms of Category A response proportions, left panel. The last three panels are test results after 16 blocks, 25 blocks, and 32 blocks, respectively, of trial-by-trial feedback learning for the standard 5-4 category structure shown on the left in Table 1 (Johansen & Palmeri, 2002; Nosofsky et al., 1994; Palmeri & Nosofsky, 1995). The dashed line is the guessing performance reference at 0.5 because there were two categories.

Figure 1. Experiment 1 training condition accuracy averaged across sets of five training blocks for all participants (top panel) and for poorer learners (bottom panel) who did not ultimately achieve the learning criterion of ≥ 0.75 in the last two blocks of training (error bars are standard errors).

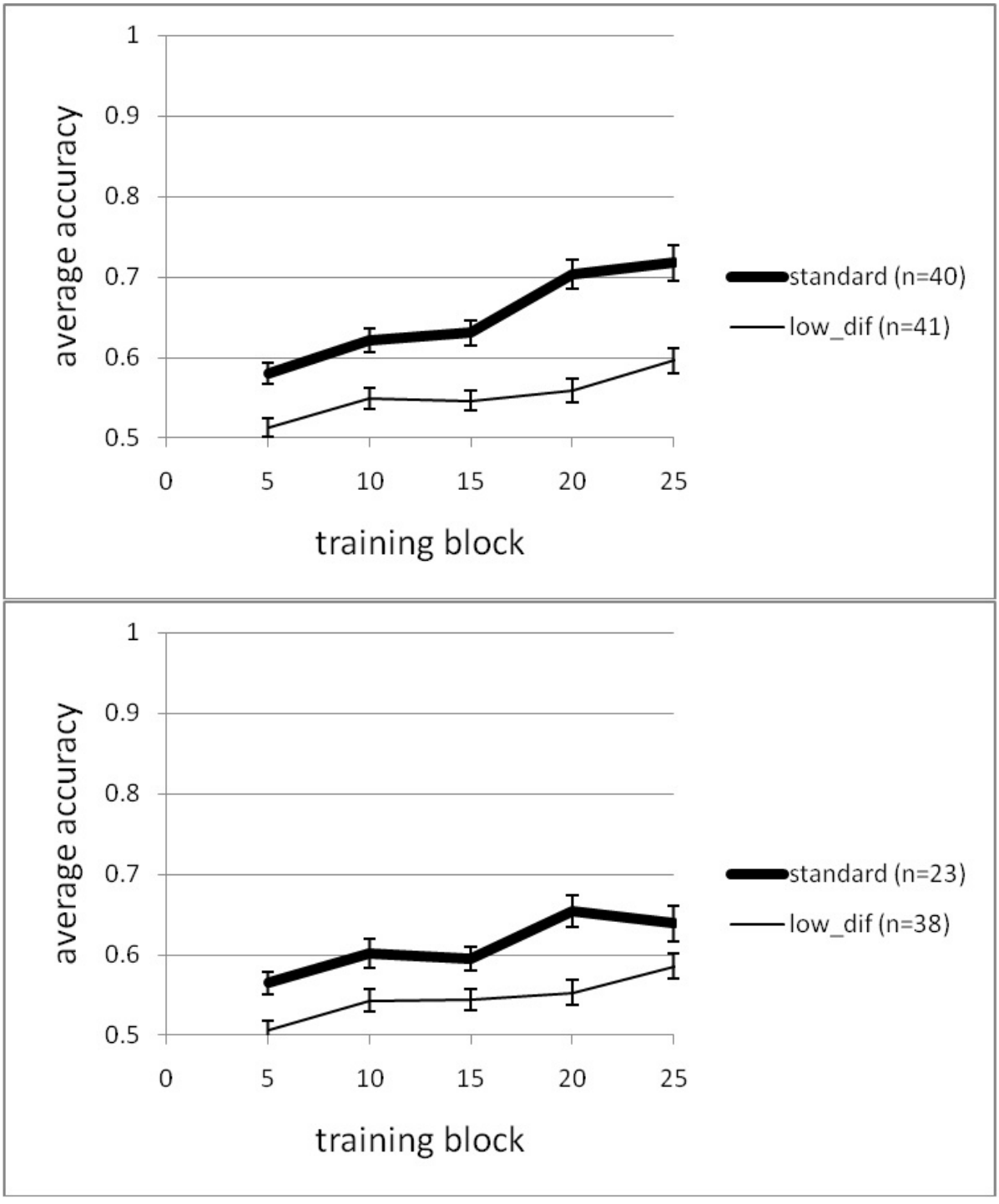

Figure 2. Experiment 1 test trial response proportions by training condition with an approximate 95% binomial confidence interval on the population proportion of 0.5 (based on $n = 40$ for simplicity).

