Some New Evidence for Concept Stability

Sam Scott (sscott@carleton.ca)
Department of Cognitive Science
Carleton University, Ottawa, ON K1S 5B6

Abstract

It is generally assumed that concepts are stored as relatively invariant chunks in long term memory. This is the default assumption in most linguistic, psychological, and philosophical work on concepts. But contrary to this assumption is the recent suggestion that both the structure and content of concepts is highly flexible and unstable. I critically review the evidence behind this assertion, then present some new experimental evidence that supports the traditional view.

Introduction

Concepts are the building blocks of human thought. They are the invariant sub-propositional mental representations of the physical and social world which we use in perception, categorization, learning, and inference. The enormous flexibility, productivity, and context-sensitivity of human thought is a result of computations performed on these otherwise invariant concepts. That's the traditional view, anyhow, and it forms the basis of most published psychological, philosophical, and linguistic work on concepts. Barsalou (1987; 1999) and Prinz (2002) are the architects of a new view of concepts – one which claims to be empiricist rather than rationalist, and one in which the flexibility of thought is explained by the flexibility and instability of concepts. In this paper, I re-examine the evidence for concept instability, and find it lacking. Then I present the results of a new experiment which demonstrates that concepts are rather more stable and inflexible than Barsalou and Prinz claim.

Arguments for Instability

The current era of empirical research on concepts was ushered in by some striking and original results from Rosch and her colleagues (Rosch, 1975; Rosch & Mervis, 1975). In a series of now classic experiments, Rosch demonstrated that concepts are graded structures bound together by overall similarity or "family resemblances". For example, DOG is considered by most experimental participants to be a better exemplar than PLATYPUS of the concept ANIMAL. This is so because the concept DOG shares more features in common with other ANIMAL exemplars than PLATYPUS does, or to put it another way, because DOG is more similar to other ANIMAL exemplars than PLATYPUS is.[1] In the experiments that were most influential in establishing this conclusion, participants were given lists of category exemplars and asked to rate each one for "goodness of example", or "typicality" using a 7-point scale. The participants tended to show a high level of agreement in their ratings, and the mean typicality scores correlated well with independent measures of within-category similarity. This pattern of results has been replicated many times using different methodologies (e.g. Barsalou, 1985; Hampton, 1979; Malt & Smith, 1984; Rosch & Mervis, 1975).

There are a number of competing interpretations of these results, but a near consensus has emerged that they are to be explained by appeal to the structure of the concepts stored in long term memory. Debates continue around the exact nature of that structure (e.g. whether concepts are best described as prototypes or mini theories), but the idea that concepts might be highly flexible and unstable structures has not often been entertained. That is what makes Barsalou's (1987) original challenge and Prinz's (2002) recent reformulation so important.

According to Barsalou (1987), typicality judgments are unreliable both within and between participants, and unstable across contexts. He concludes that stable concepts in long term memory are an "analytic fiction". According to Barsalou, concepts are constructed as needed out of a partly context-dependent subset of our full knowledge. They are not invariant structures retrieved intact from long term memory. The following sections are an attempt to outline and respond to Barsalou's main arguments for this radical conclusion – arguments from variable participant behavior, context effects, and multiple determinants of typicality.

Variable Participant Behavior

Barsalou's first claim is that the typicality ratings on which Rosch based much of her work are unstable both within and between participants. Barsalou (1987) reports that participants seem to agree with each other less than Rosch originally reported, and even disagree with themselves on different occasions. Reviewing his own unpublished work, Barsalou reports that mean between-participant correlations for 20 different groups ranged from 0.30 to 0.60. Rosch's (1975) original finding of high inter-participant agreement reflected a different, possibly misleading statistic in which the stability of the mean is measured by split-group correlation. The problem with this type of statistic is that increasing group size tends to increase the stability of the mean, potentially producing an arbitrarily high appearance of inter-participant agreement. In contrast, Barsalou's statistics resulted from calculating between-participant

[1] Words in small caps refer to concepts. An exemplar of a concept is a sub-concept (e.g. DOG is an exemplar of ANIMAL).

correlations for every possible pair of participants and then computing the mean, to predict the expected level of correlation between any two participants' typicality ratings. Even more potentially damaging is Barsalou's report (again summarizing his own unpublished work) that typicality judgments can differ on different occasions. He had participants rate typicality on the same set of items twice, with varying delays of 1 hour to 1 month. With the exception of the 1 hour delay, the participants correlated with themselves to a mean level of 0.80. Barsalou concluded that typicality judgments are too unstable to be useful in determining the underlying structure of a concept.

