# Ideals Aren't Always Typical: Dissociating Goodness-of-Exemplar From Typicality Judgments

**Aniket Kittur (nkittur@ucla.edu)**
**Keith J. Holyoak (holyoak@lifesci.ucla.edu)**
Department of Psychology, University of California, Los Angeles CA 90095

**John E. Hummel (jehummel@psych.uiuc.edu)**
Department of Psychology, University of Illinois, Urbana Champagne, Urbana IL 61820

**Abstract**

Items that are rated good examples of a category have generally been assumed to be highly typical as well. However, most previous studies have used categories defined by simple features in which exemplar goodness and typicality are strongly related. We report a study using categories based on the relationships between features instead of the features themselves, allowing manipulation of relational ideals independent of featural central tendencies. Goodness-of-exemplar (GOE) judgments were based on relational ideals, whereas typicality judgments were based on a mix of ideals and featural central tendencies. These results indicate that exemplar goodness and typicality can lead to two distinct forms of graded category structure, and should not be treated as equivalent.

**Keywords:** categorization, category learning, similarity, relations, typicality, goodness-of-exemplar

## Introduction

One of the most robust findings in categorization research is the graded structure of categories. Every member of a category is not considered an equally good example of the category; instead, items lie on a spectrum of category goodness (Rips, Shoben, & Smith, 1973; Rosch & Mervis, 1975). For example, a robin is considered by American undergraduates to be a better example of a bird than is an ostrich or a penguin.

This graded goodness-of-example effect is known as "typicality", and there is a large body of work supporting its influence on categorization. Object classification speed increases with typicality (Rips et al., 1973); for example, people are fast to affirm that a robin is a bird, but slower to affirm that a chicken (a less typical bird) is. More typical items are generated before less typical ones (Mervis, Catlin & Rosch, 1976). People are more likely to extend inferences when the source of the inference is a typical category member rather than an atypical one (Rips, 1975). Category learning is faster when typical rather than atypical items are presented earlier in the sequence (Mervis & Pani, 1980; see also Posner & Keele, 1968). In fact, the prevalence of typicality effects in categorization has led some researchers to say that "if one compares different category members and does *not* find an effect of typicality, it suggests that there is something wrong with – or at least unusual about – the experiment" (Murphy, 2002, p. 24).

The standard way in which typicality is measured is through goodness-of-exemplar (GOE) judgments; for example, "How good an example is item A of category B?" This measure is so universally accepted that the concept of typicality is often introduced as synonymous with category goodness. For example, a classic paper on categorization asserts, "Instead of being equivalent, the members of a category vary in how good an example (or how typical) they are of their category" (Barsalou, 1985, p. 629).

The reason that typicality and GOE are so often considered equivalent is quite simple: in most studies of categorization they are indistinguishable. However (and, we argue, not coincidentally), most studies of categorization also use categories structured by central tendencies and represented by simple features.

### Central Tendencies versus Ideals

Central tendency views of category structure have been dominant for the last three decades. The exemplars of a category structured by central tendency are considered better members the more similar they are to the "center" of the category. This similarity measure may be discrete, such as the number of shared and unshared properties between exemplars represented as lists of features (Rosch & Mervis, 1975), or continuous, such as the distance from a dot pattern prototype (Posner & Keele, 1970). It may be calculated by the distance between two items in a hierarchy (Lynch, Coley, & Medin, 2000), by the distance between two points in a stimulus space (Ashby & Gott, 1988), or by many other methods. However, a fundamental assumption of a central tendency structure is that the further from the "center" an exemplar gets, the worse an example of the category it becomes. Importantly, we will refer to "central tendencies" here as those metrics that depend only on the distribution of individual features. These may include such metrics as familiarity, frequency of instantiation, shared features, and distance from a prototype or exemplars[1]. All of these can be calculated without reference to other categories, goals, or relations.

