Contingent necessity vs. logical necessity in categorization

Emmanuel M. Pothos

Department of Psychology, Swansea University.

Ulrike Hahn

School of Psychology, Cardiff University.

Mercè Prat-Sala,

Department of Psychology, The University of Winchester.

in press: Thinking & Reasoning.

Running Head: critical features; **Word Count** (including abstract): 6,808

Please address correspondence regarding this article to Emmanuel Pothos, Department of Psychology, Swansea University, Swansea SA2 8PP, UK. Electronic mail may be sent to e.m.pothos@swansea.ac.uk.

Critical (necessary or sufficient) features in categorization have a long history, but the empirical evidence makes their existence questionable. Nevertheless, there are some cases that suggest critical feature effects. The purpose of the present work is to offer some insight into why classification decisions might misleadingly appear as if they involve critical features. Utilizing Tversky's (1977) contrast model of similarity, we suggest that when an object has a sparser representation, changing any of its features is more likely to lead to a change in identity, than it would in objects which have richer representations. Experiment 1 provides a basic test of this suggestion with artificial stimuli, whereby objects with a rich or a sparse representation were transformed by changing one of their features. As expected, we observed more identity judgments in the former case. Experiment 2 further confirms our hypothesis, with realistic stimuli, by assuming that superordinate categories have sparser representations than subordinate ones. These results offer some insight into the way feature changes may or may not lead to identity changes in classification decisions.

Introduction

How concepts are represented is an issue of fundamental importance in understanding human psychology. Many categorization accounts utilize some notion of similarity. For our purposes, the key characteristic of such accounts is that they predict *gradedness* in the way new instances are classified, in judgments of typicality, in assessing category membership etc. (e.g., Barsalou, 1985; Hampton, 1995; Nosofsky, 1991; Pothos, 2005; Smith and Minda, 1998). According to a radically different proposal, there are critical features (or a belief in such features), the presence or absence of which determines category membership, independently of overall resemblance. In other words, critical features automatically determine category membership, regardless of any other information. For example, consider the rule 'if an insect has yellow and black stripes, then it must be a bee'. This means that when seeing any insect with yellow and black stripes, we are compelled to conclude, by logical inference (modus ponens), that the creature we are dealing with is a bee. Category judgments that involve critical features are absolute and inflexible (e.g., the above rule is wrong; what about wasps?). However, from a cognitive efficiency point of view, they do allow quick, specific decisions. Do critical features exist, at least for some concepts? If yes, human categorization should be understood in terms of at least two, qualitatively different, systems: a system based on rules and a system based on similarity (note that the putative role of logical inference as would be implied by such rules has been criticized in other areas of cognition; e.g., Chater & Oaksford, 1993; Oaksford & Chater, 1994; regarding the distinction between rules and similarity, see, e.g., Hahn & Chater, 1998; Pothos, 2005; Sloman & Rips, 1998).

Rips (1989; Keil, 1989) presented his participants with a story of a bird that suffered from exposure to toxic waste. As a result, the appearance of the bird changed

to such an extent that it looked like an insect. This was unfortunate for the bird, however, it still managed to mate with other birds of its kind and the offspring looked normal. The mean categorization ratings provided by Rips's participants indicated the changed bird to be more likely to be a bird. Therefore, 'mating' might be considered a *sufficient* feature for the category of birds, that is, a feature that guarantees classification into the category of birds. However, things might be birds even without an ability to mate with other members of that species. *Necessary* features, by contrast, are essential for an object to be considered a member of a category; for example, being 'male' is a necessary feature for the category of bachelors. Necessary and sufficient features can be called 'critical' features, to indicate that their presence can critically affect a classification decision. This finding of Rips has been widely considered as evidence for the existence of critical features (for an overview of related research see Rips, 2001).

Note that many researchers have advocated factors that could determine the relative importance of features. For example, Sloman, Love, and Ahn (1998; cf. Rehder & Hastie, 2004) suggested that numerous causal links between a feature and other features increase its importance and used a measure of feature mutability (whether one feature can change without changing the other features) to determine feature importance. Corter and Gluck (1992; Gosselin & Schyns, 2001) provided probabilistic measures of a feature's utility in, e.g., determining category membership. However, postulating features that can be 'very important' is a very different proposal from postulating (critical) features that can logically determine category membership; the influence of the former in a classification decision can be reduced, if there is compelling evidence, but not the influence of the latter.

A variant of the critical features theory is psychological essentialism: according to this approach entities do not actually possess critical features, but people behave as if they did—and so sometimes make classification decisions consistent with a belief in critical features (e.g., Malt, 1990; Malt et al., 1999; Medin & Ortony, 1989). For example, the most fundamental feature of a living thing is its DNA; but the majority of people will have never observed a creature's DNA. The problem with essentialism as a psychological theory is that, if essences are not known (or do not even exist!), it seems difficult to construct well-controlled supporting experiments (for attempts see Gelman, 2003, Hampton, Estes, & Simmons, 2007; Pothos, Hahn, & Prat-Sala, 2009).

In our research, we have tried to examine the robustness of Rips's result and hence the evidence for critical features. Pothos and Hahn (2000) presented a number of variations of Rips's main scenario. One condition involved describing a creature that looked like a crow and could mate with other crows (normal offspring), but was in fact a space alien. Most participants considered this crow-like creature to not be a crow. Since the features of the space-crows were identical to the features of the crows, and the space-crows were *not* considered crows, mating (or any other feature) could not have been a critical feature for the category of crows. Therefore, Pothos and Hahn's (2000) finding was against critical features. Since Murphy and Medin (1985), several researchers have advocated the importance of naïve theories in categorization, and this finding appears to support such an approach over and above critical features (and possibly essentialism as well).

