Kinds of Kinds: Sources of Category Coherence

Kenneth Jeffrey Kurtz (kjk@northwestern.edu)
Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Rd
Evanston, IL 60208-2710 USA

Abstract

A fundamental question in the study of concepts is what makes sets of examples cohere as categories. We present results of three studies designed to compare standard taxonomic categories with categories that take their meaning from relationships extending outside of the individual example. An exemplar generation task is used to differentiate relational categories from taxonomic kinds and to compare possible subtypes of extrinsically cohering categories based on goals or thematic contexts. Results provide strong support for the intrinsic—extrinsic distinction and reveal signatures of underlying organization among the types of categories investigated.

Introduction

Categories play a fundamental role in cognition. The internal structure of categories supports numerous functions including classification, prediction, and reasoning. Categories give rise to an extension: the set of examples in the world that are members. The coherence of a category is the meaningful basis according to which these members go together.

One traditional view holds that the correlational structure of the environment determines category coherence due to systematic patterns of within-category similarity and between-category difference (Rosch & Mervis, 1975). Murphy and Medin (1985) propose the theory-based view that challenges the idea that similarity itself explains category coherence. 'Respects' for similarity (i.e., a basis for the selection of features and weights) must be specified in order for concepts to exist as groups of *like* examples. Additionally, they argue that category representations are richer than lists of features and must include relationships that hold within and between examples of categories. The tension inherent in the need for a constrained, yet rich basis of category coherence poses a continued challenge to theorists (Goldstone, 1994).

A useful source of inspiration is structural alignment theory—which has proven successful as an account of comparison processes such as similarity and analogy (Gentner, 1983; Markman & Gentner, 1993; Gentner, Rattermann, & Forbus, 1993). This framework offers a perspective for addressing the question of category coherence in a manner in keeping with the theory view. Respects for similarity can arise from the process of aligning corresponding predicates of two structured representations (Gentner & Markman, 1997; Medin, Goldstone, & Gentner, 1993; Markman & Wisniewski, 1997). Relational similarity drives the alignment process and largely determines the quality of a match. In the same way that structural alignment theory looks to shared relations (more than attributes or objects) to explain similarity, we can look to relationships between objects as a source of category coherence. While theory-based categories may cohere around intrinsic relationships (like the causal link between genetic material and physical features), we focus here on relationships extending beyond the individual example. For instance, the category *barriers* consists of examples that conform to the relationship: BLOCKS (X, Y).

Barr and Caplan (1987) distinguish between the intrinsic properties of a category which are true of an example in isolation versus extrinsic features which hold only in relation to other objects. As an alternative to category members bound together by common intrinsic structure (relations or attributes), category coherence can be derived from relations extrinsic to individual examples. The extreme case of extrinsic coherence is relational categories like *barriers*—members cohere based on fulfilling a core relationship. The roles of X and Y in the blocking relation can be filled by anything—so as long as the relationship holds, membership is secure. The examples of a relational category may have few or no intrinsic properties in common with one another. In this sense, relational categories are akin to analogies. Both 'prison bars' and 'raging river' are members of the category *barrier*, despite their sharing no intrinsic similarity.

Another case in which the category coherence is extrinsic is Barsalou's (1983, 1985) ad-hoc or goal-derived categories which are organized around ideals (properties that optimally promote goal resolution) rather than central tendency. Again, categories such as *things to take out of the house in case of fire* violate the correlational structure of the environment since member examples have few properties in common. Goldstone (1995) makes a useful distinction between default and directed similarity; where the former is the basis of graded structure and broad inferential power of taxonomic categories, while the latter is the focal, context-specific sense of similarity underlying ad-hoc categories or analogical relationships.

Categories may also be grounded by properties beyond the individual example that are not specific relationships. As an example, consider *items associated*

with working at an office desk. Thematic categories consist of examples that tend to cluster or co-occur in particular contexts. As Wisniewski et al. (1996) noted, such categories are often expressed as mass noun superordinates – e.g., *groceries* or *workout equipment*. There may be other relations in addition to spatiotemporal contiguity between particular pairs within a thematic category, but the members need not share any particular similarity.