However, central tendencies do not appear to define all categories. Armstrong, Gleitman, and Gleitman (1983)

[1] We do not try to distinguish between prototype and exemplar theories here, as both predict that items further from the central tendency of the category will be considered worse exemplars.

found that even in categories with supposedly all-or-none membership (such as *female*), participants would still rate the GOE of within-category members differently (e.g., they would rate *mother* higher than *princess*). Bourne (1982) found a similar difference between categorization and GOE judgments in logically-defined categories of geometric stimuli. Such data create difficulties for views of category membership as defined by graded exemplar goodness or typicality, since membership in the category does not appear to be defined by the graded structure.

Furthermore, some categories have *ideal* exemplars distinct from their central tendencies as defined by frequency of feature occurrence. For example, the best example of a food to eat when on a diet may have zero calories, even though no food may actually achieve that ideal (Barsalou, 1983). Ideals have been shown to be a better determinant of category goodness than central tendencies in goal-based categories such as "foods to eat on a diet" and "ways to hide from the Mafia" (Barsalou, 1983, 1985). Ideals may define category goodness even in natural categories. Lynch et al. (2000) found that tree experts used ideals instead of central tendencies to make GOE ratings for individual trees, whereas novices used measures such as familiarity. Similarly, Burnett et al. (2005) showed that expert fishermen based their GOE judgments of freshwater fish on ideal desirability rather than centrality. Hampton (1981) found that central tendencies were not a good predictor of GOE for abstract categories. In a study using generated faces as stimuli, Goldstone, Steyvers and Rogosky (2003) showed that people tended to categorize based on centrality when the target category was learned independent of other categories, but were best able to categorize the extreme ideals (caricatures) when the category was learned in relation to a alternative contrast category.

To characterize category goodness in ideal-based categories, it does not suffice to consider the features of a single category's exemplars; rather, such categories appear to be defined by the relations their exemplars instantiate. For example, in both the Barsalou (1983, 1985) and Lynch et al. (2000) studies, categories were defined by the relations between an exemplar and the user's goals; and in the stimuli used by Goldstone et al. (2003), category goodness was defined in relation to the prototype of the competing category.

It seems that categories defined by individual features, or without reference to other categories, naturally lead to category structures based on central tendency, as the distribution of their features provides the only possible metric by which to measure exemplar similarity. However, when categories involve relations, two distinct measures become available: how well an exemplar fits a relational ideal, and the closeness of its features to the central tendency of the category distribution. In this case it is possible that typicality and GOE judgments are not identical, and instead measure different types of graded structure. Specifically, an item with high category "goodness" may lie on the extreme end of a graded scale, whereas a very "typical" item may be a very common one near the middle. Even when theorists have posited an influence of ideals on category goodness, the basic assumption that GOE judgments and typicality are identical has remained unchallenged. For example, Barsalou (1985) and Lynch et al. (2000) found evidence that ideals influence GOE judgments, but did not distinguish GOE and typicality.

The present study examined whether the assumption that GOE and typicality judgments are equivalent holds true in categories defined by relations rather than features. Participants first learned two contrasting relational categories through classification learning. They then made forced-choice GOE or typicality judgments (varied between-subjects) for stimuli in which relational ideal and featural central tendency information were independently manipulated. Differences in choice preferences across the two tasks would reveal non-equivalencies between typicality and category goodness judgments. Preferences for specific comparisons within task would provide information on how graded central tendencies relate to category membership.

## Method

**Subjects.** 90 University of California, Los Angeles undergraduates participated as partial fulfillment of a course requirement.

**Stimuli.** Simple geometric shape stimuli were chosen to minimize the effects of prior knowledge, expertise, and goals. The stimuli were composed of two overlapping shapes, an octagon and a square. Each category was defined by the value of a single relation between the octagon and the square; e.g., the octagon being larger than the square might define an exemplar as a member of category A, whereas being smaller than the square would define it as a member of category B (see Figure 1). Either the relative size or the relative shade of the two shapes was used to define the categories (between-subject manipulation). Stimuli were presented and responses collected using a custom Matlab script on Apple Macintosh computers.