Such research casts doubt on the existence of critical features. Other researchers have also failed to find consistent evidence for critical features. For example, Larochelle, Cousineau, and Archambault (2005) observed that reported

necessary features were not actually treated as such in decisions about category membership. Also, Hampton, Estes, and Simmons (2007) argued that high variance in individual participant performance could have partly accounted for the dissociation Rips reported and also found that changing particular aspects of the stimuli Rips used eliminated essentialist classification. However, researchers have often reported effects that *look* like critical feature effects (for example, Rips, 1989). Our purpose is to shed light on such effects. Specifically, under what circumstances will a categorization decision appear as if it involves a critical feature, even if it does not?

In Rips's study, it is possible that 'mating' is not a critical feature in general, but merely seems like one when few other features are present. A general similarity model, Tversky's (1977) contrast model, readily predicts that the fewer the features in the representation of an object, the more likely it is that changing a single feature will alter the object's identity. For example, consider stimuli composed of four features, that we indicate as $A_1A_2A_3A_4$ or stimuli composed of eight features, indicated as $A_1A_2A_3A_4A_5A_6A_7A_8$. In both cases we change one feature, so that the changed stimuli can be represented as $A_1A_2A_3B$ or $A_1A_2A_3A_4A_5A_6A_7B$. Everything else being equal, we believe that the feature change from A_4 to B is more likely to look like a critical feature compared to the feature change from A_8 to B. In other words, we suggest that a change in object identity will be less likely when many other features are present and intact (a case of rich or complex representation) and more likely when few other features are present (a case of sparse or simple representation). For example, a straightforward application of Tversky's (1977) ideas in the above example would suggest that in the sparse case similarity is proportional to 3 (common features) – 1 (different feature) = 2, while in the complex case 7 –1 = 6. Therefore, the similarity between the original and the changed object would be greater in the complex case

compared to the sparse case and so we predict more judgments of ‘identity’ in the complex case. Note that Tversky (1977) provided several empirical demonstrations that complexity impacts similarity in this way.

This raises a question in relation to Rips’s results: Given that Rips did observe results which appeared to indicate that ‘mating’ was a critical feature, are we forced to conclude that the representation for the birds/ insects in Rips’s experiment was sparse (so that the change of a single, prominent feature looked like a chance in a critical feature)? This is likely because participants in Rips’s experiment had no information about the presented bird, other than that it could still mate with other birds after the accident. That is, the presentation of information in the experiment possibly encouraged a sparse representation of the bird in Rips’s story. Having said this, as noted above, the conceptual representation of natural kinds is a complicated issue in that, for example, classification of natural kinds may be (partly?) driven by a belief in essences—in the relevant Experiment 2 reported here we restrict the examination to artifacts.

Hampton (1995) presented an argument broadly similar to ours. He showed how highly weighted features in a prototype can look like defining features, in that, unless a test instance has these features, it is impossible for the test instance to be similar enough to a prototype. In other words, Hampton’s suggestion was that critical feature effects could be understood in the context of a similarity-based theory of categorization.

Experiment 1 is a straightforward examination of our idea with artificial stimuli, required to validate the experimental design. Experiment 2 is the critical test of our hypothesis. In Experiment 2, the same objects and corresponding feature changes are described to participants. However, in one condition the objects are

described in a superordinate way and in another in a subordinate way; this is a manipulation which affects the number of category features. Will exactly the same feature change look like a critical feature change in the superordinate case, but not in the subordinate case? If yes, then we will have provided a simple similarity account of how categorization decisions, which look like they involve critical features, can occur.

Experiment 1

In this experiment we examine the intuition that feature changes are more likely to be perceived as critical if the corresponding object representation is sparse. In other words, we predict that $A_1A_2A_3A_4$ and $A_1A_2A_3B$ are less likely to be considered to belong to the same category than $A_1A_2A_3A_4A_5A_6A_7A_8$ and $A_1A_2A_3A_4A_5A_6A_7B$, as would be predicted by Tversky's (1977) feature contrast model.

Participants and design

One-hundred undergraduate students from Bangor University or Winchester University took part in the study for a payment or course credit. The number of participants in the two conditions of the study (between participants design), that we shall call simple (sparse) and complex (rich), were 51 and 49 respectively. The two conditions were identical but for the complexity of the stimuli used: in the simple condition, each of the stimuli we employed was comprised of four features, while in the complex one of eight features.

Materials

Each participant in both the simple and the complex condition saw six training items and six test items. All participants saw the same training items. However, there were

three different sets of (six) test items each, corresponding to counterbalancing the form of the test items (described shortly). All training items had the same features. However, in different items the *texture* of the features varied. For example, in one item the texture for a feature might be a continuous thick line, while in another item it would be a dotted line. As there were six training items, each feature was instantiated with six different textures. The rationale for adopting this approach is that we wanted to model a situation in which there would be different instantiations of the same features. For example, think of a dog: different dogs have different legs, ears, tails etc. Likewise, the training items in this experiment were meant to have the 'same' features, but instantiated somewhat differently. Of course, there is no single perfect design to capture such an intuition, however, the consistency of the results across participants suggests that our assumptions were broadly valid.

The form of each training item for the simple condition was a square enclosed by a circle with two triangles on the side of the circle and a cross on top of it. The training items in the complex condition were analogous and were created by adding four features to each of the items in the simple condition. In this way, the simple and complex conditions are as comparable as possible. The additional features used for the complex items were a diamond beneath the circle. The diamond enclosed a star and had two rectangles on its sides. An ellipse was attached below the diamond. The features used for both the simple and complex condition were selected so that intuitively they were roughly equally salient (this was informally assessed by independent judgments from the authors). Examples of the stimuli are shown in Figure 1.