In the present work, we use an exemplar-generation task to investigate and compare these possible bases for what makes things "go together" as a category. The sources of category coherence are: default similarity to the central tendency, directed similarity to ideals or to a core relationship, and patterns of contiguity in spatio-temporal contexts. We believe that there are kinds of categories that are best explained in terms of each of these sources of coherence, while there are also categories that are grounded in mixed forms of coherence. Barsalou (1985) shows that ideals account for a significant portion of the variance in typicality not only for goal-derived categories, but also taxonomic categories. Furthermore, different kinds of experts have been shown to organize the taxonomic category *tree* in terms of ideals derived from their experience (Lynch, Coley, & Medin, 1999).

Our main goals are: 1) to assess the psychological reality of extrinsically cohering categories in contrast with the intrinsic coherence attributed to standard taxonomic categories; 2) compare types of extrinsically cohering categories; and 3) address the non-uniformity of coherence in real-world categories. We use an exemplar generation task along with several follow-up measures to determine whether these different posited sources of category coherence are made evident in the behavior of the category.

Experiment 1: Exemplar generation

We begin by using exemplar generation as a means of indexing category coherence. Through measuring the content and dynamics of responding, a picture can develop of the organization of the knowledge being accessed. This technique has been used sporadically in the categorization literature. Goal-derived categories have been shown to support exemplar generation, but produce less output and show a lesser degree of correlation of output dominance with typicality than taxonomic categories under short time intervals (Barsalou 1983, 1985; Vallee-Tourangeau, Anthony, & Austin, 1998). In addition, greater output consensus was found for taxonomic than ad-hoc categories.

Several lines of evidence lead to the prediction that taxonomic categories should be easier and more natural than relational categories. As noted above, relational categories are akin to analogies: their members need share only relational similarity, not overall literal similarity. In contrast, the members of taxonomic categories share overall similarity. For example, two instances of *vegetable* are likely to have considerable intrinsic similarity (seeds, skin, etc.), as well as some extrinsic similarity (sold in stores, provide nourishment, for people, etc.). There is considerable evidence that relational similarity is more difficult to access in memory than object similarity (Gentner et al, 1993; Holyoak & Koh, 1995); and acquired later in development (Gentner & Rattermann, 1991). Further, Barr & Caplan (1987) showed that categories characterized by extrinsic features possess a greater degree of graded structure—possibly due to lower category validity of extrinsic properties. Thus we expect that relational categories will be less fluent (fewer runs of responses with minimal inter-item delay), less generative (fewer responses produced), and less consistent (lower agreement between participants) than taxonomic categories.

Evidence on the behavior of goal-derived categories leads us to expect that they should also be less generative than taxonomic categories. The investigation of thematic categories is more exploratory. The fact that members of thematic categories – such as 'ticket' and 'popcorn' for *things associated with going to the movies* -- lack not only intrinsic, but even relational similarity, might lead one to expect low generativity. On the other hand, the fact that thematic associates share spatiotemporal contiguity suggests that members might readily prime one another.

Method

Participants. 75 undergraduates from Northwestern University served as participants in order to fulfill an introductory course requirement.

Materials and design. Eight category cues (see Table 1) were selected for four different types of categories: taxonomic (all count noun superordinates), thematic, goal-derived, and relational. Natural language labels for the categories were determined for optimal clarity. Each participant generated exemplars for two category cues of each type. Item assignment was accomplished by random selection. The experiment used a within-Ss design with four item conditions corresponding to the types of categories.