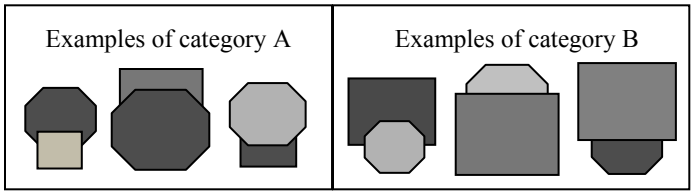


**Figure 1.** Examples of category members where the defining relation is relative size. These are only a small subset of the possible exemplars for each category.

**Procedure.** Participants first learned to classify stimuli into their respective categories through trial and error, with feedback after each trial. They performed a minimum of 200 classification trials, stopping after they reached a criterion of 12 trials correct in a row. During this phase stimuli were generated from a limited range of feature values.

Following classification learning, participants engaged in a forced-choice test phase in which they chose between two exemplars based on either typicality or on category goodness. For each exemplar, participants in the typicality condition were asked "which of these is more typical?" of a particular category, while those in the category goodness condition were asked "which of these is a better example?" of a given category.

Both the relational idealness and the featural central tendency of each exemplar could be separately manipulated during the test phase. An exemplar's defining relation could be either consistent (R+) or inconsistent (R-) with the given category. It could also have features that were from either the limited range of values seen during training (F+) or extreme values outside of it (F-). The resulting exemplar possibilities (R+F+, R-F+, R+F-, R-F-) were crossed to generate six possible combinations of exemplar pairs (see Figure 2). Two exemplar pairs were different in both their features and relations, two differed only in their relations, and two differed only in their features. Each participant judged 128 exemplar pairs, with the order of the pairs and the placement of each exemplar in the pair pseudorandomized for each participant.

Two key groups of comparisons test how GOE and typicality judgments relate to relational and featural information. The first group consists of the four comparisons in which one exemplar has a category-consistent relation while the other exemplar's relation is category-inconsistent (the four comparisons on the left in Figure 2). This comparison set allows us to examine the interactions between relations and features by always having the relations differ while manipulating feature centrality. For example, if GOE judgments relied on relational information but typicality judgments relied on featural central tendencies (or vice versa), we would expect a difference in participant choice in the R+F- vs. R-F+ condition (with the central tendency measure favoring the R-F+ option and the relational measure favoring R+F-). Less marked differences would be manifested in an advantage in the R+F+ vs. R-F- comparison for the R+F+ item (which dominates on both features and relations), and in a disadvantage in the R+F- vs. R-F+ condition for the R+F- item (which would be poorer due to its non-central features).

The second comparison group comprises the two exemplar pairs in which the relations were either both consistent with the given category or both inconsistent. In such cases the only manipulated difference was whether the items had novel, extreme features or seen, central features. However, there is a subtle difference, which is evident from examining the exemplars in the R+F+ vs. R+F- pair (the top right pair in Figure 2). In terms of central tendencies the R+F+ item dominates: its features have been seen before, and they instantiate a relation of similar magnitude to those previously seen. However, the R+F- item better exemplifies the relational ideal that defines its category: its octagon is *more* bigger than its square than is the R+F+ item. Thus even though both items are members of the category, the two possibilities for graded structure lead to different possible choices, based on featural central tendencies (R+F+) or relational ideals (R+F-). If GOE judgments differ in within-category graded structure