Related studies by Bellezza (1984a; 1984b; 1984c) also lend general support to Barsalou's results. Bellezza showed that participants tend to be highly variable in their performance on two types of production tasks related to categorization: exemplar listing for categories (Bellezza 1984a), and property listing for nouns and proper names (Bellezza, 1984b; 1984c). In this series of experiments, participants performed an identical task on two separate occasions, one week apart. The degree of overlap between the two sessions on the exemplar-listing task, as measured by mean common-element correlation, was 0.69 within participants and 0.44 between participants. For the property-listing tasks, the numbers were approximately 0.50 within participants and 0.20 between-participants.

What should we conclude from all this? First of all, the fact that participants disagree with each others' typicality ratings is not particularly surprising, and does not necessarily invalidate the mean ratings over a large group of participants, particularly when these mean ratings prove to be predictive of other psychological effects. Experimental psychology is a statistical science designed to measure tendencies across large groups. Between-participant variance of the kind Barsalou reports is neither unusual nor troubling. Nor should it be particularly troublesome that participants show a high variability in production tasks, which are probably affected by all sorts of subtle priming effects as well as the participants' moods and preoccupations on the day tested. It is only the finding of within-participant disagreement on the typicality ratings in particular that might really pose a problem. If participants can't even agree with themselves across time, then maybe typicality ratings do not result from the structure of relatively invariant concepts after all.

But there are two reasons to doubt the significance of the problem. First of all, within-participant reliability for typicality, as measured by Barsalou, seems significantly higher than within-participant reliability on Bellezza's production tasks. This shows that participants are more variable in free recall than they are in rating given stimuli, which is exactly the sort of trend to be expected if the act of rating exemplars causes the participants to access some kind of invariant underlying structure. Secondly, Barsalou (1987) reports that high- and low-typicality items showed little variance within participants. Rather, the majority of the within-participant disagreement was on the intermediate-typicality examples. This may simply reflect the way the participants were using the rating scale. If they tend to rate high on the scale for typical exemplars and low for atypical exemplars, then that leaves them with a lot of room to maneuver in the middle of the 7- or 9-point scales that are usually used. If so, participants can be expected to produce different values in this range on different occasions. Later work by Barsalou (1989) supports this interpretation: when participants rank exemplars in order of typicality, they show less variance, both within- and between-participants, than when they use a rating scale. So the reliability of typicality ratings may not be a major cause for alarm after all.

Context Effects

Barsalou's (1987) second claim is that sentential and situational context will affect typicality judgments. As evidence, he cites Roth and Shoben (1983), who found that when a concept word appears in a sentence, the sentential context affects participants' typicality ratings. For example, participants were shown the sentence, "During the midmorning break the two secretaries gossiped as they drank the beverage", and were then asked to use a 9-point scale to rate how well given exemplars fit their "idea or image of what the category term [beverage] refers to in the sentence" (Roth & Shoben, 1983, pp. 358-359). Normally atypical exemplar names for BEVERAGE, such as "coffee" or "tea" were rated much more typical in context than when no context was provided. In other experiments, Roth and Shoben also found differences in reaction times that were consistent with a shift in typicality across contexts. Similarly, Barsalou (1987), again reviewing his own unpublished work, reports that participants changed their typicality judgments when instructed to take another person's point of view. For example, American students judged robins and sparrows to be highly typical birds from an American point of view, while swans and peacocks were highly typical from a Chinese point of view.

The problem with these types of studies is that it is not at all clear that the concept itself is the same across contexts. It seems quite likely that participants are inferring what Barsalou would call "ad hoc" concepts such as BEVERAGE THAT SECRETARIES DRINK IN THE MORNING or BIRD FROM A CHINESE POINT OF VIEW and rating typicality against those concepts rather than the target concepts BEVERAGE and BIRD. If so, then these studies do not in any way prove that conceptual content is dependent on context. This is not to say that context doesn't play any role in the behavior of the participants, but no evidence to date conclusively shows that the concepts themselves can actually change in context.