Each test item was identical to a training item but for a single feature. Recall, there were three sets of test items; these corresponded to a (partial) counterbalancing

of which features were changed. In the first set of test items for the simple condition we created six stimuli that were identical to the training items except that there was no square in the middle but rather a whirling pattern; all six test stimuli had a whirling pattern center and for each stimulus the texture of the whirling pattern was different. Two more sets of test items were generated in an analogous way, by changing different features, so as to accommodate for the possibility that participants would perceive the change of a particular feature as more important than the change of other features. The three sets of test items for the complex condition were created simply by attaching the extra four features to the test items for the simple condition (Figure 1).

To clarify, each item had a 'simple' version and a 'complex' one. The difference between the two is that in the latter case four additional features were added. However, crucially, in both cases the same feature was changed. In other words, the simple and complex conditions were completely matched in terms of which features were changed. Finally, each stimulus in both conditions was printed individually on an A4 sheet.

Procedure

Participants received printed instructions informing them that they were about to see a set of objects, all of which belonged to the same category, which we called the category of 'Chomps'. They then received the training items in a folder and were allowed to observe them in any way they wished. Once they had done so, they were given new instructions telling them that they would shortly see another set of items and that they would have to decide which of the new items belonged to the same category as the training ones. Participants indicated their response for an item by

writing it down on the sheet on which the item was printed. Participants were tested individually and the experiment lasted about five minutes.

Results

We were interested in participants who considered the test items to be primarily in the same category as the training ones vs. participants who did not do so (i.e., more test items classified as 'Chomps' vs. 'Non-Chomps'), since the former could be assumed to be categorizing (primarily) on the basis of similarity and the latter (primarily) on the basis of a critical feature. Some participants classified the same number of test items as 'Chomps' and 'Non-Chomps'. Such intermediate responses are neutral with respect to the hypotheses of interest, and so were eliminated from this analysis. There were nine intermediate responses in the simple condition and seven in the complex one, so that the results of 16 participants (out of 100) were not considered. Participants were more likely to consider most of the test items as 'Chomps' in the complex condition (many features) than in the simple one (few features): $chi^2(1)=8.64$, $p=.003$ (Table 1).

----------Table 1----------

The above analysis allows a quick appreciation of the pattern of results. An alternative analysis, which does not lead to the exclusion of any participants, is possible by assigning a score of 0—6 to each participant, depending on how many test items he/she considered in the same category as the training items (note that this information was not available for one participant in the simple condition). The scores of participants in the simple and complex conditions was then compared with a *t*-test, which was found highly significant: $t(97)=3.276$, $p=.001$. The mean score of participants in the simple and complex conditions were 1.64 (SD=2.05) and 3.10

(SD=2.38) respectively, showing that participants in the latter condition made more 'Chomps' (positive) responses.

Discussion

We used artificial stimuli composed of four or eight features and investigated how participants categorized these stimuli when we changed one of their features. Overall, participants were reluctant to consider the test items in the same category as the training ones (for both conditions together only 31 participants considered the majority of test items to be 'Chomps', compared to the 53 who did not). Importantly, however, participants were more likely to consider the test items as belonging to the same category as the original ones, with the complex items as opposed to with the simple items, consistently with our prediction and the application of Tversky's (1977) model of similarity. The converse, null hypothesis would have been that changing some features would result in changes in category membership, while changing other features would not, regardless of how many other features were present. If that were true, then there would have been no systematic differences in the proportion of object identity judgments between the simple and complex conditions.

One can ask whether individual feature salience might undermine confidence in our conclusions. It is hard to see how this is possible. For example, we partly counterbalanced the changed feature within each of the two conditions. Also, in both the simple and the complex condition the same features were changed. Given that across these manipulations participants' performance indicated a consistent preference for 'same' judgments in the complex condition, we conclude that our methodological assumptions were valid. Another issue is whether the instructions might have biased participants to seek an equal number of 'Chomps' and 'Non-Chomps' classifications

in test. However, this was not the case. For example, in the simple condition, there were 27 participants who selected no test items as Chomps, 6 who selected 5 or 6 items as Chomps, with the rest in-between. So, apparently, participants did not feel constrained to select an equal number of test items as Chomps and Non-Chomps.

Experiment 2

With artificial stimuli there was no reason to expect a violation of the predictions from Tversky's (1977) model, because for artificial stimuli there would be little basis to consider a particular feature as critical. With realistic stimuli, by contrast, we often have strong intuitions that certain features might be critical ones. Accordingly, eliminating a critical feature should *always* lead to a change in the object's identity, regardless of whether the representation of the object is manipulated to be sparser or more extensive. Thus, Experiment 2 is the main test of our account of how effects mimicking critical feature effects can arise.

How can we determine whether an object's representation is sparse or complex? Asking participants to provide feature lists may be good for determining characteristic or diagnostic features, but possibly less appropriate for identifying *all* the features relevant to a representation. Moreover, the representation for a concept is bound to be different for different people. For example, if a person is unfamiliar with a concept, her representation of that concept would plausibly be sparse, even if averaged data suggests otherwise. In other words, we required a way for determining certain concept representations to be sparse, while others complex, for each one of our participants.

We made the minimal assumption that the more general a concept, the fewer the features the objects within this category have in common, simply because the

more abstract or general a concept the more diverse the range of objects that comprise it (Komatsu, 1992). That is, more general concepts should be represented with fewer features relative to *corresponding* more specific ones (cf. Rosch et al., 1976). Thus, if two concepts are related to each other as subordinate and superordinate, then the latter will necessarily have a sparser representation relative to the former. Consider, for example, the concept of 'vehicle' in relation to the concept of 'car'. The former concept includes as members bicycles, buses, trains, motorcycles etc., as well as cars. For such a diverse range of objects to cohere together, the concept representation has to involve relatively few features. For example, while all vehicles are used for transportation, they could have any number of seats, accommodate different numbers of passengers, etc. By contrast, the members of the concept of car have several common features with each other; in other words, the concept car will have a more extensive representation compared to the concept vehicle. These ideas are illustrated in Figure 2.