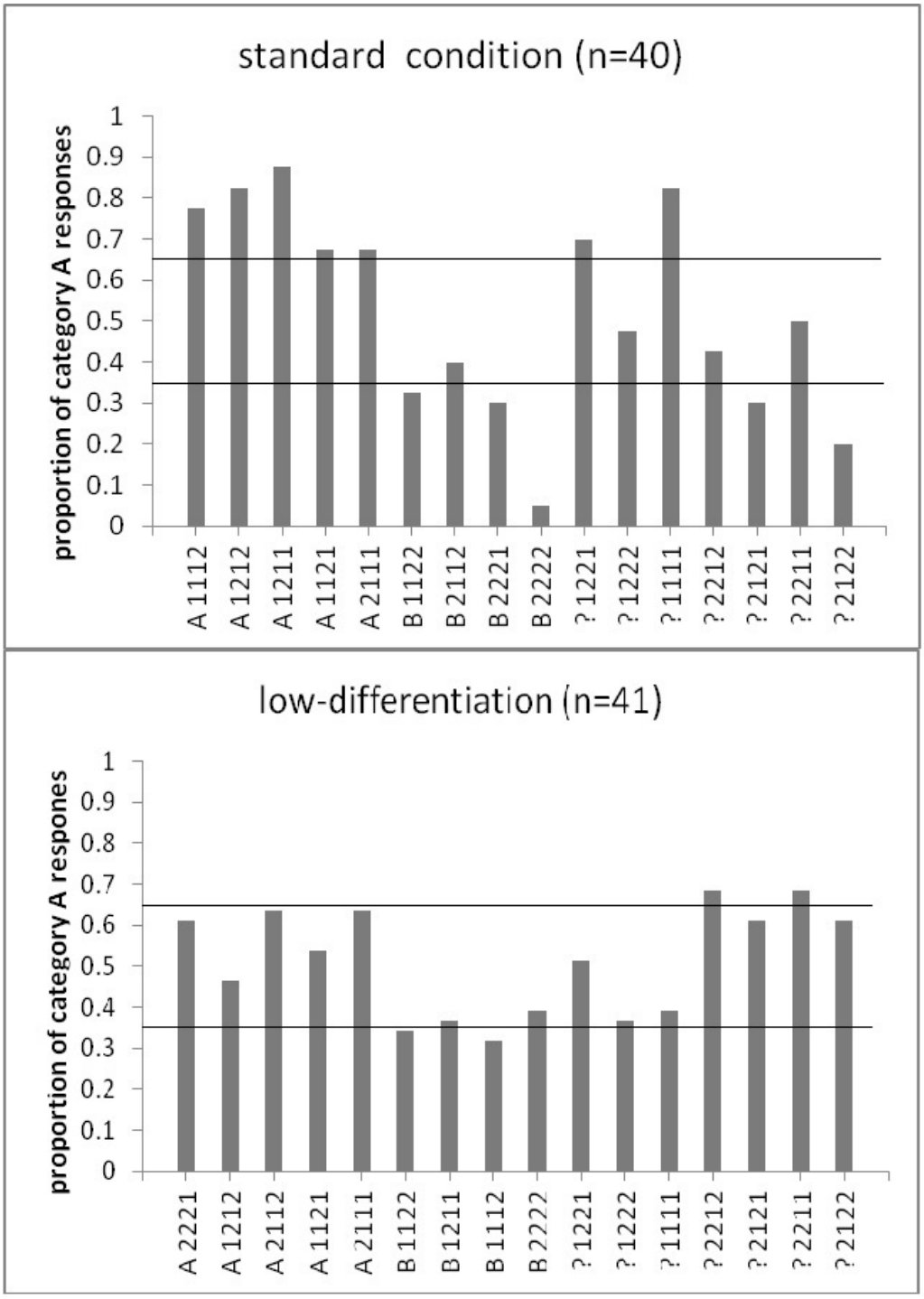

Figure 3. Experiment 1 test trial response proportions by training condition separated for high and low accuracy participants in reference to a learning criterion, ≥ 0.75 average accuracy in the last two training blocks. Data from Medin and Schaffer's (1978) Experiment 3 are shown for reference.