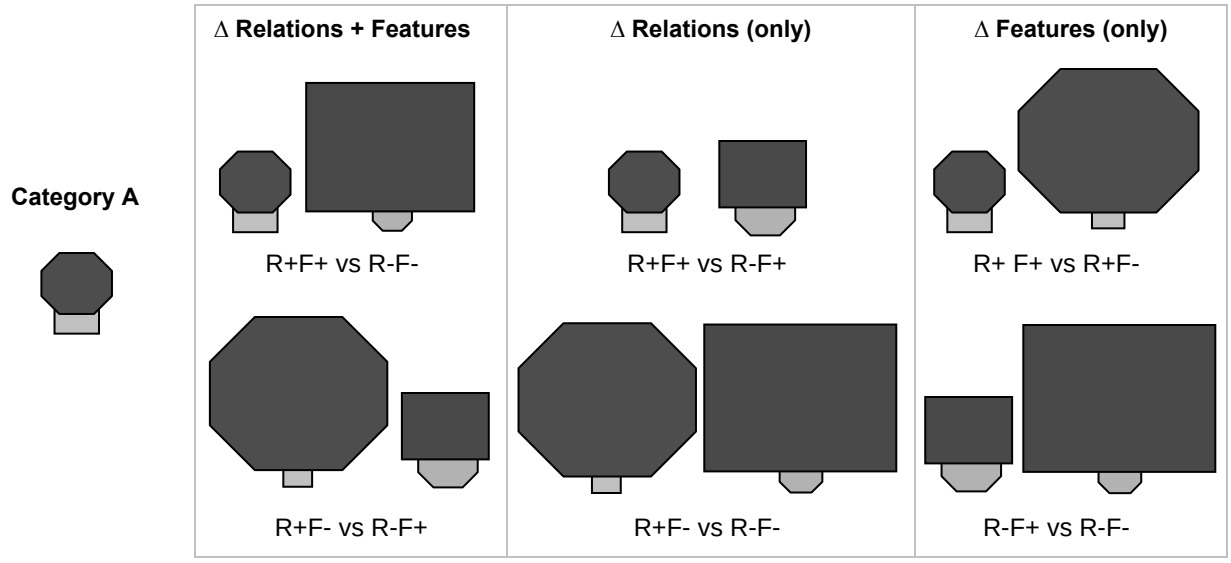


**Figure 2.** Diagram of **e**xemplar pairs combinations shown during test for a category defined by the octagon being bigger than the square. R+ = relation consistent with category; R- = relation inconsistent with category; F+ = seen, central features; F- = new, extreme features. Actual pairs were instantiated with heterogeneous feature and relation values; examples are for illustration purposes only.