Procedure. Participants read a set of instructions appearing on the computer screen. They were told they would be shown category cues (words or phrases) and asked to generate as many examples as they could during 4-minute intervals. To illustrate the nature of the task, a sample was shown: examples of *beverage* include water and milk. Participants were asked to work

as quickly and accurately as possible. The category cue remained visible for the duration of the trial. Participants typed into a response window on the screen. For each response, the time of the initial keystroke and the time of typing the return key were recorded. All responses entered for that cue remained accessible to the participant in a history window. This is comparable to a pencil-and-paper version of a listing task, but allowed recording of precise timing information. The eight category cues were presented in a random order. The entire experiment took approximately 35 minutes.

Table 1: Categories used.

Taxonomic	an animal
	a plant
	a fruit
	a vegetable
	a vehicle
	a household appliance
	a type of dwelling
	a musical instrument
Thematic	an item associated with:
	dining out at a restaurant
	going to the movies
	working at an office desk
	preparing for sleep at night
	working out at the gym
	going to the beach
	a party
	taking an airplane trip
Goal-derived	an item to take on a camping trip
	an item to remove from the house in case of fire
	an item not to eat while dieting
	a picnic activity
	a thing to do for weekend entertainment
	a way to advertise something
	an item to sell at a garage sale
	a thing that makes someplace desirable to live
Relational	a weapon
	a trap
	a guide
	a signal
	a barrier
	a tool
	a filter
	a shield

Results

The results are summarized in Table 2. All analyses were conducted by item since each participant only responded to two out of the eight items in each condition. The number of presentations of each item was not equal due to the random selection, but the number of presentations of each type of item was equal. All items were presented at least seven times.

Two analyses of response dynamics were performed on the entire data set (N = 75): response fluency and clustering. *Response fluency* was a measure of how long it took to generate each response. The "downtime" between any two responses was computed as the amount of time between the initial keystroke of each response. Item fluency was determined by the median downtime between responses. Mean fluency (in milliseconds) varied across item condition, as shown in Table 2. Comparison of the means using a one-way ANOVA showed a reliable difference between conditions, $F(3,31) = 6.44$, $p = .002$. As predicted, the Relational condition ($M = 10832$) was significantly less fluent according to post-hoc comparisons (all such tests we report were performed using the Bonferroni correction) than both the Taxonomic ($M = 6370$), $p < .01$) and the Thematic ($M = 6793$), $p < .01$) conditions.

We assessed *clustering* of responses in several ways yielding convergent results. In one analysis, any two responses that occurred within less than 67% of the median downtime for all responses by that participant were considered to be clustered together. To measure the degree of clustering, a ratio was constructed between the number of clustered responses and the number of isolated responses. A value greater than one indicates more clustered than isolated responses.

The clustering ratios are shown in Table 2. A one-way ANOVA was used to assess the differences between conditions, $F(3,31) = 6.76$, $p = .001$. Post-hoc comparisons showed reliably less clustering for Relational ($M = .64$) than for Taxonomic categories ($M = 1.38$), $p < .02$. In addition, significant differences were found between Relational and Thematic ($M = 1.48$), $p < .01$, as well as between Goal-derived ($M = .82$) and Thematic ($p < .04$).

Table 2: Summary of Results of Experiment 1.

	Tax	Thematic	Goal	Relation
Productivity	23.2	21.9	19.1	14.2
Consensus	25.8	18.1	17.3	13.8
Paragons	5.1	2.1	1.1	0.8
Easy-Access	2.5	3.0	1.5	0.6
Fluency	6370	6793	8140	10832
Clustering	1.4	1.5	0.8	0.6

Analyses of response content. A subset of the responses (the first 46 of the 75 participants) was analyzed intensively using a scoring procedure performed by trained undergraduate research assistants. Responses were removed from the analysis on the basis of a clear failure to understand the task or to undertake it seriously. Repeated and blank responses were also removed. A conservative coding of responses was performed: pure synonyms, abbreviations, and minor

syntactic variations (e.g., singular versus plural) were treated as the same response.