Multiple Determinants of Typicality

The evidence in support of Barsalou's (1987) third claim comes from previous work in which he found typicality effects in ad hoc and goal-directed concepts, such as THINGS

TO DO ON THE WEEKEND (Barsalou, 1983; 1985). Rosch had shown that typicality effects in object concepts correlated consistently with family resemblance, rather than any simple measure of frequency or familiarity. But this does not seem to be a possibility for ad hoc and goal-directed concepts, whose exemplars tend to bear little resemblance to one another. For example, what do money, children, jewelry, photo albums and pets have in common? On the face of it, they seem to have very little in common at all, but they are all intuitively good exemplars of the concept THINGS TO SAVE FROM A BURNING HOUSE. On the other hand chairs, televisions and pots do not seem like such good exemplars, though they could all conceivably be rescued as easily as the good exemplars. What could the sources of typicality effects be in these domains?

To answer this question, Barsalou (1985) looked at four possible predictors of graded typicality judgments for a given concept exemplar: 1) *Central tendency*, a measure of within-category similarity, like Rosch's "family resemblance"; 2) *Closeness to an ideal*, such as "zero calories" for the concept THINGS TO EAT ON A DIET; 3) *Frequency of instantiation*, measured by participants' own estimates of how often an exemplar occurs as a member of a given category; and 4) *Familiarity*, measured by participants' own estimates of how familiar they were with the exemplar. In accord with many earlier studies, Barsalou found that central tendency was most predictive of typicality for object concepts[2], but both frequency of instantiation and closeness to an ideal were much more predictive for the goal-directed concepts. Of course, this result is easily explainable given that there is no reason to believe that ad hoc concepts are in any way psychologically basic. Whereas the object concepts he studied are in common use and tend to be lexicalized, ad hoc concepts are uncommon and generally have to be expressed using long phrases. It is quite possible that the ad hoc concepts function just as Barsalou suggests. That is, they have little invariant structure, but rather are constructed on the fly when participants' attention is called to their existence. But the predictive value of family resemblance and central tendency seems to argue against the same conclusion for lexicalized object concepts.

However, Barsalou (1987) went on to try and prove that the determinants of graded structure can change. Having established that ideals can determine graded structure for certain concepts, he constructed an experiment involving two groups of imaginary individuals defined so that they could be sorted equally well by overall family resemblance or using a single attribute. Each individual was represented by a list of 5 attributes, encoding how often they engaged in activities such as dancing, renovating houses, writing poetry, watching movies, and so on. The defining attributes for the two categories were how often the individual read the newspaper and how often they jogged. In the "ideals" condition, participants were told the categories were CURRENT EVENTS TEACHER and PHYSICAL EDUCATION TEACHER. In the "central tendency" condition, they were told the categories were Z PROGRAMMER and Q PROGRAMMER. After studying the descriptions of individuals in the two categories, they were asked to rate them for typicality. Not surprisingly, in the ideals condition, participants' ratings correlated with the values of the relevant defining dimension (that is, how often they jogged or read the newspaper). In the central tendency condition, their ratings correlated with overall family resemblance. Barsalou concluded that this was further evidence for the flexibility of concept structure.

The main problem with this study was that the participants in the ideals condition were not learning arbitrary concepts from scratch, whereas they were in the central tendency condition. Suppose, as is quite likely, that the participants already had pre-existing concepts for PHYSICAL EDUCATION TEACHER and CURRENT EVENTS TEACHER that were organized around family resemblance principles. The participants were then placed before descriptions of individuals that show values for a mere five dimensions. It is likely that very few of these dimensions would seem relevant to categorization decisions based on these pre-existing concepts. (Does a physical education teacher watch more or fewer movies than a current events teacher?) Under these conditions, it's plausible that only the target features of the individuals seemed salient given the concepts' pre-existing central tendencies. If so, the fact that the participants appeared to sort using that feature says nothing about the general organization of their concepts. By contrast, participants in the central tendency condition were learning two new concepts, and thus approached the problem as a new learning situation, basing their typicality judgments on all available attributes at once.

