Stability in Concepts and Evaluating the Truth of Generic Statements

James A. Hampton
Department of Psychology
City University, London

Contribution to *New Directions in Cognitive Science*, Ed. F.J.Pelletier. Oxford University Press.

Revised Draft: February 2007

Address for correspondence:
Department of Psychology, City University, Northampton Square, London EC1V OHB
hampton@city.ac.uk

Introduction

In this paper I present arguments and evidence in favor of an extended version of prototype theory of concepts (Hampton, 1979; 1995; 1998; Rosch & Mervis, 1975). I first present a brief outline of the theory, and then focus on two particular sources of psychological evidence that support the theory – vagueness in categorization, and the effects of modifiers on the truth of generic sentences.

Prototype Theory

The prototype theory of concepts was introduced by Eleanor Rosch and Carolyn Mervis as an account of a number of phenomena that failed to be explained by classical theories of concepts. The notion of prototype representations had already been proposed in the psychological literature (Dennis, Hampton & Lea, 1973; Posner & Keele, 1968), but it was Rosch and Mervis (1975) who provided the most substantial evidence that the concepts underlying the meaning of many substantive terms may be represented in this way. Related work on prototypes in linguistics can be found in Aarts et al., 2004, Lakoff, 1987, and Taylor, 1995.

In the light of research on prototypes from the last three decades, four main effects can be identified:

A) The membership of conceptual categories is often *vague* in the sense that even though the objects and concepts in question are perfectly familiar it is difficult to decide if the object belongs in the category or not. The case of tomatoes being fruit is a classical example, but there are many others (see McCloskey & Glucksberg, 1978; Hampton, 1979; 1998).

B) The different members of a category differ in their *typicality*, even though they may all be clearly in the category. For example an apple is considered a more typical fruit than a coconut.

C) When people turn their attention to *why* certain objects fall in a particular category, they are frequently unable to provide a definition that satisfactorily explains the difference between items that are category members and others that are not (Hampton, 1979; McNamara & Sternberg, 1983). If there is a definition of the term represented in the mind it appears to be “opaque” – there is a sense that one knows the meaning of the term, but it is not transparent how it is applied.

D) When asked to describe characteristics of the category in question, people find it easy to generate many relevant attributes (they may say for example that fruit is sweet or that it grows on trees), but most of these attributes are not in fact true of all members of the category. In fact they reflect *generic* attributes of the concept.[1]

Put together these four sources of evidence point strongly to the notion that when people represent a concept such as FRUIT, what they have in mind is some prototypical notion of the typical attributes to be found in the category. What they do *not* have in mind is some rule that tells them what is in the category and what is not.

The four phenomena of vagueness, typicality, opaque definitions and genericity are readily explained by the central assumption of prototype theory – namely that people represent concepts in terms of a cluster of attributes that are typically true of the conceptual category. [2]

Extended Prototype Theory

[1] There are many forms of generic sentence that have in common that their truth appears to be independent of the existence of counterexamples (see Krifka et al., 1995). Most of the descriptions generated in describing conceptual contents refer to attributes that are frequently true (“birds fly”) or relevant ways in which the concept can be distinguished from other contrasting categories (“birds lay eggs”).

[2] In linguistic semantics there is evidence for the need to differentiate different domains of concept such as natural kinds, artifacts, life forms etc., (Wierzbicka, 1985). The psychological literature contains no strong evidence for such a fine grain of distinctions, although there are important differences in how people understand the ontology of natural kinds versus artifacts. In particular , the four phenomena described here are easily demonstrated for a wide range of semantic categories from different domains.

Rosch and Mervis (1975) demonstrated the prototype nature of concepts with a simple "feature listing" task in which people had to report the features that they thought might explain category membership. Hampton (1979) adopted a more in-depth interview method in which people were asked to generate attributes of concepts in a number of different ways, but the representation used for each category was again a simple list of features. More recently, evidence has accumulated that the weight that a feature carries in determining the membership of a concept is not just a function of its statistical distribution (as had been originally assumed). Features may be equally strongly associated with a concept class, but differ in the degree to which they are important for membership. For example almost all motor tires are black and round, but it is clearly more likely that a white round object could be a tire than that a black square object could be a tire. In order to explain this important effect, it is now commonly assumed that concept representations also include information about causal dependencies amongst the features (Barsalou & Hale, 1993; Murphy & Medin, 1985). Being round is causally involved in the function of a motor tire in a way that being black is not. Hence the representation of concepts includes not only diagnostic information (cues statistically associated with the category) but also explanatory information (why cues co-occur in a particular cluster).

To some writers, this newer theory of concepts leaves prototypes behind. There has been a tendency to see a "prototype" as a visual image corresponding to the most typical individual member of a category, (see for example the definition of prototype in Osherson & Smith, 1981), and it is clear that such a representation would be far from adequate in explaining our knowledge of the contents of the concept. If seen as a cluster of properties with associated dependency information, then the extended notion of a prototype is the same notion as a *schema* or *frame* representation. Why then still claim that these richer representations are prototypes? Because the four phenomena listed above still need an account, and the most obvious account is that people represent the central tendencies of classes and not their boundaries. The vagueness of categories and the variation in typicality of members are entirely consistent with the schematic representation of the concept including causal dependency information. Similarly the difficulty in pinning down a precise definition and the ease of producing generic information, regardless of whether it is universally true or not, are both consistent with a representation of a rich prototype concept.

Having outlined the theoretical framework within which the research is set, the remainder of this paper will describe two new lines of research that explore the nature of prototype representations. The first relates to the stability of semantic judgments, and the second to how one judges the likely truth of generic sentences. Further details of these and related studies may be found in Hampton (2006).

Stability of semantic judgments

People have a hard time agreeing about what items belong in what class. Furthermore they may frequently change their own mind. A classic study by McCloskey & Glucksberg (1978) presented students with lists of items related to different categories. For example there were lists of possible furniture, possible fruits or possible vehicles, including some items that were clearly category members, some that were clearly not, and others that lay somewhere in between.