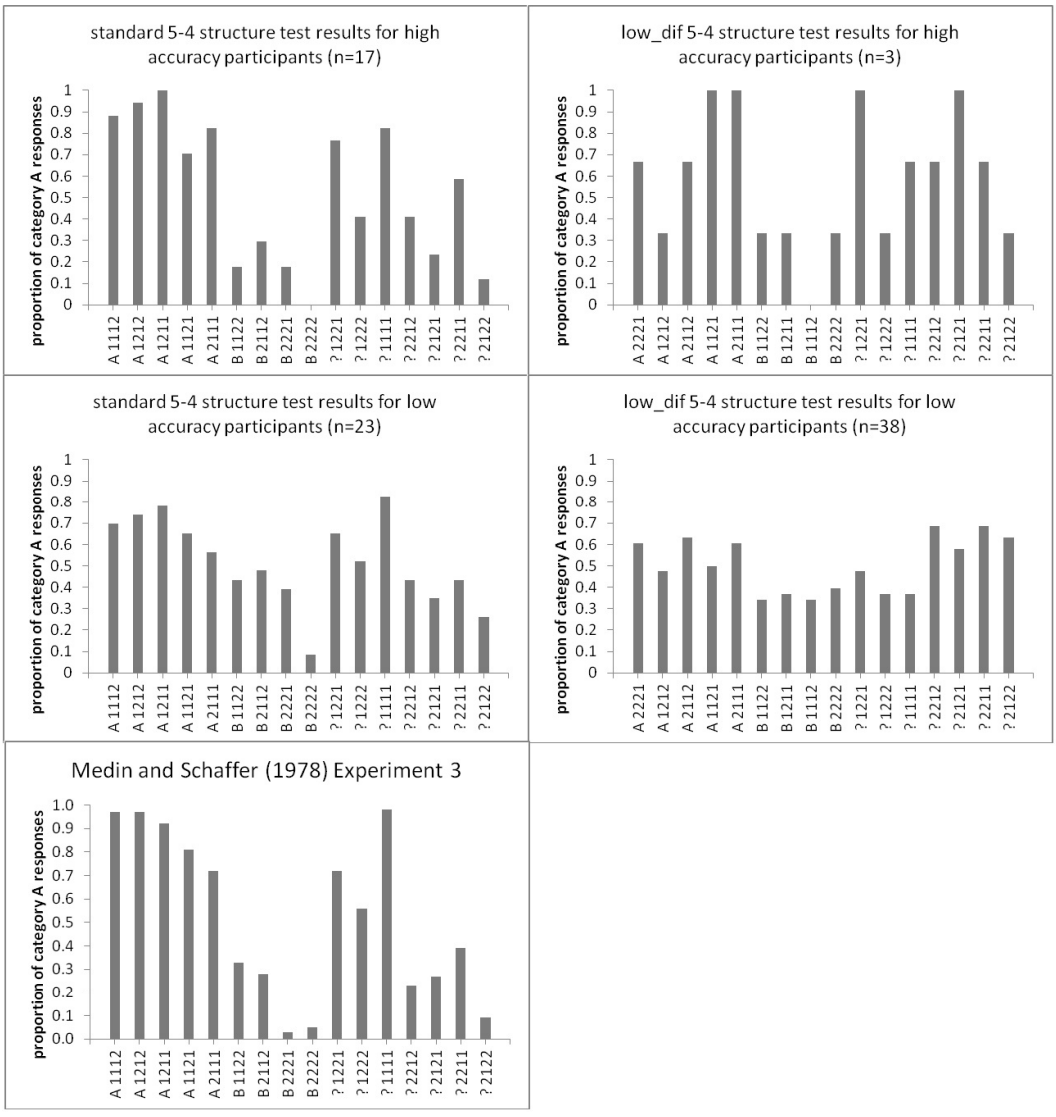

Figure 4. Experiment 1 category A response proportions by condition and with mean block accuracy averaged across sets of five training blocks for all participants (errors bar are standard errors).