Default Concepts

In the end, none of Barsalou's three conclusions are warranted, hence neither is his suggestion that invariant concepts stored in long-term memory are an "analytic fiction". Nevertheless, Barsalou and (more recently) Prinz have continued to argue that much of the structure of concepts in working memory is assembled on the fly rather than included in an invariant structure retrieved from long term memory. Both Barsalou's (1999) "simulator" and Prinz's (2002) "proxytype" concepts are skeletal structures consisting of just a small amount of context-invariant information. They are default concepts that are retrieved and integrated with other context-dependent information to form a newly modified concept in working memory. On these accounts, concepts are highly flexible structures, and no two retrievals of the same default concept from long term memory are likely to result in exactly the same concept in working memory. Though weakened a little from Barsalou's original position, these theories still predict much more flexibility and variability in concepts than most other

[2] Barsalou called them "common taxonomic categories", but of the 9 categories, 8 were object categories.

theories. In the next section, I present new empirical evidence that I hope better addresses the debate than the work reviewed above.

Some New Evidence

Almost all typicality rating tasks, from Rosch onwards, have used the same methodology: participants are tested in groups, filling out booklets in which each page shows a category name at the top, followed by a list of items to rate for typicality in that category. Thus the context of each individual typicality rating is determined by the list of items in which it appears. Participants do not just rate the items against the category, but also against the other items that the experimenters have chosen to include in the list. If the participant is not sure what the experimenters meant by, for instance, "animal", she can scan the list of stimuli to figure out an intended meaning, then rate each item based on that.

Although Barsalou's arguments failed to establish his claim of massive concept flexibility, the usual method of collecting typicality ratings for family resemblance studies does not rule it out either. It is possible that participants form new concepts for each category name by scanning the list of stimuli and forming a prototype based on what the exemplars and instances in that list have in common. Then they rate for typicality based on that contextually-determined prototype. But if this is what they are doing, then the correlation with family resemblance, a direct measure of how much the exemplars and instances have in common, becomes a self-fulfilling prophecy.

The goal of the experiment described in this section was to check whether manipulating the context of the typicality rating task would have a significant effect on the behavior of the participants. Typicality ratings were collected using a task designed to minimize the effect of context. Instead of seeing a list of items on a page, participants saw the items one at a time, displayed on a computer screen in random order. In this condition, items were presented randomly from 10 different concepts to ensure that, on average, participants would have to process 9 items from different concepts between each pair from the same category. The immediate context of each individual typicality judgment under these conditions is quite different between participants. If, contrary to Barsalou and Prinz, typicality ratings do cause participants to access invariable default concepts, then between-group agreement should be high, and within-group agreement should be comparable to previous studies.

A note on the design of the experiment: The original idea was to do a between-groups comparison against the data reported in Rosch (1975). However, typicality data for four categories collected using Rosch's method was also available locally. The inclusion of this third source of data (Group A below) allowed for more detailed within-group comparisons, since Rosch only reported the mean typicality.

Methods

Participants

The Group A participants were 32 student volunteers, 12 male and 20 female, with a mean age of 25.3 years. They performed the task either for extra academic credit or for pay. The Group B participants were 31 student volunteers, 13 male and 18 female, with a mean age of 20.5 years. They performed the task for extra academic credit.

Stimuli

The Group A participants rated 20 items in each of four categories, named "animal", "monster", "fish", and "superhero". The 20 stimuli for these categories were selected from an associative frequency study in which participants generated exemplars for each of the categories. The Group B participants saw the same stimuli as Group A plus stimuli from another 6 categories originally studied by Rosch (1975). 20 items were selected at random from Rosch's lists for the categories named "furniture", "weapon", "vegetable", "sport", "toy", and "clothing".

Group A Procedure

For both groups, the instructions were very similar to those used by Rosch (1975) – see the Appendix for an example. Group A participants provided their ratings on paper in a booklet. Each page had a category name printed at the top and items in that category listed on the left hand side. Each item had a 7-point scale for rating goodness of example, as well as an "X" for participants to circle if they felt the item did not belong in the category at all, and a "?" for participants to circle if they did not know what the item was. Participants received the items within the categories in one of two random orders, and received the categories in a random order. No more than 2 participants saw the same random ordering of categories.