One group were asked to judge how typical each item was in its category. A second group simply had to judge if the item belonged in the category or not with a Yes/No binary decision. This second group returned a few weeks later and made the judgment for a second time. Figure 1 illustrates their results. The data have been grouped according to ranges of mean typicality derived from the first group along the horizontal axis, while on the vertical axis is shown the mean probability that the second group categorized the item as a category member. Note how the probability of saying "yes" rises in a very systematic fashion as typicality increases. Notice also how for a wide range of items the probability of categorization is neither 1 nor 0. This region of the borderline of a category is where the phenomenon of vagueness is to be found. Items within this region are neither clearly in the category nor clearly not in it, at least on the basis of the lack of agreement among the students. Furthermore, McCloskey & Glucksberg demonstrated that for these same borderline items people in the second group were likely to change their categorization from one week to the next. The vagueness was not therefore simply a matter of different strokes for different folks – many individuals were unable to maintain a consistent decision about the categorization at the borderline.

In relation to the question of stability, three studies will be presented. The first concerns the question of what underlies the instability. People were asked to choose between different justifications of why they chose to say that an item was neither clearly in a category nor clearly out. The second study considered whether the region within which instability occurs might itself be well-defined. Are people clear about what is "clearly" in or out of a category and what is not? The final study concerned the stability of typicality judgments, and in particular whether they follow a prediction of prototype theory that typical items should show greater stability than atypical items.

Study 1: the reasons for vagueness

A number of different theoretical accounts have been offered to explain the problem of vagueness in categorization (see Hampton, 2007). In this study I focussed on three possibilities:

a) People are vague about categorization because they feel that they don't know enough about the concept or the item being categorized. This view is known as the epistemological account of vagueness (Williamson, 1994) by which there is actually a correct meaning for a term, but people are *ignorant* of it.

b) People are vague about categorization because the terms are inherently *ambiguous*, and the precise answer would require more information about the exact context involved.

c) People are vague about categorization because the concept is itself vague – there just is no hard and fast way of determining the "correct" categorization status of an item, even if one knew all about it and even if the context was fully specified. One way that this can be manifested is if membership in a category is graded or *partial*, so that whether one counts an item as belonging or not would depend on whether a broad or a narrow view of the category was taken.[3]

In order to test these three ideas, students were given lists of items to categorize, including many that were on the borderline. They were given three responses to choose from: "clearly yes", "intermediate" and "clearly not", and worked through the lists categorizing each item with one of these responses. When they reached the end they were unexpectedly asked to revisit each of the responses where they had marked the categorization as "intermediate" and tick off one or more possible explanations from a list provided. The list included two possible types of Ignorance (e,g, "because I don't know enough about the category/item to say"), four reasons that related to Ambiguity (e.g. "because some examples of the item are in the category and some aren't" or "because it depends on the context in which you have to categorize the item"), and one that related to Partiality ("because it depends on whether you take the category in a broad or in a narrow sense"). Table 1 shows the options given to the participants together with the percentage of decisions for which they were chosen, and Table 2 gives the relative percentages broken down to the three main types of justification for each of the 8 categories. Ignorance, Ambiguity and Partiality were each selected about one third of the time as explanations for an intermediate response, but the frequencies varied across categories.

There was a clear differentiation between the categories that may be considered to have some kind of technical definition (the biological categories Vegetable, Fruit, Fish and Insect and the category Science) and the rest. Ignorance was used as an explanation 49% of the time for these 5 categories, and only 8% of the time for the rest (the categories of Sports,

[3] There is also the possibility that the contents of the concept are in some sense "indeterminate" – a person's knowledge of the term is "woolly" and they possess no recognizable set of descriptors that would enable it to be applied in any consistent way – perhaps because no such descriptors exist. Since the experimental materials here employed well-known familiar concepts, I do not consider this explanation for vagueness any further here.

Tools and Furniture). Table 1 confirms that ignorance largely related to particular items, rather than to the category as a whole. This result is consistent with a finding reported by Estes (2004) that people are inclined to assume that membership of natural kind categories is all-or-none, while membership in artifact categories may be partial. In line with Estes' results, Ambiguity and Partiality by contrast were more commonly selected in the latter 3 categories. Ambiguity was selected 51% of the time for the non-natural kinds and 27% of the time for the natural kinds, while Partiality was selected 40% and 24% of the time respectively.

A final analysis looked at the correlation across items of the frequency with which different justifications were selected. The two Ignorance options were uncorrelated with the rest, whereas the different forms of ambiguity and partiality all correlated positively. The degree to which people stated that there was ambiguity in the item, category or context correlated with the degree to which people stated that the answer depended on whether the category was taken in a broad or a narrow sense. There was therefore evidence in the correlations that there are principally two underlying reasons for people being unsure of a categorization – one relating to ignorance and the other relating to the fact that the concepts are vague and context dependent so that there is no determinate answer to the question. The latter explanation clearly fits well with a prototype account of concepts. Because we represent concepts via their central cases, and do not hold consistent information about the rules for determining the boundaries, we are often aware that membership in a category can be a matter of how broadly or narrowly one draws the criterion for category membership. Naturally this study does not provide definitive evidence of the reasons that people may *actually* have for giving an intermediate judgment – we are relying on their ability to introspect about their concepts and their conceptual judgments. However this kind of metalinguistic judgment has proved of considerable value in other fields, and to the extent that the reasons chosen differ systematically between items and domains they provide an additional source of behavioural evidence that a psychological account of vagueness needs to address.

Study 2: second order vagueness

A second question that arises about the vagueness of semantic categories is whether in fact the borderline region in which categorization is uncertain is itself more tightly defined. Accounts of the logic of vagueness have been offered that use three-valued logics according to which a statement can be true, false, or have an undefined truth status (Halldén, 1949; Williamson, 1994). It is obviously important for such logics that the question of whether or not a statement has a truth value should not itself be subject to vagueness. A recent example of this type of three-valued account is Kamp and Partee's (1995) use of "supervaluations" to explain the logic of vague statements (Fine, 1975; Kamp, 1975). They hypothesize that the category scale can be divided into 3 distinct regions – clear members, partial members and clear non-members. They then offer their supervaluationist account of how one can still deduce that "x is an A or not an A" is true, and that "x is an A and not an A" is false, even when the question of whether x is an A or not is itself undetermined (because x is borderline to the category A). Their account thus aims to explain how I may be stumped by the question of whether a person living in France with assets of $100,000 should be categorized as "rich", but would be happy to agree that they are definitely either rich or not rich, and that they are definitely not both rich and not rich. In order for this account to work however the borderline indeterminate region of membership needs to be clearly demarcated. The second study examined whether this was the case with borderline cases of category membership.