----------Figure 2----------

In related previous work, Hampton (1982) found violations in transitivity in class inclusion relations; in other words, he identified cases such that instance X would be a member of Y, Y or Z, but X would *not* be a member of Z. Concept X would be most specific, Y of intermediate specificity, and Z the most general. Hampton's result is analogous to the one we are trying to obtain, but for a crucial difference. We are interested in whether a classification of an object would change as a result of changing a particular (assumed critical) feature. Hampton did not employ any feature changes and, indeed, he was only interested in identifying *some* concepts for which a violation of transitivity could be observed. By contrast, we sought to identify an effect that would *consistently* apply in a group of stimuli.

We decided to use a range of simple artifact concepts. If artifact categories have critical features, then these should correspond to their intended function (Bloom, 1996; note, Estes, 2003). For example, the category of hammers would not exist without the notion of driving nails into wood—it appears that many artifact categories are created from a need to express certain functions.

Participants

One hundred and twenty two experimentally naïve participants, all second year psychology undergraduates at Swansea University, took part in the study for course credit.

Materials

We identified everyday objects with a view to alter one of their features that we deemed necessary for the way they are categorized. For example, for an object to be a hammer, the object must be sturdy enough to enable the applications a hammer is typically used for. The validity of our assumptions regarding which features are important for object identity was assessed post-experimentally, in terms of whether eliminating these features resulted in changes in object identity.

An important issue in the design of the materials is to ensure that the features that are to be changed have equal relevance to both subordinate and superordinate categories. The key assumption is that, if critical features exist, then subordinate and corresponding superordinate categories must have analogous critical features. In other words, if a feature can be considered as critical for a subordinate category, then a correspondingly more general feature would be likewise considered as critical for the parent superordinate category. For example, *assume* that the critical feature for a

‘hammer’ is its function, namely its capacity to be used to drive nails into wood. We can also reasonably suggest that the critical feature for the same hammer categorized at the superordinate level as a ‘tool’ is this functional capacity. In this case the capacity is now an instance of the general tool function of ‘being used to build or make things’. Therefore, if the critical features assumption of categorization is correct, then a hammer-like object that turns out not to be a hammer, cannot still be a tool, or vice versa, since in both cases effectively the same critical feature would be missing.

Of course, one could change a feature at the subordinate level in a way that leads to an alternative subordinate level classification under the same parent superordinate category. For example, one could imagine a tool like a hammer, but which has a pointy head, rather than a flat head. Such situations are trivial, since all they show is that we can have several different subordinate level categories under the same parent superordinate category. We do not consider such situations in the present experiment.

A question is how we could support the assumption that critical features at the superordinate and corresponding subordinate category levels are equivalent. Empirically, if the critical feature hypothesis is wrong (as it turned out to be the case), then we expect that the importance of any particular feature will be contextual: If there are many other features present, changing a particular feature may not be very significant, but if there are few other features present, then any particular feature change may lead to a change in object identity. We think the most convincing route for addressing this issue is arguing a priori that, *if critical features exist*, and if a feature can be assumed to be critical at the subordinate level, then a closely analogous feature can be assumed to be critical at the superordinate level.

Table 2 shows the 11 objects we used and the actual question description. The critical column is the ‘changed feature’ one. It can be seen that the changed feature readily implies a sensible critical feature at the subordinate possible classification, and a directly analogous critical feature at the superordinate classification. For example, for the piano/ musical instrument pair, the assumed subordinate/ superordinate features are the ability to produce music of a particular kind/ the ability to produce music; for the football/ equipment pair, the features are used in soccer/ used in a sport; etc. Note, finally, that in specifying the subordinate/ superordinate pairs, we took care to ensure that the category terms we considered superordinate were fully inclusive of the corresponding subordinate ones (i.e., all the members of the subordinate concept were a subset of the members of the corresponding superordinate one),

----------Table 2----------

The objects were described either in terms of a subordinate category (e.g., a hammer would be described as a ‘hammer’) or a superordinate one (e.g., a hammer would be described as a ‘tool’). Note that we are not interested in whether the description we provided for our stimuli corresponds to a basic level categorization or not, but rather in whether the two category terms are such so that one is subordinate to the other.

The actual materials used in the experiment had the form of a series of the 11 stimuli printed individually on A5 sheets. The order of stimuli was randomized for each participant. Each stimulus consisted of a line or two describing it. For example, for the currency stimulus the description we provided was ‘After a series of financial disasters, the economy of a country collapses so that the country’s currency is worthless, and cannot be used to buy anything.’ We tried to make the descriptions as

concise and specific as possible. Below the description, participants were prompted to decide whether the stimulus belonged or not to the category we were interested in. Continuing with the same example, (different) participants would read ‘Are items of the currency still money?’ or ‘Are items of the currency still coins?’ ‘Yes’ and ‘No’ boxes were printed next to the questions so that participants could indicate their response. The stimuli were further illustrated with a picture in the cases of the doll, the football, the hammer, the piano, and the fakirs’ bed. In the other cases, it was not considered necessary to provide a picture, or it was not possible to identify an appropriate intuitive picture.

Procedure

We wanted all participants to make some subordinate and some superordinate categorization decisions, to avoid any confound that might arise from individual biases regarding more general or more specific categorizations. Also, we were reluctant to ask the same participant to determine the categorization of a stimulus at both the superordinate and the subordinate level, since the outcome of one judgment might bias the other judgment. Therefore, the approach we adopted was to request each participant to make about half the categorization decisions at the subordinate level and about the other half at the superordinate level. So, for example, if a participant had to decide whether the piano stimulus was a ‘piano’ (subordinate) she would not be asked to decide whether it would be a ‘musical instrument’ (superordinate) as well. Each participant received a booklet with the 11 stimuli. Approximately half the participants were asked to make subordinate classifications for the items ‘loudspeaker’, ‘MS Word’, ‘bed’, piano’ and ‘hammer’ and

superordinate classifications for ‘car’, ‘football’, ‘figurine’, ‘tunnels’, ‘box’, and ‘money’; vice versa for the rest of the participants.