Group B Procedure

Group B participants provided ratings on a computer keyboard. Stimuli were displayed in black on a white background. On each trial, participants first saw a category name (all lowercase) displayed using a 24-point bold-faced Arial font, centered horizontally, and 50 pixels above vertical center on a 15 inch screen. This was followed, after 500 ms, by the name of the item to rate (all lowercase), centered both horizontally and vertically and displayed using a 30-point bold-faced Arial font with no indefinite article. The category name remained on the screen during the entire trial, and participants had an unlimited time to respond, using a standard computer keyboard with keys relabeled to present a 7-point response scale numbered in increasing order from left to right as well as the "X" and "?" options. The participants were given 10 practice trials involving category names and items not included in the main study. The main block of trials involved all 200

category/item pairs presented in a different random order for each participant.

Results

To assess agreement between groups, Pearson correlations of mean typicality between groups were computed for each of the four categories rated by both Groups A and B. These correlations were substantial and significant, ranging from $r = 0.91$, $p < .01$ to $r = 0.95$, $p < .01$. Table 1 displays these correlations along with the correlations of mean typicality between the Group B participants and Rosch's (1975) 209 participants. Again all correlations are significant and substantial, ranging from $r = -0.81$, $p < .01$, to $r = -0.97$, $p < .01$, (these correlations are negative because Rosch used an inverted 7-point scale).

Table 1: Mean Typicality Correlations Between Groups

Category	A vs. B	B vs. Rosch
Animal	0.95	
Clothing		-0.94
Fish	0.90	
Furniture		-0.93
Monster	0.91	
Sport		-0.89
Superhero	0.92	
Toy		-0.81
Vegetable		-0.84
Weapon		-0.97

As for within-group agreement, Table 2 shows the available data for split-half correlations over all 10 categories. No exact values are available for Rosch's six categories, except for her report that all split-half correlations were above $r = 0.90$. The values from Groups A and B reflect split-half correlations based on random splits. The mean split-half correlation for the four categories rated by both Groups A and B is 0.84 for Group A and 0.87 for Group B (the overall mean for Group B was 0.90).

Table 2: Random Split-half Typicality Correlations

Category	Rosch	Group A	Group B
Animal		0.94	0.94
Clothing	>0.90		0.97
Fish		0.78	0.84
Furniture	>0.90		0.87
Monster		0.81	0.78
Sport	>0.90		0.98
Superhero		0.83	0.93
Toy	>0.90		0.94
Vegetable	>0.90		0.83
Weapon	>0.90		0.96

Table 3 shows the available mean between-participant correlations, computed using the method of Guildford and Fruchter (1973) as suggested by Barsalou (1987). They range from 0.22 to 0.70, but most fall within the range of 0.30 to 0.60 reported by Barsalou. The mean of these mean correlations for the four categories rated by both Groups A and B is 0.46 for Group A and 0.39 for Group B (the overall mean for Group B was 0.51).

Table 3: Mean Pair-wise Typicality Correlations

Category	Grp. A	Grp. B
Animal	0.64	0.52
Clothing		0.70
Fish	0.52	0.39
Furniture		0.64
Monster	0.24	0.22
Sport		0.58
Superhero	0.45	0.42
Toy		0.43
Vegetable		0.61
Weapon		0.56

Discussion

For all categories, the between-group mean typicality correlations shown in Table 1 are substantial and significant (most striking are the correlations between Group B and Rosch's participants, separated by 30 years and 3000 miles). For both Groups A and B, within-group agreement, whether measured by split-half correlations (Table 2) or mean correlations between participants (Table 3), is comparable to that reported in previous studies. One potentially anomalous result is that the mean between-participant agreement measured using Barsalou's method appeared to be a little lower in Group B than in Group A on the four categories that could be compared. But over all of the 10 categories rated by the Group B participants, most mean between-participant correlations were well within the range reported by Barsalou for this statistic and several exceeded it. So the appearance of lower within-subject agreement is probably an artifact of the choice of categories given to Group A. In particular, the words "fish" and "animal" are ambiguous in a way that the other category words are not. In common usage, "fish" can mean a kind of organism or a kind of meat, and "animal" can include humans or not, and sometimes refers only to mammals. These different meanings correspond to quite different concepts. So the lower agreement in Group B likely stems from confusion over which concept was intended, not from context-dependent modifications to a default concept. In summary, there is little evidence to be found in this study that modifications to the context of the typicality rating task have a significant effect on the structure or content of the concepts employed by the participants.