The data collection was very simple. Two groups of students were used, and each returned after 2 weeks to repeat exactly the same task. The degree to which they changed their responses on the second occasion was taken as a measure of the vagueness of the decision. One group were given a standard yes/no categorization to make about a list of cases like those used by McCloskey & Glucksberg (1978). The second group were allowed to make one of three responses, corresponding to the Kamp and Partee (1995) three-valued logic. Thus they could say "definitely yes", "maybe", or "definitely no" to each category-item. If it is the case that we clearly recognize when we know something to be true or false for sure, then one would expect a lower rate of inconsistency for the second group when they returned to make their judgments a second time. Intuitively it seems quite plausible that I could know an apple to be a fruit and a carrot not to be a fruit, and I know equally well that a tomato is unclear. So two weeks later I shouldn't be likely to change my mind. On the other hand if I can only say "yes" or "no", I may say yes to the tomato the first time round, and then change it to a no on the second occasion, since I really am unclear about the correct answer. A well delimited region of uncertainty should therefore correspond to a reduced level of inconsistency in the three-response group compared with the two-response group.[4]

The study has now been run twice with students at City University London, once with Bayo Aina, and once with Gurinder Jai. On the replication study we tightened up the instructions and asked people in the second group only to say "yes" or "no" if they were sure in their minds about the answer. The results were quite clear and consistent. There was no observable change in either study in the rate of inconsistency between the two groups. In order to compensate for the fact that the second group had greater opportunity for change (having 3 rather 2 response options), the consistency was first counted for both groups as the probability of a "yes" or a "clear yes" being unchanged on the second occasion. The same calculation was then made for a "no" or a "clear no", and the average of yes and no stability calculated for each group.

In Aina's experiment, with 58 participants and a list of 132 items, the percentage of consistent responses was 83.2 ± 1.8% for the two-response group and 83.7 ± 1.5% for the three-response group[5]. The difference (0.5%) was clearly not significant, and it was estimated that the design had a power of 90% to detect a significant difference between the participant groups of as little as 4%, with alpha set at .05.[6]

In the smaller replication by Jai (with 40 participants and 96 items) the two-option group showed average stability of yes and no responses of 67%, while the three-option group showed average stability for clear-yes and clear-no responses of 69%. Again there was no significant difference. This study had an 80% power to detect a difference of 10% by subjects and 5% by items, with alpha at .05.

There was therefore no evidence that people have any better idea about whether an item is vague than they do about whether an item is in the category in the first place. The result is entirely consistent with the view of concepts that has been proposed here. We represent central cases rather than rules for categorization. Borderline cases are not all just "known exceptions" or problem cases –an item that seems borderline on one occasion may appear quite clear on the next occasion. In fact in the Jai study an item identified as borderline on the first occasion had only a 20% chance of being still considered borderline at retest.

Study 3: Stability of typicality judgments.

The last of the studies of stability in concepts was an experiment that tested a direct prediction of prototype theory concerning typicality judgments. The typicality of an item is usually measured by asking people to rate how "representative" or "typical" it is of a particular category. People commonly agree for example that a chair is a typical kind of furniture whereas a piano is atypical. The account offered of this phenomenon (Rosch, 1975) is that people find it quite easy to compare an item with the prototype representation of the category and so to judge how similar the two concepts are. Indeed it is assumed that a similar process is involved in judging typicality as in judging category membership – both involve assessing the degree of match between the item and the category – although there is evidence that the

[4] A reviewer correctly pointed out that inconsistency across time is not incompatible with Kamp and Partee's supervaluationism – since the determination of the vague region only has to be held constant between the evaluation of the proposition "x is an A" and the proposition "x is not an A". However the question of whether we "know what we don't know" about categorization remains an important and interesting empirical question. The existence of fairly precise borders for the region of vagueness would place a useful constraint on theoretical accounts of instability and vagueness in categorization.

[5] 95% Confidence intervals are quoted.

[6] The analysis by items had a 90% power to detect a difference as small as 2.2%.

weight accorded to deeper features may be greater in the case of categorization (Hampton, 1998; Rips, 1989).

The experiment tested the simple notion that variability in typicality judgments (which Barsalou, 1987 showed to be quite considerable) could be explained by variability in the weight of different attributes in the concept representation. (This same variability may also be a primary source of the instability in categorization reported above). If this were the case, then one would predict that the closer the match of an item to a category, the less variable the rating of typicality should be. For example an item that matches all of the attributes of the category will always have maximum similarity to the concept, regardless of attribute weights. The typicality judgment should therefore always be maximal. But an item that matches, say, two-thirds of the attributes will be subject to instability depending on whether the two-thirds that it has are given high or low weight on a particular occasion. A similar stability should also be found for items that match very few attributes – but in that case they would clearly not be category members, and so questions of typicality would not arise.

The experiment was again quite simple. A single group of students rated typicality of lists of items in a number of categories on two occasions two weeks apart. Items were chosen to fall at equal intervals on a scale of mean typicality obtained in pretesting. The likelihood of the same rating being given again was calculated as a function of the initial rating, and is shown in Figure 2. There was clearly greater stability at both ends of the scale – as reported by Barsalou (1987). However this can easily be explained by an end anchor effect. If, for example an item was considered very atypical initially and given a rating of 9, then on the second occasion it could be seen either as more or as less typical than before. If more typical, then the rating would improve to 6 or 7, but if less typical the rating would remain at 9. The end points should therefore show less instability than mid points on the scale for this reason. Our prediction however related to the difference between typical and atypical items, and this difference is also seen in Figure 2. The mean stability for the high typicality end of the scale (ratings 1-4) was significantly greater than that for the low typicality end of the scale (ratings 6-9), in keeping with the prediction. Being close to the centre of a category appears to provide greater stability to the judgment of similarity to the prototype.