Participants were told that they would receive some A5 sheets with short descriptions for a series of items and that they would have to make a classification decision for each of these items. The materials presented the items and prompted the participant for a response for each item, so that after the initial instructions no further interaction was required between the experimenter and the participants. To clarify, all participants saw the same item descriptions but different participants were asked to make a different combination of (superordinate or subordinate) classifications for the items.

The experiment was run at the start of a psychology class. Participants were provided with the instructions and then they received a little booklet with the items. They were told to mark their responses and give the booklet to the experimenter, before leaving the lecture theatre. Participants were not rushed to finish their responses (those who did not want to participate simply did not return the booklets). The experiment lasted for about five minutes.

Results

There were four missing responses out of the total of 122*11=1342 responses. The objective of the analyses is to examine the hypothesis that an item would be more likely to be endorsed (that is, accepted as a category member) at the subordinate level (rich representation) compared to the superordinate level (sparse representation). We conducted an item-based analysis and a participant-based one. The conclusion from both analyses was the same and supports our hypothesis.

In the item-based analysis, we compared, for each item, the percentage of times it was endorsed at the subordinate level with the endorsement rate at the superordinate level (Table 3). For example, how often was the piano item accepted as a 'piano' and how often was it accepted as a 'musical instrument'? Recall that for a particular item different participants would make the subordinate classification from the participants who made the superordinate one. A Wilcoxon Signed Ranks Test for paired samples comparing mean endorsement rates at the superordinate and subordinate levels was significant (Z=1.956, p=.05, two-tailed). The choice of a non-parametric test is justified in this case, since there was no a priori reason to expect that the mean endorsement rates for different items would be uniform: the mean endorsement rate for different items would depend on the perceived importance of the feature change. As Table 3 shows, in all cases the endorsement rate at the subordinate level was higher than the endorsement rate at the superordinate level, as hypothesized, with two exceptions: in the case of the 'fakir's bed' item, endorsement at the two levels was nearly equal. Also, in the case of the 'MS Word' item, endorsement at the superordinate level was much *higher* compared to endorsement at the subordinate level (0.80 vs. 0.24). Informal debriefing indicated that participants had misunderstood our description for this item: they considered the computer virus to be an instance of 'software', even if the malfunctioning word processor may no longer be classifiable as 'software'. Note that when eliminating the 'MS Word' item the comparison of the endorsement rates between superordinate and subordinate levels is significant at the .007 level (Z=2.701; same test as above).

In the participant-based analysis, we computed the average endorsement rate of each participant at the superordinate and subordinate level. Recall that each participant made six or five classification decisions at the superordinate level and six

or five classification decisions at the subordinate level. The average endorsement rate at the subordinate level was .47 (SD=.23) and at the superordinate level .38 (SD=.21; an endorsement rate of 1 implies that a participant endorsed all items as members of their respective categories), a difference which was significant with a paired-samples *t*-test: $t(121)=3.448$, $p=.001$. Note that in this case there would be no justification to employ a non-parametric test, since, under the experimental hypothesis, all participants ought to be more likely to endorse an item at the subordinate level, compared to the superordinate level.

----------Table 3----------

Discussion

It appears that when the same feature is changed in a superordinate category (fewer features) it is more likely to lead to a category change for an object, than when it is changed in a subordinate category (more features). This result is consistent with our suggestion, that a single feature change is more likely to *look* critical when a representation is sparser, compared to a situation when it is richer. In Experiment 2, we are therefore led to the same conclusion as in Experiment 1, despite the differences in methodology and materials.

At the subordinate level, around 53% of object classification decisions indicated a change in object identity as a result of the feature change (at the superordinate level, this percentage was about 62%). This indicates that the feature we changed for each object was considered important for this object's identity by many of the participants, thus partly validating the design. Note that different participants are likely to consider different features as important for an object's representation (cf. Larochelle, Cousineau, & Archambault, 2005). Equally, the a priori salience of the

different features which were assumed critical might vary between participants. Consider what would happen even if there were huge differences in the salience of the assumed critical features (noting that we do not think this is the case, based on the obtained results): if an assumed critical feature was completely not-salient, then changing it should not reduce the identity judgments, regardless of whether the corresponding object was described in a superordinate or subordinate way. If an assumed critical feature was extremely salient, then changing it might lead to uniformly 'different' judgments, also regardless of whether the corresponding object was described in a superordinate or subordinate way. Therefore, at the very worst, differences in feature salience might increase the likelihood of the null hypothesis, but, crucially, this does not affect the validity of the dependent variable (which was the difference in identity judgments, depending on whether an object was described in a superordinate or subordinate way). Table 3 shows the item-by-item endorsement rates at the subordinate and superordinate level. While there were differences in whether the transformed objects were considered to be in the same category as the original ones, in nearly all cases the endorsement rate at the superordinate level is lower compared to the subordinate level (a notable exception concerns the item 'MS Word', but, as noted above, it appears that participants interpreted this item in an unintended way).

Another potential problem with Experiment 2 is this: suppose that classification has nothing to do with sparsity but rather with the ease in which participants can imagine alternative classifications for an object. For example, suppose that participants produce more 'same' responses in relation to subordinate classifications rather than superordinate ones, simply because it is less easy to imagine alternative subordinate classifications compared to alternative superordinate

ones. For example, an object that is described as a hammer does not really have any alternative classifications. However, an object which is described as a tool might conceivably be described as a toy as well (we thank an anonymous reviewer for this suggestion). In general, while there are certain superordinate classifications which are widely applicable (such as toy), this is not universally true. For example, what is an alternative superordinate categorization for currency or furniture? It seems as difficult to imagine alternative classifications with respect to such superordinate categories as it is for their respective subordinate ones. Therefore, we do not think that this is a confounding factor in Experiment 2.