Productivity or item output was measured as the mean number of responses produced. A one-way ANOVA was performed to test for differences between Goal-derived (M = 19.1), Relational (M = 14.2), Taxonomic (M = 23.2), and Thematic (M = 21.9). An effect of item condition on productivity was found $F(3,31) = 3.87$, $p = .02$. The effect appears to be driven by the low mean productivity in the Relational condition. Post-hoc comparisons revealed a significant difference between Relational and Taxonomic ($p < .03$). A marginal difference was also found between Relational and Thematic ($p < .07$).

In order to evaluate whether participants tended to generate the same responses to the categories, output *consensus* was measured in two ways. For each item, the percentage of participants who produced each response was computed and the mean was taken across all responses generated for that item. By this measure, output consensus varied as follows: Goal-derived (M = 17%), Relational (M = 14%), Taxonomic (M = 26%), and Thematic (M = 22%). A one-way ANOVA showed a reliable difference between item conditions $F(3,31) = 6.41$, $p = .002$. Post-hoc comparisons showed significant differences between Taxonomic and both Relational ($p = .001$) and Goal-derived ($p < .03$). The difference between Taxonomic and Thematic was marginally significant ($p < .07$).

As a convergent measure, output consensus was also analyzed by computing the percentage of responses that occurred frequently (generated by at least 60% of the participants receiving the item). Few responses were widely agreed upon by participants: Goal-derived (M = 4%), Relational (M = 2%), Taxonomic (M = 12%), and Thematic (M = 5%). Group means were compared using a one-way ANOVA that revealed an effect of condition on output consensus $F(3,31) = 4.50$, $p = .01$. On both measures, agreement was greatest for Taxonomic and lowest for Relational.

The content of the exemplar generation data showed particular responses that occurred with great regularity (produced by at least 85% of participants). To give an example from each type of category: 'car' was a paragon of the Taxonomic category *vehicle*, 'wall' was a paragon of the Relational category *barrier*, 'tent' and 'sleeping bag' were paragons of the Goal-derived category *an item to take on a camping trip*, and 'check' was a paragon of the Thematic category *an item associated with dining out at a restaurant*. The prevalence of such high-agreement responses was computed by counting the number of paragons for each item. This measure is reported as a frequency, not as a percentage of the total set of responses, since the presence of special responses is not likely to follow from the overall breadth of responding. The mean number of paragons is shown in Table 2. A one-way ANOVA showed a significant difference between groups, $F(3,31) = 4.36$, $p = .01$. As confirmed by post-hoc comparisons, Taxonomic categories yielded significantly more paragons than Relational ($p < .02$) or Goal-derived ($p < .04$).

In addition, certain responses were found to occur both early and often in the exemplar generation task. The presence of such easy-access items was determined according to mean list position (normalized by list length). Frequent responses with a mean position score of less than 0.3 were considered easy-access responses. Easy-access responses sometimes, but not always, corresponded with category paragons. For example, the easy-access responses for the Taxonomic category *vehicle* included the paragon 'car' plus 'truck.' 'Wall' was the only paragon as well as the only easy-access response for the Relational category *barrier*. 'Tent' and 'sleeping bag' were both paragons and easy-access responses for the Goal-derived category an *item to take on a camping trip*. For the Thematic category *an item associated with dining out at a restaurant*, the paragon was 'check', but the easy-access responses were 'waiter' and 'menu'. A one-way ANOVA showed an effect of item condition on the frequency of easy-access responses, $F(3,31) = 5.98$, $p < .005$. Post-hoc comparisons showed Relational categories produced reliably fewer easy-access items than Thematic ($p < .005$) and Taxonomic ($p < .03$) categories.

Discussion

A basic pattern can be discerned across the set of results. The Relational and Taxonomic categories are reliably different on nearly every measure tested. This provides strong support for the predicted differentiation of intrinsic and extrinsic forms of category coherence. The analysis of response content reveals that Relational categories (and to a lesser degree Goal-derived categories) are less productive, less consistent, and less likely to have paragons or easy-access responses. The thematic categories are distinct in the high frequency of easy-access responses.