Conclusions from the stability experiments

Three studies have been presented investigating how and why semantic judgments tend to be unstable. In the first it was shown that when people said that a categorization was uncertain they then tended to choose a variety of explanations. In keeping with the notion that people represent the center of categories rather than their borders, it was found that people often considered that uncertainty arose because conceptual categories can be interpreted in a broad or a narrow sense, and can be ambiguous and context dependent. In the second study it was shown that borderline cases of categories are not a separately identifiable class of item – when people have to judge if a case is borderline or not they are just as likely to change their response as when they have to judge if it is in the category or not. The borderlines are just fluid and vague, and three-valued logics cannot be readily applied to categorization. The final study showed that, as predicted by prototype theory, typicality judgments are more stable for high typical items than for low typical items. In the second part of this article, I turn to an account of some experiments on a second phenomenon associated with prototype theory – genericity.

Genericity – why do people say that birds can fly?

Study 4 – the modifier effect

Genericity is the aspect of prototypes that relates to the attributes that people produce when describing members of a conceptual category. When asked to say why something is a bird, or what makes an activity a sport, people typically generate “features” or attributes that are typically true of the category. Not only do different things belong to categories to different extents, but attributes may also be true of categories to different degrees. When we say “birds fly” or “sports provide exercise” we have in mind that these are important parts of the information that anyone who possessed these concepts should know. They are part of the general culturally moderated cluster of knowledge that is tied to these particular concepts.

There may well be more scientific ways in which conceptual categories can be defined, particularly for natural kind terms like bird or gold, but the average person's notion may be little more than this cluster of interrelated attributes. There is some debate about whether people also subscribe to the notion of a scientifically discoverable "essence" that constitutes the real meaning of natural kind terms. Hampton, Estes & Simmons (2007) found evidence that there may in fact be individual differences in how willing people are to make this assumption.

The studies of genericity to be discussed form part of a series of experiments reported by Jönsson and Hampton (2006, 2007), who followed up an interesting result discovered by Connolly, Fodor, Gleitman and Gleitman (2006). The effect that Connolly et al. discovered we called the "Modifier Effect". It turns out that if one group of people are asked to judge how true it is that (say) ravens are black, and another group are asked to judge how true it is that young jungle ravens are black, then the second group will give lower ratings. The modifier phrase "young jungle" has an effect of reducing the judged truth of a property of the head noun "ravens". This effect was replicated by Jönsson and Hampton (2007), who showed (as did Connolly et al.) that the effect is moderated by the typicality of the modifier. That is to say that there is a small effect for a typical modifier ("feathered ravens are black") but a larger effect for an atypical modifier ("jungle ravens are black").

The theoretical significance of this effect relates to what it has to say about the process of combining (or modifying) concepts. According to Connolly et al., the modifier effect shows that the prototype of a modified concept (e.g. jungle raven) bears no direct relation to that of its unmodified parent concept (raven). They argued that to assume that some atypical subset of ravens would have the same stereotypical properties as the general class would be a mistake, and so they claimed that the reduced truth for modified sentences shows that people do not usually make this mistake. According to the theory of conceptual combination that they espouse, concepts such as "jungle" and "raven" combine first through intersection of their referential sets (presumably the class or species of ravens that are found in or adapted to living in jungles). All the possessor of the concept is then allowed to assert with any confidence is that jungle ravens are ravens and that they bear some important relation to jungles. They cannot be assumed to be black with the same confidence with which it is known that ravens are black. This theoretical position is opposed to one commonly adopted in cognitive science which suggests that there are a set of processes by which the prototypes of two concepts may be combined to yield a composite (e.g. Cohen & Murphy, 1984; Hampton, 1987; 1988; Murphy, 1988; Smith, Osherson, Rips & Keane, 1988; Wisniewski, 1997). According to prototype combination models, modified concepts should indeed "inherit" stereotypical properties from their "parent" concepts. Hampton (1987) provided detailed data on precisely how the degree to which attributes are judged true of modified concepts can be derived from their judged truth for the simple concepts from which they are combined.

In our first experiment, (Jönsson & Hampton, 2007, Experiment 1) we showed that the effect seems to be one that generates a systematic reduction in the truth of properties. There was a significant correlation such that the more true a statement was rated when unmodified, then the more true it was also rated when modified. In a second experiment, we showed that there are possibly three different sources of the effect. We gave students pairs of sentences, one modified and one unmodified and first asked them to judge which (if either) was more true. About 60% of the time they claimed that the two sentences were equally true, supporting the idea that the modification had been incorporated into the concept prototype with little additional effect to the other prototypical properties. On the remaining occasions, the modifier effect was seen – most often the modified sentence was judged less likely. We then asked our participants to tell us the reason that they had made this choice, and obtained three main types of justification.[7] The first, and most common, was simply to say that the modified sentences didn't sound sensible, because the property was in any case true of the

[7] As before, we are assuming that participants have relevant intuitions about the basis of their judgments that should be used to constrain theoretical explanations.

whole class. For example, one participant preferred to say “seaweed is green” rather than “Indian seaweed is green” because seaweed (in her view) was all green anyway. The modified sentence violates Grice’s maxim of quantity (“Be as informative as you can”), see Grice (1975), Levinson (2000). The second most common justification depended on the participant thinking of a way in which the modifier might affect the property directly or indirectly. The modifiers had been chosen to be independent of the properties, but participants often came up with unanticipated connections. For example it was considered less likely that bitter nectarines would be juicy than that nectarines would be juicy, on the grounds that the bitterness might signal unripeness, and unripeness would lead to less juice.

The two types of justification found so far do not really provide any reason for supposing that people didn’t consider that (in the absence of any knowledge-based reasoning) modified concepts inherit the properties of their parts. The third type of justification however did support Connolly et al.’s position to a degree. This class of reasons related to the uncertainty that an unknown subclass would have the stereotypical properties of the general class. So for example one participant did not think that Brazilian doves were as likely to be white as doves in general, because “I don’t know any Brazilian doves, so they may be any color”. The uncertainty of properties of atypically modified concepts was clearly insufficient as a general account of the modifier effect however, since it accounted for less than 10% of the judgments that participants made in the task.