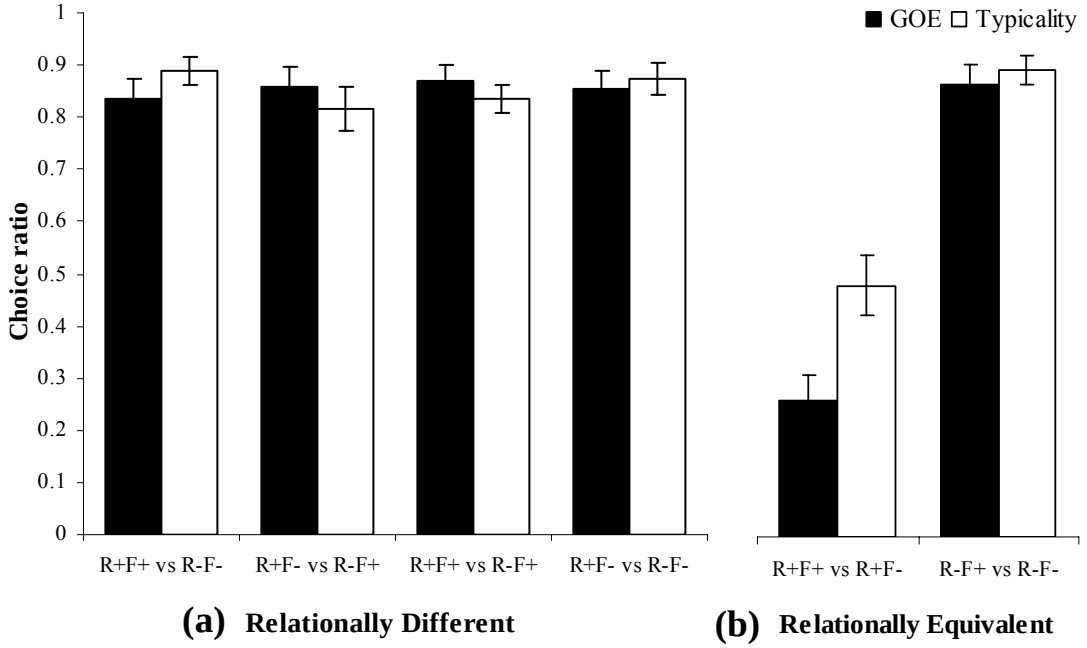


**Figure 3** (a). Proportion of responses favoring the relationally consistent item. R+ = relation consistent with category; R- = relation inconsistent with category; F+ = seen, central features; F- = new, extreme features. Error bars are one standard error above and below the mean. (b) Proportion of responses for items with seen, central features in comparisons where both items were either relationally consistent or inconsistent. The R+F+ vs. R+F- condition was a strong test between ideals and central tendencies.

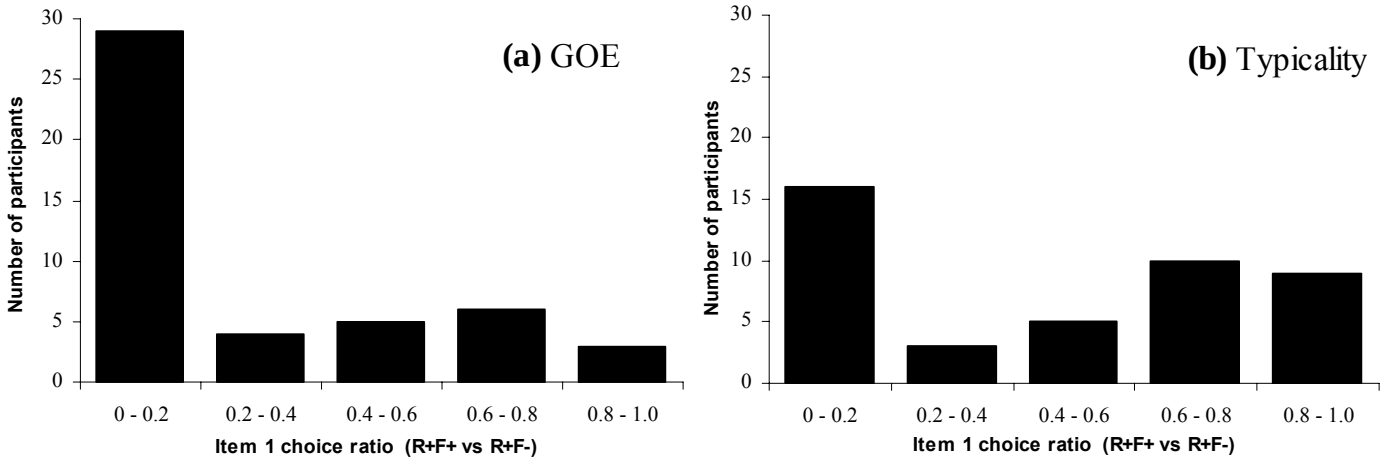


**Figure 4**. Histogram across quintile of mean (per subject) item responses to the R+F+ vs. R+F- comparison for (a) GOE and (b) typicality conditions. Low values indicate participants who consistently chose using ideals (the R+F- item) and high values indicate participants who based their choice consistently on central tendencies (the R+F+ item). Middle values correspond to participants who gave mixed responses.

from typicality judgments, we would expect to see a difference in participant choice for this comparison.