A statistical concern regarding the analytical approach we adopted in Experiment 2 relates to whether we are justified as considering participants' decisions for different items as independent. The assumption of independence would be supported by the fact that the items were unrelated to each other. Not making this assumption would imply that we believe that if a participant responded 'not a category member' for one item he/she was more likely to respond 'not a category member for another item'. However, this seems implausible (cf. Hampton et al., 2007).

Finally, we can consider whether if one was asked to classify a soft rubbery thing that looked like a 'hammer', what alternative is there but to call it a 'hammer'? Crucially, we did not ask participants to produce a name for the changed object, but rather examine whether it should be classified in a particular category or not. There are countless instances of items which we decide not to classify into one of our existing categories, but for which we do not have an alternative classification readily available.

Overall, a simple explanatory framework, based on Tversky's (1977) theory of similarity, proved sufficient to account for when feature changes are more likely to lead to changes in object identity and when they are not.

General discussion

There has been an increasing consensus against critical features in concept representation. However, effects which *seem* like critical feature effects do exist and beg the question of how they occur. The purpose of this work was to provide some clarification along these lines, by modifying Rips's (1989) experimental paradigm and utilizing a standard theory of similarity (Tversky, 1977). Our suggestion was that certain feature changes might seem critical, because the object representation is sparse (so that changing any particular feature might trigger a change in identity), but not if it is complex (so that when there are several features present, changing any one of them would have a relatively weak effect; cf. Hampton, 1995). In Experiment 1, we confirmed these expectations with schematic, artificial stimuli. Experiment 2 provided the more compelling test of our hypothesis, with real stimuli.

The null hypothesis in this investigation was that a feature change is considered either important (and hence leading to a change in identity) or unimportant, regardless of whether an object's representation is sparse or complex. However, in both experiments we found that (the same) feature changes were more likely to lead to an identity change when the object representations were sparse. Thus, our findings provide some insight into both the way critical feature effects arise and also into the nature of object representation and similarity theory. Of course, this is not to say that research like ours 'proves' that critical features do not exist for all concepts. Clearly, there are practical constraints in the range of concepts which can be

considered in any specific study and one cannot preclude the possibility that there are specific (artifact) concepts for which category membership is defined by critical features.

Another limitation in the generality of this conclusion is that we considered only artifact categories. One can reasonably ask whether our conclusions would extend to natural kinds. Methodologically, the emphasis on artifact concepts makes sense: with artifacts, it is fairly straightforward to specify putative critical features (e.g., their function; Bloom, 1996). It is much less clear what would be the critical features for most natural kind categories (Rips, 2001). Following Rips (1989), one might suggest that for living organisms 'mating' might correspond to a sufficient feature for certain categories. However, as we have seen, not all investigations have supported this conclusion (e.g., Pothos & Hahn, 2000). More importantly, some researchers have argued that critical features do not apply at all in the case of natural kinds, rather what drives categorization is a belief into 'essences', that is, hidden, underlying characteristics that make the members of natural kind categories to be what they are (Malt, 1990; Medin & Ortony, 1989; cf. Medin, Wattenmaker, & Hampson, 1987). As one might expect, contrasting similarity and essentialist accounts of categorization is complicated by the fact that, typically (and almost by definition), we do not know what the relevant essences for different categories are. A recent attempt (Pothos et al., 2009) has produced support for both purely similarity-based categorization and essentialist categorization, but these investigators did not employ the sparse-complex methodology developed here. With future work we hope to carry out such an investigation and so examine whether the sparse-complex framework might explain (part of) essentialist categorization effects.

Acknowledgements

This research was partly supported by ESRC grant R000222655 and EC Framework 6 grant contract 516542 (NEST). We would like to thank Nick Chater, James Close, Amotz Perlman, and Ilias Tzimkas for their help at various stages of this project.

References

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. Cognition, 54, 299-352.

Barsalou, L. W. (1985). Ideals, central tendency and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology: Learning, Memory and Cognition, 11, 629-654.

Bloom, P. (1996). Intention, history, and artifact concepts. Cognition, 60, 1-29.

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning. Mind and Language, 8, 72-89.

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

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. Memory & Cognition, 31, 199-214.

Gelman, S. A. (2003). The essential child. Oxford University Press.

Gosselin, F. & Schyns, P. G. (2001). Why do we SLIP to the basic-level? Computational constraints and their implementation. Psychological Review, 108, 735-758.

Hahn, U. & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? Cognition, 65, 197-230.

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: essence, appearance, and behavior in the categorization of natural kinds. Memory & Cognition, 35, 1785-1800.

Keil, F. C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.

Komatsu, L. K. (1992). Recent views of conceptual structure. Psychological Bulletin, 112, 500-526.

Larochelle, S., Cousineau, D., & Archambault, A. (2005). Definitions in categorization and similarity judgments. In H. Cohen & C. Lefebvre (Eds.) Handbook of Categorization in Cognitive Science. Amsterdam: Elsevier, p. 278-303.

Malt, B. C. (1990). Features and Beliefs in the Mental Representations of Categories. Journal of Memory and Language, 29, 289-315.

Malt, B.C., Sloman, S.A., Gennari, S.P., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language, 40, 230262.

Medin, D. L. & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou and D. L. Ortony (Eds.) Similarity and analogical reasoning. Cambridge: Cambridge University Press.

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

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

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception and Performance, 17, 3-27.

Oaksford, M. & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. Psychological Review, 101, 608-631.

Pothos, E. M. (2005). The rules versus similarity distinction. Behavioral & Brain Sciences, 28, 1-49.