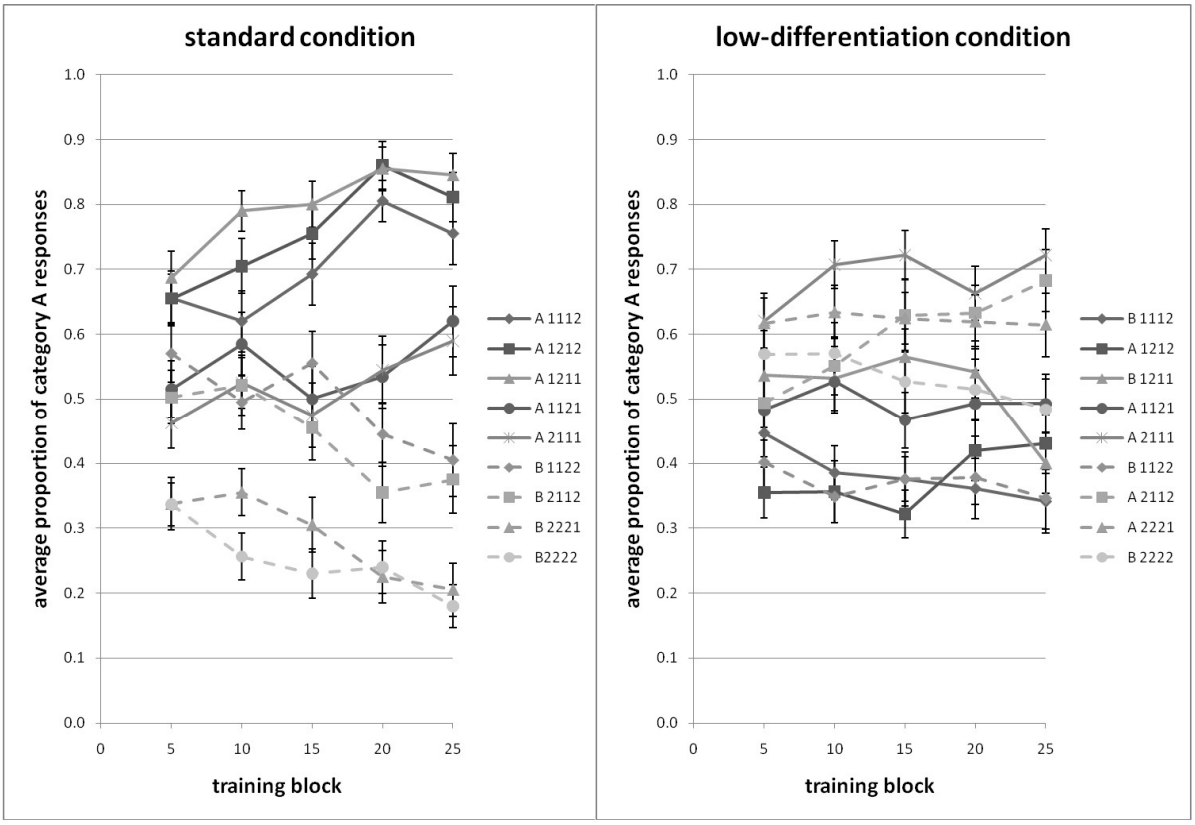

Figure 5. Exemplar model best fit predictions for the standard and low-differentiation conditions of Experiment 1.

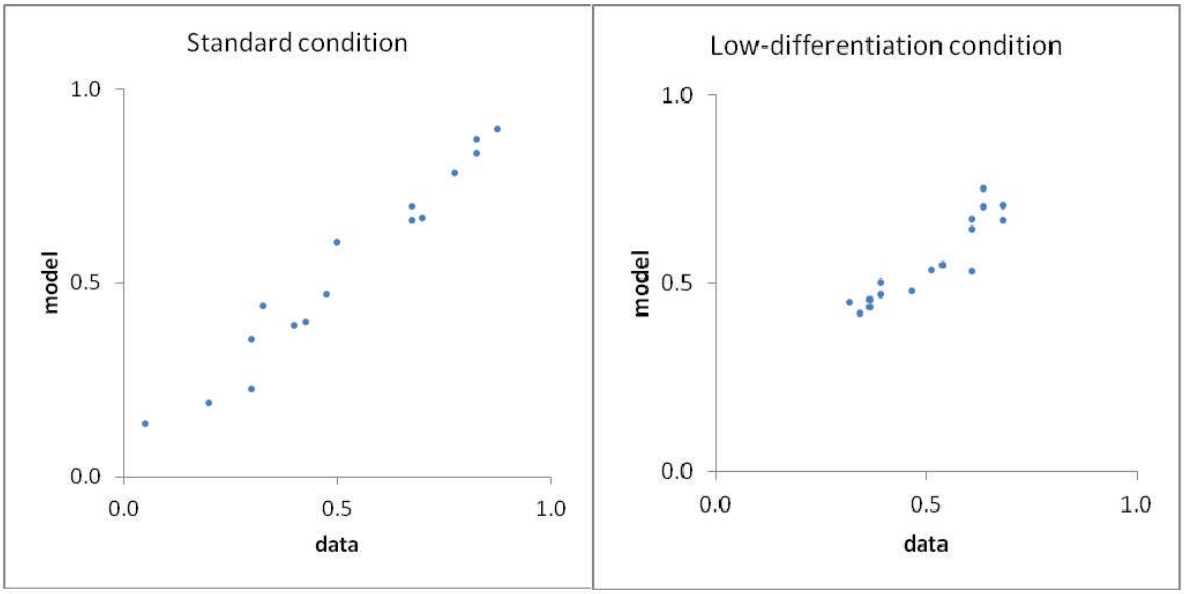

Figure 6. Category summary sheet from Experiment 2 with instances from two categories corresponding to the abstract category structure for the standard condition in Table 1.