A further experiment in Jönsson and Hampton (2007) showed another connection between the properties of a concept and those of an atypical subset of the concept. We manipulated the degree to which properties were considered immutable or mutable for the concept. The notion of mutability of a property of a concept (Sloman, Love, & Ahn, 1998) is that properties differ in terms of how easy it is to imagine the concept, otherwise unchanged, but lacking that property. It is assumed that mutability reflects the degree to which a property is seen as causally linked to other properties. Thus it is easy to imagine a tire that is not black or made of rubber, but harder to imagine a tire that is not round or with no means of attaching it to a wheel or axle. In our experiment we showed that mutability affects the degree to which a property is considered true for modified versions of a concept. Thus a property that was more mutable for the unmodified concept (“ravens are black” as opposed to “ravens have feathers”) was considered less likely to be true of the modified concept. Hence “young jungle ravens have feathers” was considered more likely to be true than “young jungle ravens are black”, because of the lower mutability of the former property for ravens.

Study 5: The Inverse Conjunction Fallacy

If prototype theory is right, then people should find it difficult to make consistent decisions about whether one class is completely included within another. For example, if one decides that chairs are a kind of furniture on the basis of the match between one’s representation of the concept of chair and one’s representation of furniture, what is to make sure that there aren’t some atypical chairs that are not actually furniture? Focussing on the descriptive content of the central cases of a category makes it hard to determine what the extent of the denotation of the concept should be, and harder still to determine whether a strict relation of class inclusion holds between two concepts. Hampton (1982) demonstrated this effect with artifact categories. In that study I showed that people were quite happy to say that a chair is a kind of furniture, even though they judged car seats or ski lifts to be chairs that are not furniture.[8] In Hampton (1988, Experiment 1) I similarly showed that the class of things called “school furniture” or “office furniture” were not fully contained in the class of things called “furniture”.

[8] It may be objected that although people may categorize a car-seat as a chair, they would never call it such in everyday discourse. Different accounts may therefore be needed to explain judgments of concept subordination and naming behaviour. However both are equally relevant to the question of how our concepts are represented. Fortunately Hampton’s (1982) intransitivity result did not depend on this single case, but was also found with other kinds of furniture (clocks, shelves, mirrors, lamps, etc), and other general categories (sports, pets, plants, tools etc.) where the objects (e.g. Big Ben) could also be named by the concept term (e.g. clock).

As applied to the modifier effect discussed above, these earlier results suggested that we might find that people continue to consider a modified sentence less likely to be true, even when it is universally quantified. Given the generic semantics of a sentence such as “ravens are black” it is quite permissible for some unusual kinds of raven to be white without thereby rendering the statement false. “Ravens are black” says something about what people generally believe is true of typical ravens (it expresses a prototypical property of ravens), but it does not say that *all* ravens are black. On the other hand the statement “All ravens are black” should logically entail that there are *no* kinds of raven that aren’t black. So if one judged this sentence true, one should also logically judge all sentences of the form “All *M* ravens are black” where *M* is some modifier that selects some non-empty subset of ravens.

Jönsson and Hampton (2006) gave students a choice of two sentences of the following form and asked which was more true (or were they equally true):

All sofas have backrests
All uncomfortable handmade sofas have backrests

There was a strong tendency for people to judge that the unmodified sentence was more likely to be true than the modified sentence, regardless of the illogicality of this position. (After all, “all sofas” should include all sofas, even those that are uncomfortable and handmade.) This effect, which we termed the Inverse Conjunction Fallacy, is similar in many respects to Tversky and Kahneman’s well-known Conjunction Fallacy. The conjunction fallacy (Tversky & Kahneman, 1983) involved the belief that a person was more likely to be found in a subset of a class than in the class itself. For example, a woman (Linda) with liberal politics was judged more likely to be a feminist bank teller than to be a bank teller. The inverse conjunction fallacy turns the effect upside down by considering whether a *property* is true of a whole class or a whole subclass, and in this case the fallacy is to believe that it is more likely for the property to be true of the whole class than of the subclass.

The explanation for both kinds of conjunction fallacy may be the same. Tversky and Kahneman referred to something they called “intensional reasoning” – that is judging the likelihood of an event or fact on the basis of similarity or representativeness[9]. Because Linda is more typical of a feminist than of a bank teller, it somehow seems more likely that she should be in the conjunction of the two sets than in just one of them alone. The inverse conjunction fallacy also seems to involve intensional reasoning. Whatever led to the modifier effect with generic sentences also seems to apply to universally quantified sentences.

Jönsson and Hampton (2006) went on to investigate the fallacy further. First they showed that it does not depend on use of the word “all”. Very similar fallacious responding was found with the phrases “All X are Y”, “All X are always Y”. “100% of X are Y”, and “Every single X is Y”. In other words, no matter how the meaning of the universal quantifier was expressed it tended to be ignored. For example people judged it more likely that “every single sofa has a backrest” than that “every single uncomfortable handmade sofa has a backrest”.

A second control was to check whether people believed that the subclass was in fact in the class. A group of students judged sentences such as “All jungle ravens are ravens” or “All uncomfortable handmade sofas are sofas”. The large majority of the sentences showed near universal agreement to these statements. A third control looked at whether people thought that some subsets might not exist. If there are no such things as jungle ravens, then one would be justified to say it is unlikely to be true that all jungle ravens are black (in the sense that it would not be true because it would be meaningless). A few of the modified concepts turned out to generate some doubt as to their existence, but there was no evidence that these concepts were any more likely to generate the fallacy than others.

[9] Intensional reasoning more generally will depend on what one assumes to be the “intension” of a concept. Tversky and Kahneman’s account clearly presupposes that people make judgments of likelihood on the basis of how similar an object or situation is to the prototypical case. For them an intension thus includes some kind of prototype.