Pothos, E. M. & Hahn, U. (2000). So concepts aren't definitions, but do they have necessary *or* sufficient features?. British Journal of Psychology, 91, 439-450.

Pothos, E. M., Hahn, U., & Prat-Sala, M. (2009). Similarity chains in the transformational paradigm. European Journal of Cognitive Psychology, 21, 1100-1120.

Rehder, B. & Hastie, R. (2004). Category coherence and category-based property induction. Cognition, 91, 113-153.

Rips, L. J. (1989). Similarity, typicality and categorization. In S. Vosniadou and A. Ortony (Eds.) Similarity and analogical reasoning. Cambridge, UK: Cambridge University Press.

Rips, L. J. (2001). Necessity and natural categories. Psychological Bulletin, 127, 827-852.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyles-Brian, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8, 382-439.

Sloman, S. A. & Rips, L. J. (1998). Similarity as an explanatory construct. Cognition, 65, 87-101.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature Centrality and Conceptual Coherence. Cognitive Science, 22, 189-228.

Smith, J. D.& Minda, J. P. (1998). Prototypes in the mist: the early epochs of category learning. Journal of Experimental Psychology: Learning, Memory, & Cognition, 24, 1411-1436.

Tversky, A. (1977). Features of Similarity. Psychological Review, 84, 327-352.

Tables

Table 1. Number of participants considering most of the test items to be ‘Chomps’ or not, in Experiment 1.

	Test items ‘Chomps’	Test items not ‘Chomps’
Simple condition	9	33
Complex condition	22	20

Table 2. The stimuli used in Experiment 2.

Object	Subordinate/ Superordinate classifications	Changed feature	Actual question used in the task
Piano*	Piano/ musical instrument	Converted to a mini-bar.	The object above has had all of its internal workings removed and replaced by a minibar.
Football*	Football/ sporting equipment	The football that is part of the World Cup Trophy.	In the above picture, part of a famous sporting trophy is highlighted with an arrow.
Fakirs bed*	Bed/ furniture	Bed with nails.	The picture above is of an object on which Indian fakirs are sometimes seen to lie down.
Doll*	Doll/ plaything	A Chinese figurine made of porcelain.	Look at the object above; it is Chinese a figurine and it is made of porcelain.
Hammer*	Hammer/ tool	A soft rubber hammer.	On his retirement, a builder is given the above object by his colleagues. It is made entirely of very soft rubber.
Loudspeaker	Loudspeaker/ stereo equipment	Converted to a nest for birds.	The back of an old loudspeaker is removed, and it is filled with straw to make a bird's nest.
Word	Word-processor/ software	A virus means that every time a key is pressed beep sounds and nothing else happens.	A computer virus has changed Microsoft Word into a program which just makes a beeping sound whenever any keys on the keyboard are pressed, so that no writing is possible.
Coin	Coins/ money	Economy of a country collapses so that the currency cannot be used for monetary exchanges.	After a series of financial disasters, the economy of a country collapses so that the country's currency is worthless, and cannot be used to buy anything.
Porsche	Car/ vehicle	A steel replica of a Porsche car outside the Porsche headquarters	Outside the headquarters of the Porsche car company, a stainless steel replica of a Porsche stands on a pedestal.
Tunnel	House/ building	A network of tunnels is converted so that people can live in it.	A network of tunnels is dug into the ground, and furnished so that people can live in it.
Box	Cardboard box/ container	A cardboard is flattened.	A cardboard box is cut down the sides so that it becomes a flat piece of cardboard.

Note: A '*' indicates that the description of an item was supplemented with a picture.

Table 3. The percentage endorsement rate for each item, when participants were asked to classify it in a superordinate or corresponding subordinate category.

Item	Subordinate	Superordinate
piano	0.5	0.25
football	0.43	0.31
fakirs bed	0.32	0.35
doll	0.77	0.18
hammer	0.63	0.37
loudspeaker	0.40	0.33
MS word	0.24	0.80
coin	0.80	0.70
Porsche	0.35	0.27
tunnel	0.47	0.37
box	0.30	0.19
Average	0.47	0.37

Figure captions

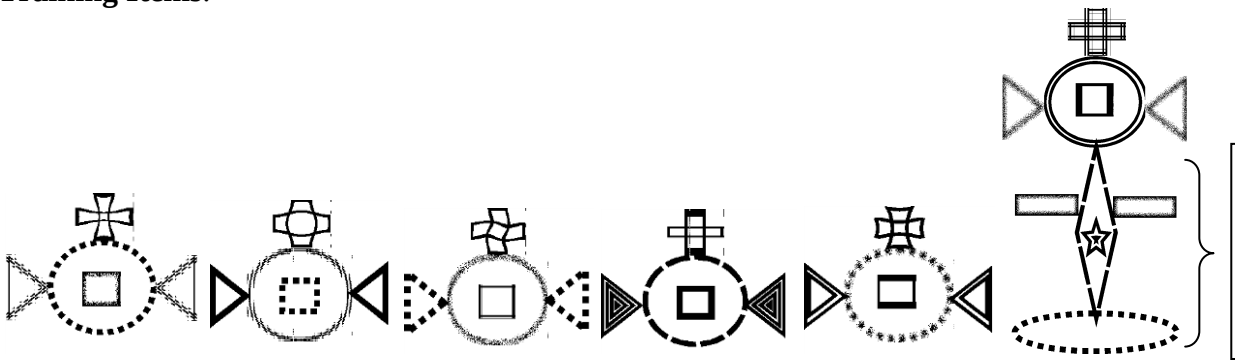
Figure 1. A list of all the stimuli employed in the Simple condition of Experiment 1. Each participant was tested with only one set of Test Items. The stimuli in the Complex condition were the same, except for the fact that (the same) four additional features were added below each of the stimuli in the Simple condition (this is illustrated for one of the Training Items). The three sets of Test Items partly counterbalance which feature was changed.