## Results

During the training phase there was no significant difference between the GOE and typicality groups on the number of trials needed to reach criterion, $t(89) = .19$, $p = 0.66$. This finding is unsurprising, since the training phases were identical for the two groups. However, it does verify that group differences in learning were not responsible for differences between conditions in the test phase.

There were two main groups of comparisons in the test phase: the four comparisons in which the relations differed between exemplars, and the two in which the relations were equivalent. Every comparison in the first group has one exemplar with a category-consistent relation and one with a category-inconsistent relation. By manipulating the featural information for each comparison we can test how featural central tendencies affect category membership decisions. Participants' choices for this group are shown in Figure 3a. Participants in both the GOE and typicality conditions showed a strong tendency to choose the exemplar with the

category-consistent relation (by *t*-test, all $p < .001$). A 2x4 ANOVA (response type x comparison) revealed no significant main effects or interactions ($p > 0.10$).

Of particular interest are choice judgments on the R+F- vs. R-F+ exemplar pair, which directly pitted relational category membership against featural central tendencies. For this pair choice preference was strongly based on relational category membership, with no discernible negative effect of poorer match to central tendencies in comparison to other conditions. Nor was there any noticeable enhancement effect in the R+F+ vs. R-F- case, where one exemplar dominated the other in both relational goodness and featural central tendencies. In summary, both GOE and typicality judgments appeared to be based exclusively on relational information when differentiating category members from non-members, with no measurable effect of featural information.

The second comparison group was comprised of the two exemplar pairs in which the relations were either both consistent with the given category or both inconsistent (the two comparisons on the right in Figure 2). In such cases the only manipulated difference was whether the items had novel, extreme features or seen, central features.

For the R-F+ vs. R-F- comparison both relations were inconsistent with the category; however, the R-F+ exemplar had central features while the R-F- exemplar had extreme features. In this case the R-F+ exemplar is both the better central tendency choice (with previously seen features) and the better ideal (since its relation is "less" against the ideal than that of the other exemplar). As expected, participants in both conditions showed significant preference for the R-F+ exemplar (see Figure 3b).

The R+F+ vs. R+F- exemplar pair provided the key comparison in this study. Any differences between central tendency and ideal judgments would likely be expressed here, as the R+F+ exemplar was the better central tendency choice while the R+F- exemplar was the better ideal choice (as its relation was of greater magnitude than the other exemplar). Since both were members of the same relational category, participants' choices would also reveal differences in within-category graded structure. Analysis of the R+F+ vs. R+F- comparison showed a significance difference in participants' choices (see Figure 3b). In the GOE condition there was a strong preference for the ideal (R+F-). This preference did not hold for the typicality condition, which generated mean choice judgments close to chance. However, this metric is insufficient to characterize participants' choices: it does not indicate whether the near-chance performance was due to participants changing their minds about which response to make within the session, or to different participants using different types of information to make consistent within-session responses that average to near-chance overall. To test these alternatives, we analyzed responses to the R+F+ vs. R+F- comparison on a per-subject basis. Each participants was assigned to a quintile corresponding to the mean of their responses, resulting in the histograms shown in Figure 4 The two distributions were significantly different, $\chi^2(16, N = 44) = 28.55$, $p = 0.027$. The bimodal distribution of participants in the typicality condition suggests that most participants consistently used either ideals or central tendencies to make their judgments, but few used a blend of strategies.

## Discussion

This experiment demonstrates that GOE judgments are distinct from judgments of typicality in categories not defined by featural central tendencies. Specifically, when graded category membership could be based on either relational ideals or featural central tendencies, participants asked to make GOE judgments based their responses on ideals whereas those asked to make typicality judgments used both ideals and central tendency information.