In our final experiment we aimed to discover conditions under which the fallacy would be reduced. Placing the sentences side by side and asking which (if either) was more true did not change the incidence of the fallacy compared with having different groups of students judge each separately. However when people were given the two sentences as a pair one after the other and simply had to say whether each one in turn was true or false, the incidence of the fallacy was significantly diminished. Committing the fallacy is not therefore inevitable, but it appears to be a very powerful cognitive illusion, resisting most of our efforts to alleviate it.

Study 6: Comparing the modifier effect and the inverse conjunction fallacy

The final study to be reported is an unpublished experiment conducted with Ou Lan, a student at City University, London (also reported in Jönsson & Hampton, 2007). In this experiment we considered what effect the introduction of universal quantifiers would have on the modifier effect. Does the word “all” simply reduce general confidence in the statements, in a similar way to the “atmosphere effect” described for syllogisms by Woodworth and Sells (1935)? Or would the difference between modified and unmodified sentences be reduced when the logical constraint of “all” is introduced. The experiment involved two factors. Sentences could be modified or unmodified as before, and in addition they could be generic or universally quantified. Figure 3 shows the results of the experiment in which students judged the likely truth of sentences of each of the four types. The statistical analysis confirmed significant main effects and a significant interaction. Modified sentences were judged less likely to be true, both when generic and when universally quantified. Introduction of the quantifier reduced the judged likelihood of both modified and unmodified sentences. The interaction reflected the fact that the modifier had a larger effect on the generic sentences than on the universally quantified sentences. Universal quantifiers made people more cautious about judging a sentence likely to be true, and also diluted the effect of the modifier. People still however thought it more likely that “All X are Y” than that “All MX are Y”.

Conclusions

In this article I have endeavoured to show how the way in which people use concepts follows systematic patterns that do not necessarily conform to the strict requirements of logic. Human thinking can be classed as logical, illogical and non-logical (Evans, 1982). The first two kinds of thinking are correct or incorrect attempts to derive inferences based on the logic of a problem, whereas the last is the use of a process which is neither logical nor illogical – it is a form of rationality that involves approximate, heuristic reasoning. The kinds of results obtained in the studies described here fall best into the latter class of thinking. People need words in order to communicate, and they frequently are not completely clear about what they are trying to say. Communication only needs to be as precise as the context requires. Most situations involve typical cases in which the names of things are clear, and the categories they belong to are uncontroversial. It is usually more important to know what is generally true of a type of thing, than to know whether something is universally true – more useful to know that toadstools are generally poisonous than that toadstools are not all poisonous. We learn the generality of things first, and only with sufficient expertise would we expect to know of detailed exceptions.

The effects of this way of building and representing knowledge are that semantic judgments can often be unstable or vague – we don’t know (or necessarily care) whether certain things belong in certain categories, because what we represent are the clear cases. Unclear cases have to be decided by comparison with clear cases in the context of the problem to be solved. Similarly, a judgment of whether or not a property is true of a class may be made on the basis of how similar the class is to other classes for which the property is known to be generally true. We find it hard to differentiate our intuitions about what is generally true and what is always true. That is a sign of thinking intensionally rather than extensionally.

References

Aarts, B., Denison, D., Keizer, E., & Popova, G. (2004). *Fuzzy Grammar: A Reader*. Oxford: Oxford University Press.

Barsalou, L. W. & Hale, C. R. (1993). Components of conceptual representation: from feature lists to recursive frames. In I.van Mechelen, J.A.Hampton, R.S.Michalski, & P.Theuns (Eds.), *Categories and Concepts: Theoretical Views and Inductive Data Analysis* (pp. 97-144). London: Academic Press.

Barsalou, L. W. (1987). The instability of graded structure: implications for the nature of concepts. In U.Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101-140). Cambridge: Cambridge University Press.

Cohen, B. & Murphy, G. L. (1984). Models of concepts. *Cognitive Science, 8,* 27-58.

Connolly, A., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2006). Why stereotypes don't even make good defaults. *Cognition*, in press.

Dennis, I., Hampton, J. A., & Lea, S. E. G. (1973). New Problem in Concept Formation. *Nature, 243,* 101-102.

Estes, Z. (2004). Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural. *Psychonomic Bulletin and Review, 11,* 1041-1047.

Fine, K. (1975). Vagueness, truth and logic. *Synthese*, *30*, 265-300.

Grice, H. P. (1975). Logic and conversation. In P.Cole & J. L. Morgan (Eds.), *Syntax and Semantics* (3rd edition, pp. 41-58). New York: Academic Press.

Halldén, S. (1949). *The Logic of Nonsense*. Upsala: Universitets Arsskrift.

Hampton, J. A. (1979). Polymorphous Concepts in Semantic Memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441-461.

Hampton, J. A. (1982). A Demonstration of Intransitivity in Natural Categories. *Cognition, 12,* 151-164.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition, 15,* 55-71.

Hampton, J. A. (1988). Overextension of conjunctive concepts: Evidence for a Unitary Model of Concept Typicality and Class Inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 12-32.

Hampton, J. A. (1995). Testing Prototype Theory of Concepts. *Journal of Memory and Language, 34,* 686-708.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65,* 137-165.

Hampton, J. A. (2006). Concepts as Prototypes. *The Psychology of Learning and Motivation: Advances in Research and Theory, 46,* 79-113.

Hampton, J.A. (2007). Typicality, Graded Membership and Vagueness. *Cognitive Science, in press.*

Hampton, J. A., Estes, Z., & Simmons, S. (2007) Metamorphosis: Essence, Appearance and Behavior in the Categorization of Natural Kinds. *Memory & Cognition, in press.*

Jönsson, M. L., & Hampton, J. A. (2006). The Inverse Conjunction Fallacy. *Journal of Memory and Language, 55,* 317-334.

Jönsson, M. L., & Hampton, J. A. (2007). The modifier effect in within-category induction: On prototypes and default inheritance. MS under review, City University London, September.

Kamp, H. & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57,* 129-191.

Kamp, H. (1975). Two theories about adjectives. In E.L.Keenan (Ed.), *Formal Semantics of Natural Language*, (pp. 123-155). Cambridge: Cambridge University Press.