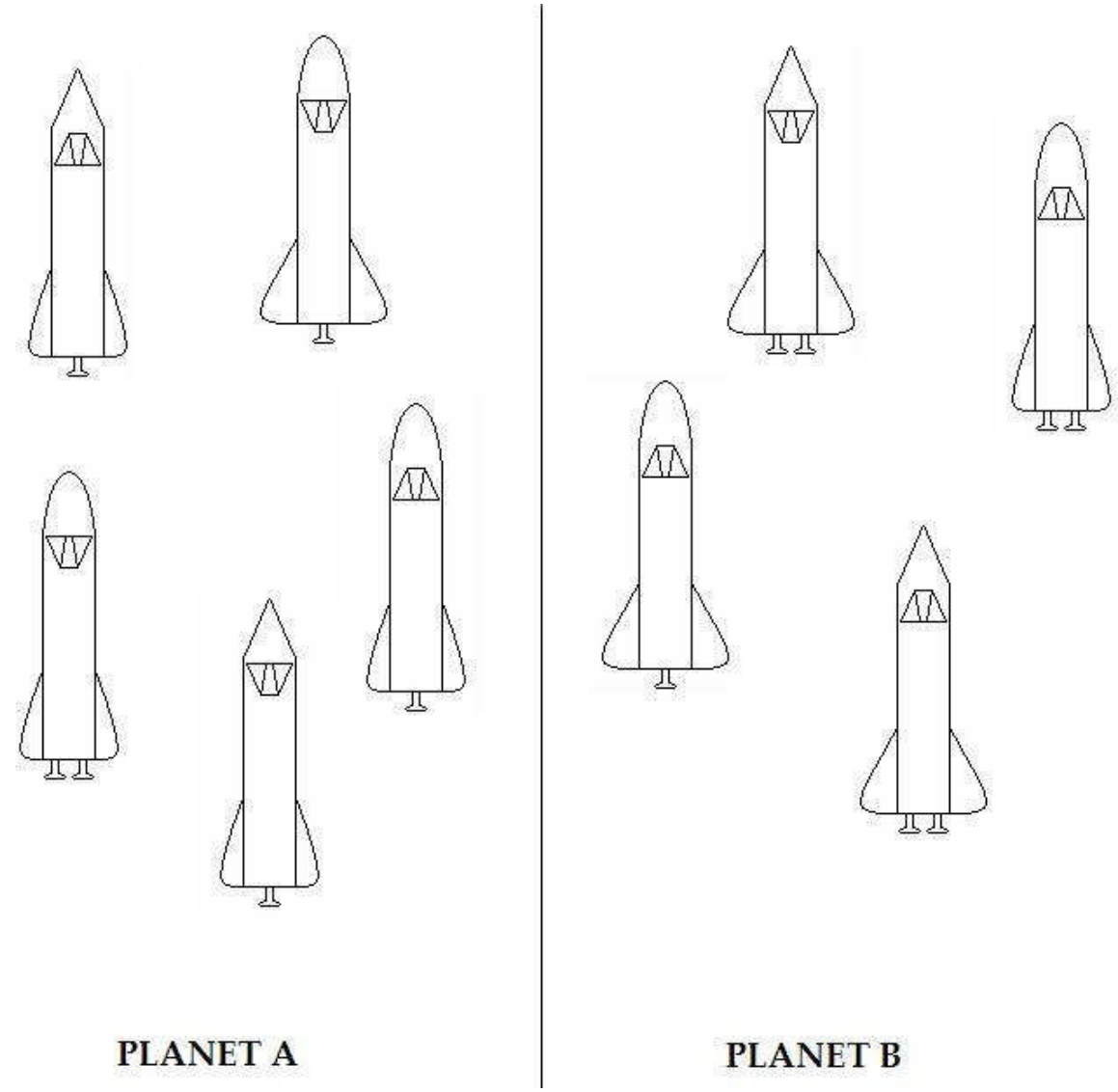

Figure 7. Average data for the explicit memorization task from Experiment 2 in terms of Category A response proportions, left panel. The last three panels are test results after 16 blocks, 25 blocks, and 32 blocks, respectively, of trial-by-trial feedback learning for the standard 5-4 category structure shown on the left in Table 1 (Johansen & Palmeri, 2002; Nosofsky et al., 1994; Palmeri & Nosofsky, 1995). The dashed line is the guessing performance reference at 0.5 because there were two categories.