Figure 2. An illustration of the idea that superordinate classifications involve fewer features than subordinate ones. Each circle represents a binary feature, that is the feature is either present or absent. Black circles correspond to features which are a) present b) common to all the members of a particular category. Therefore, the black circles for each category indicate the representation of the category (in terms of features). The bottom part of the figure shows two subordinate categories. Their members have quite high overlap, there are four features common to all members. The top part shows a corresponding superordinate category. This category contains all the exemplars of the two subordinate categories. However, one can readily see that the only feature which is present and shared by all the exemplars of the superordinate category is the middle one. This illustrates the idea that the superordinate category has a sparser representation compared to the subordinate ones.

Figures

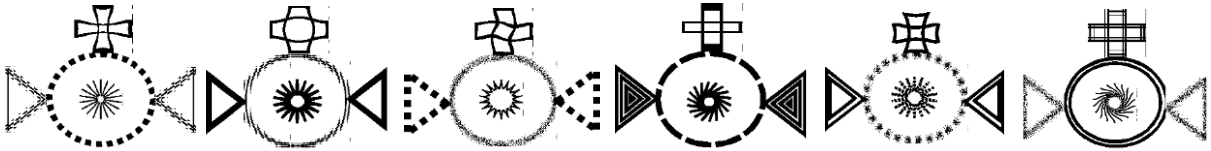
Figure 1 [on the following page]

Training Items:

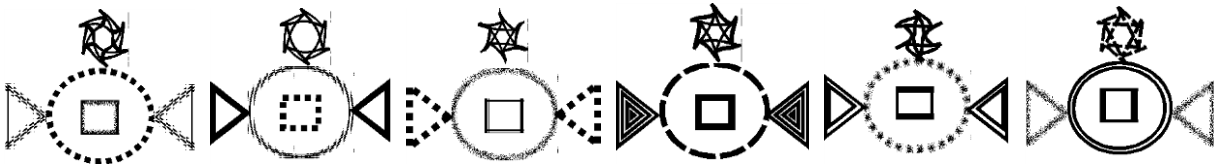


Complex items were created from Simple items with this addition.

Test Items – Set 1:



Test Items – Set 2:



Test Items – Set 3:

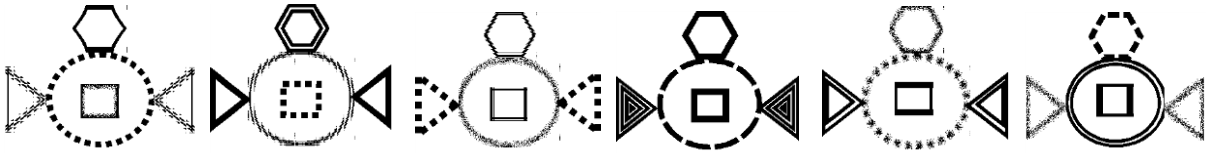
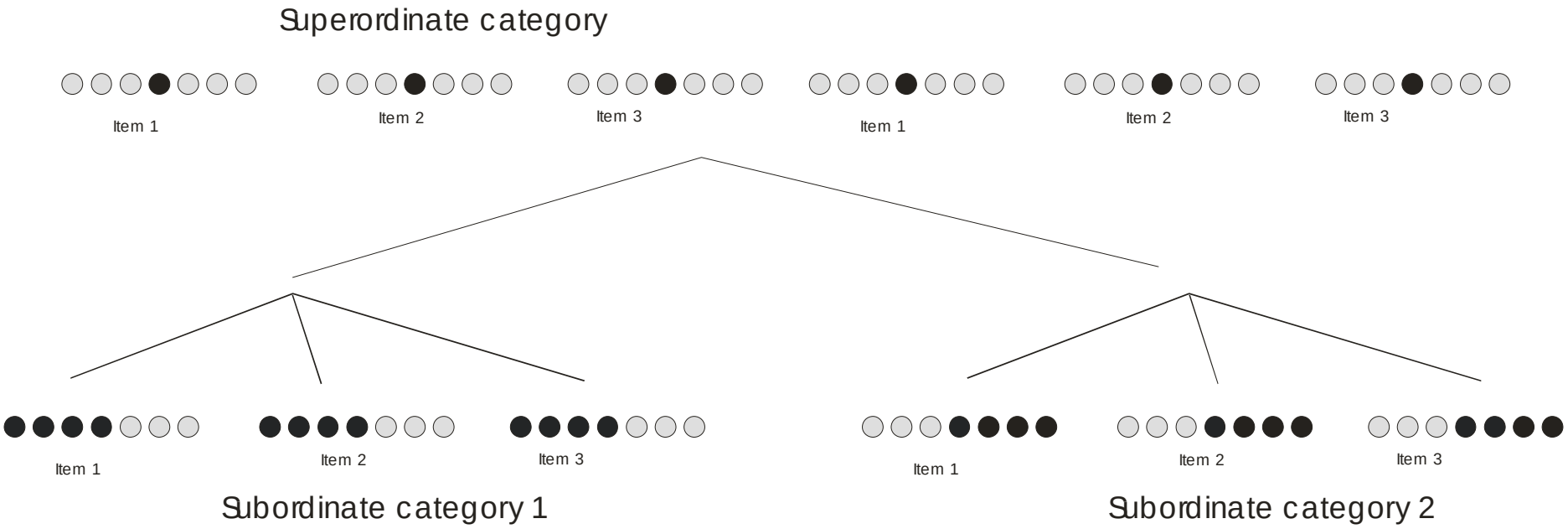


Figure 2.


Superordinate category
Item 1
Item 2
Item 3
Item 1
Item 2
Item 3
Item 1
Item 2
Item 3
Item 1
Item 2
Item 3
Subordinate category 1
Subordinate category 2