Krifka, M., Pelletier, F. J., Carlson, G. N., ter Meulen, A., Chierchia, G., & Link, G. (1995). Genericity: An Introduction. In G.N.Carlson & F. J. Pelletier (Eds.), *The Generic Book* (pp. 1-124). Chicago: University of Chicago Press.

Lakoff, G. (1987). *Women, Fire and Dangerous Things*. Chicago: University of Chicago Press.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* Cambridge: MIT Press.

McCloskey, M. & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition, 6,* 462-472.

McNamara, T. P. & Sternberg, R. J. (1983). Mental models of word meaning. *Journal of Verbal Learning and Verbal Behavior, 22,* 449-474.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Murphy, G. L. (1988). Comprehending Complex Concepts. *Cognitive Science, 12,* 529-562.

Osherson, D. N. & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 11,* 35-58.

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

Rips, L. J. (1989). Similarity, typicality and categorization. In S.Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 21-59). Cambridge: Cambridge University Press.

Rosch, E. R. & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605.

Rosch, E. R. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192-232.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189-228.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining Prototypes: A Selective Modification Model. *Cognitive Science, 12,* 485-527.

Taylor, J.R. (1995). *Linguistic categorization: Prototypes in linguistic theory* (2nd Edition). Oxford: Clarendon.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293-315.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Wierzbicka, A. (1984). Apples are not a kind of fruit: The semantics of human categorization. *American Ethnologist, 11*, 313-328.

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin and Review, 4,* 167-183.

Woodworth, R., & Sells, S. (1935). An atmosphere effect in syllogistic reasoning. *Journal of Experimental Psychology, 18,* 451-460.

Table 1: reasons offered to participants in Study 1 for choosing an “intermediate” categorization response. A and B were classed as Ignorance, C through F were classed as Ambiguity, and G was classed as Partiality. Percentage is percentage usage across items and participants.

	REASON	Percentage
A.	because I don’t know enough about the category to say	7%
B.	because I don’t know enough about the item to say	25%
C.	because some examples of the item are in the category, and some aren’t	3%
D.	because it depends on how you define the category (otherwise than G)	15%
E.	because it depends on how you define the item	11%
F.	because it depends on the context in which you have to categorise the item	7%
G.	because it depends on whether you take the category in a broad or in a narrow sense	31%
H.	some other reason (please specify at the bottom of the page)	1%

Table 2 Percentage of reasons given, broken down by three main types of reason and individual category.

	Ambiguity	Partiality	Ignorance
Vegetable	15%	23%	62%
Fruit	27%	13%	60%
Fish	30%	33%	37%
Insect	26%	20%	53%
Science	37%	29%	35%
Sport	49%	43%	8%
Tool	55%	35%	9%
Furniture	50%	41%	8%
Mean	29%	31%	32%

Figure 1

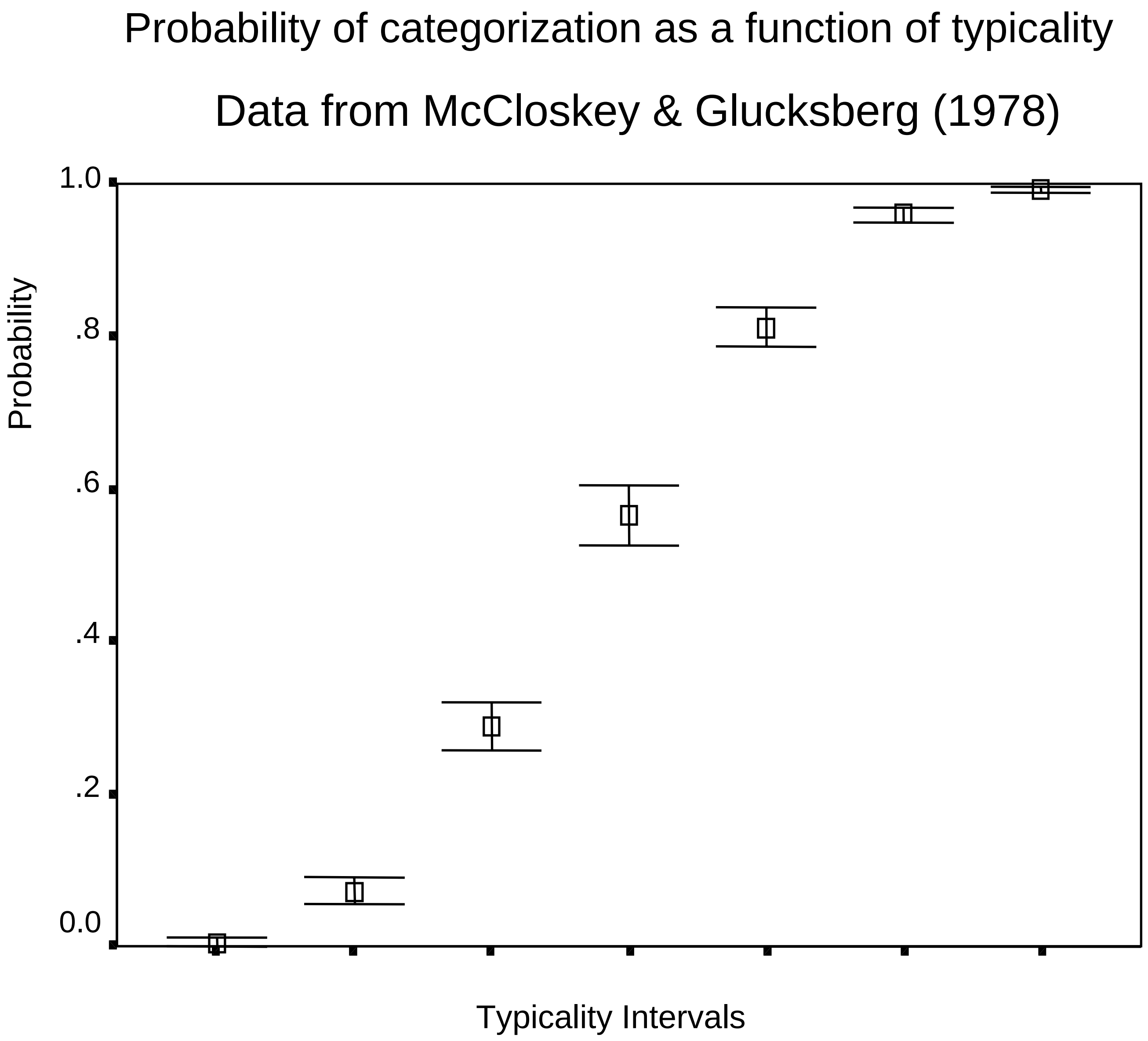

Probability of categorization as a function of typicality
Data from McCloskey & Glucksberg (1978)
1.0
.8
.6
.4
.2
0.0
Probability
Typicality Intervals