Importantly, although in typicality judgments many participants used feature central tendency information in deciding between two members of the same category, when the choice was between two members of different categories participants chose based on relational information – even if that meant choosing the item with the poorer match to central tendencies. Thus typicality only appeared to play a role in determining the graded structure within a category, whereas category membership operated at an earlier stage. These data suggest a two-stage model of categorization, in which items are first judged for category membership and then graded within the category. Furthermore, it is striking that graded structure can be based on either ideals or centrality depending on task demands, even though learning procedures for the GOE and typicality conditions were identical. Our results provide strong evidence that the two types of graded structure were learned in parallel.

These findings are inconsistent with the hypothesis that GOE judgments are equivalent to typicality, despite this equivalence being assumed in many studies of category learning. We believe there are two reasons why such a fundamental assumption has previously remained unchallenged. First, most studies examining typicality and GOE as separate measures used categories based on properties of individual features (e.g., Nosofsky, 1991; Rips & Collins, 1993). These studies found typicality and GOE judgments to be very similar. For example, Rips and Collins found that typicality and category likelihood were either indistinguishable or that typicality was a blend of category likelihood and similarity judgments (neither of which accounts for the data in the present experiment). These investigators were trying to distinguish between centrality and frequency, both of which are dependent on individual feature distributions. In contrast, we found that typicality and category goodness have different graded structures dependent on central tendencies and ideals.

Second, those studies that did use categories in which typicality and GOE can be separated (i.e., where category ideals do not match central tendencies) used GOE but not typicality ratings (e.g., Barsalou, 1983,1985; Lynch et al., 2000). For example, Lynch et al. found that GOE ratings for trees by experts were based on ideals such as tallness;

nonetheless, the title of their study begins: "Tall is typical." We argue that tallness may not be typical; instead, tallness may be an ideal used for GOE judgments, yet at the same time typicality could be based on central tendency measures.

Our data may also account for the puzzling findings mentioned of Armstrong et al. (1983) and Bourne (1982), in which graded GOE judgments were not consistent with all-or-none category membership judgments. The two-stage categorization process suggested here may account for these results, since exemplars are first categorized as members or nonmembers (according to their relations) and then rated for typicality or GOE (through graded structure). According to this hypothesis we would expect to see graded structure for within-category comparisons, and all-or-none judgments for between-category comparisons, which is the pattern found in our data.

In the present study we controlled for prior experience and knowledge by using artificial categories and stimuli, whereas most prior studies involving categories based on ideals have been based on user goals (Barsalou, 1983, 1985), prior knowledge, and pre-experimental expertise (Lynch et al., 2000; Proffitt et al., 2000). We believe that the present methodology provides an important step towards identifying the specific conditions giving rise to the dissociation between GOE and typicality judgments. However, it will be important to test whether the results found here generalize to real-world categories.

In conclusion, we have shown that GOE and typicality judgments are not the same, and may be based on different kinds of graded structure. Our results suggest a reinterpretation of typicality for those categories that are not based on properties of individual features. The present findings support a two-stage model of categorization, blending classical all-or-none category membership with multiple types of within-category graded structure. More work is needed to characterize the interactions of feature-based and relational information as they relate to common measures of categorization such as typicality, GOE, category membership, and similarity.

## Acknowledgments

Preparation of this paper was supported by NSF grant SES-0350920 to KH. Special thanks to Patricia Cheng and Barbara Spellman for helpful comments and discussion.

## References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*, 263-308.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 33-53.

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

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.

Bourne, L. E. (1982). Typicality effects in logically defined categories. *Memory & Cognition, 10*, 3-9.

Burnett, R. C., Medin, D. L., Ross, N. O., & Blok, S. V. (2005). Ideal is typical. *Canadian Journal of Experimental Psychology, 59*, 3-10.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition*, *31*, 169-180.

Hampton, J. A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition, 9*, 149-156.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition*, *28*, 41-50.

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*, 283-284.

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology*, *12*, 496-522.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT Press.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19*, 131-150.

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

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*, 304-308.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*, 811-828.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665-681.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. J*ournal of Verbal Learning and Verbal Behavior, 12*, 1-20.

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