Conclusions

I argued that Barsalou and Prinz's claims for concept instability rest on a bad interpretation of previous experimental results. But past studies on family resemblance were conducted in such a way that unstable concepts could not entirely be ruled out. The experiment described in this paper provided evidence that even when the immediate context of the task is disrupted, neither within- nor between-group agreement is affected. It is reasonable to conclude from this that participants are in fact able to access an invariant concept when presented with a category name, and thus there is little reason to believe that the massive flexibility of thought is due to flexibility or instability built into the concepts themselves. Variability in categorization and other tasks likely results from interaction effects between a number of cognitive systems and sources of knowledge.

Acknowledgements

Thanks to Robert J. Stainton and Craig Leth-Steensen for repeated discussion of the material presented in this paper, and to Edina Torlaković for comments on an earlier draft.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, 11(3), 211-227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal Experimental Psychology: Learning Memory and Cognition*, 11(4), 629-654.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.) *Concepts and Conceptual Development.* Cambridge University Press.

Barsalou, L. W. (1989). Intra-concept similarity and its implications for inter-concept similarity. In S. Vosniadou & A. Ortony (Eds.) *Similarity and Analogical Reasoning*. Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Bran Sciences*, 22, 577-660.

Bellezza, F. S. (1984). Reliability of retrieval from semantic memory: Common categories. *Bulletin of the Psychonomic Society*, 22(4), 324-326.

Bellezza, F. S. (1984). Reliability of retrieval from semantic memory: Information about people. *Bulletin of the Psychonomic Society*, 22(6), 511-513.

Bellezza, F. S. (1984). Reliability of retrieval from semantic memory: Noun meanings. *Bulletin of the Psychonomic Society*, 22(5), 377-380.

Guildford, J. P., & Fruchter, B. (1973). *Fundamental Statistics in Psychology and Education*. New York: McGraw Hill.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18, 441-461.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *J Verb Learn Verb Be*, 23, 250-269.

Prinz, J. J. (2002). *Furnishing the Mind: Concepts and their Conceptual Basis.* Cambridge, MA: MIT Press.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192-233.

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

Roth, E. M., & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology*, 15, 346-378.

Appendix

The following instructions were given to Group B Participants:

> This study has to do with what we have in mind when we use words which refer to categories. Let's take the word *red* as an example. Imagine a true red. Now imagine an orangish red … imagine a purple red. Although you might still name the orange-red or the purple-red with the term *red*, they are not as good examples of red (not as clear cases of what *red* refers to) as the clear "true" red. In short, some reds are redder than others. The same is true for other kinds of categories. Think of birds. Everyone has some notion of what a "real bird", a "birdy bird" is. To me a sparrow or a robin is a very birdy bird while an ostrich is a less birdy bird. Notice that this kind of judgment has nothing to do with how well you like the thing; you can like a purple red better than a true red but still recognize that the color you like is not a true red. You might think that the ostrich is much more interesting than other kinds of birds without thinking that it is the kind of bird that best represents what people mean by birdiness. This judgment also has nothing to do with how frequently you see or think about the thing; you could live on an ostrich farm and deal with ostriches every day, but still think that they are a pretty bad example of what people generally mean when they talk about birds.
>
> In this study you are asked to judge how good an example of a category various instances of that category are. The computer will present a category followed by an item. You are to use the buttons labeled 1 to 7 to rate how good an example of the category each instance is. A 7 means that you feel the item is a very good example of your idea of what the category is. A 1 means you feel the member fits very poorly with your idea or image of the category. A 4 means you feel the member fits moderately well. It is also possible that you may not know what the example is, in which case you should press the button labeled "?". You also may feel that the instance is not actually a member of the category at all, in which case you would press the button marked "X". You can go at your own speed, and you will be given a chance to practice before the experiment starts.
>
> Don't worry about *why* you feel that something is or isn't a good example of the category (and don't worry about whether it's just you or people in general who feel that way) – just mark it the way you see it.