Figure 2

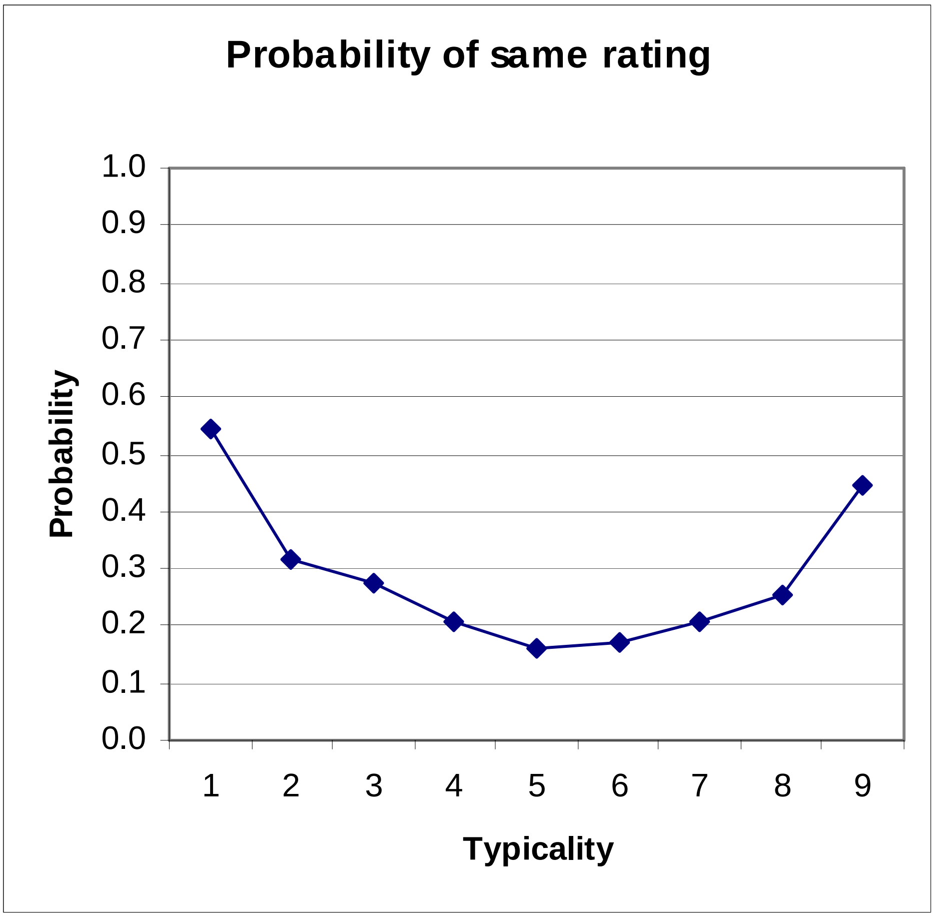
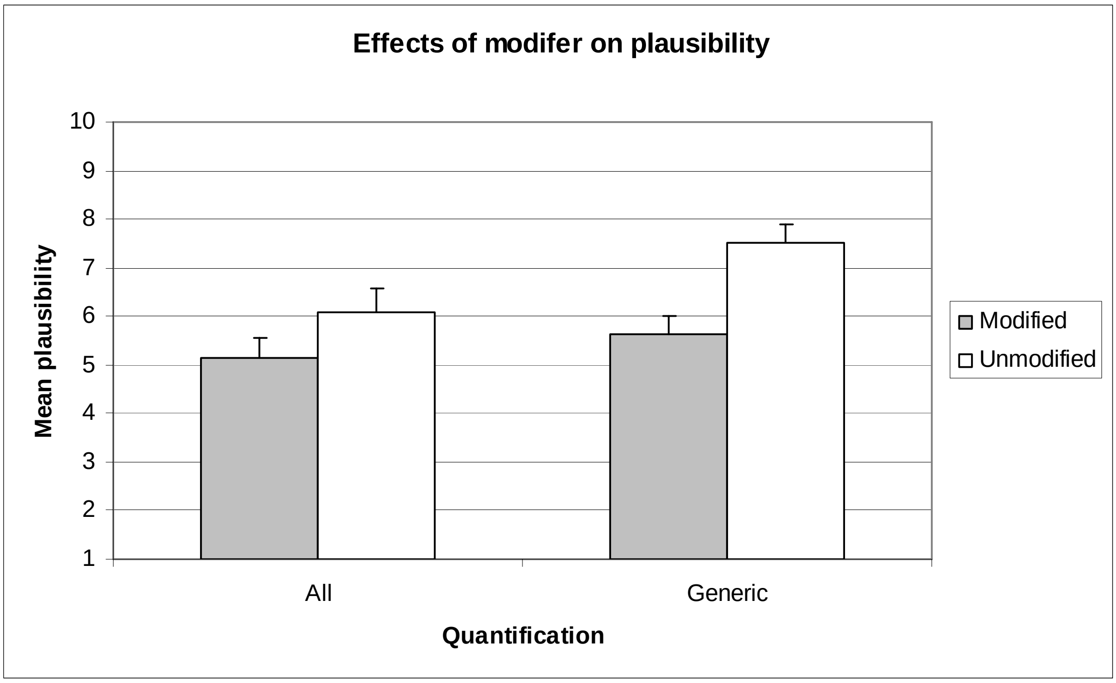

Probability of same rating
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
Typicality

Figure 3


Effects of modifer on plausibility
Mean plausibility
10
9
8
7
6
5
4
3
2
1
All
Generic
Quantification
▨ Modified
□ Unmodified