Experiment 2a: Pairwise Similarity of generated exemplars

We suggested above that only the Taxonomic categories possess coherence based on intrinsic similarity. To confirm this claim in terms of the actual responses generated by participants, we obtained similarity ratings for within-category pairs.

Method

Participants. 37 undergraduates from Northwestern University served as participants in order to fulfill an introductory course requirement.

Materials. Stimulus materials were sets of the six responses generated with the highest consensus on half of the category items in Experiment 1 (the half selected were those items yielding the most high-consensus responses). All within-category pairs were tested.

Procedure. Participants received instructions to rate the similarity of pairs of items on a scale from low (1) to high (5). An example of both high and low similarity was provided. All possible within-category pairs were presented in pseudo-random order (no pairs from the same category were presented consecutively). Each pair was presented in random left-right order. Participants used the mouse to click on the button labeled with the numerical rating. A response could be changed by re-selecting before clicking "OK" to continue.

Results and Discussion

Mean pairwise similarity was computed across the fifteen within-category response pairs for each item. Across all participants, Taxonomic ($\underline{M}$ = 3.8) pairs showed the highest mean similarity while the other conditions were nearly equal: Relational ($\underline{M}$ = 2.3), Goal-derived ($\underline{M}$ = 2.4), Thematic ($\underline{M}$ = 2.4). This difference was confirmed by a one-way ANOVA, $\underline{F}(3,15) = 13.31$, $\underline{p} < .001$. Post-hoc comparisons showed highly significant differences between Taxonomic and the other conditions (all $\underline{p} < .01$).

The results are consistent with our prior findings in showing an advantage for Taxonomic categories over Relational categories. In addition, neither Thematic nor Goal-derived categories showed high within-category similarity. This pattern is consistent with the view that taxonomic categories are based on overall default similarity while the other types are grounded in alternate forms of category coherence.

Experiment 2b: Category transparency of generated exemplars

A measure of the nature of a category's coherence is how readily the common basis can be perceived given a large set of examples. In this study, participants were presented with sets of generated category examples and asked to say what they had in common. As before, an advantage was predicted for Taxonomic categories.

Method

Participants. 25 undergraduates from Northwestern University served as participants in order to fulfill an introductory course requirement.

Materials. The same materials were used as in Experiment 2a. Instead of pairs, the six high-consensus responses for each category were presented together. A packet was prepared with one page for each category. The six responses were displayed on the page in three staggered columns of two to minimize spatially organized sub-groups within the set. The order of the six responses was fixed (alphabetical). The pages of each packet were randomly ordered.

Procedure. On each page of the packet to be completed, participants were given a blank line on which to answer: "What would you say the following examples have in common?" Additionally, three blank lines were provided to: "Try to think of a few more examples that fit well with the group."

Results and Discussion

The results of the commonality task are of principle interest since participants were almost always able to generate consistent additional examples. Participants routinely interpreted the commonality judgement as if the task were to induce the category from the examples. Category transparency for each item was computed as the percentage of participants whose response was scored as a match to the initial category cue used in Experiment 1. Responses that captured the meaning of the category, but differed in word choice, were accepted as matches. Taxonomic ($\underline{M}$ = 100%) and Thematic ($\underline{M}$ = 97%) items showed very high mean category transparency. Relational ($\underline{M}$ = 74%) and Goal-derived ($\underline{M}$ = 68%) items showed considerably less transparency. A one-way ANOVA revealed an effect of item condition, $\underline{F}(3,15) = 3.71$, $\underline{p} < .05$. A planned contrast between Relational and Taxonomic showed a reliable condition difference, $\underline{t}(12) = 2.19$, $\underline{p} < .05$.

Taxonomic categories are highly transparent, as should follow from their high intrinsic similarity. In contrast, Relational categories, as expected from their extrinsic similarity grounding, show lower transparency—likewise for Goal-derived categories. While participants were sometimes able to instantiate these original categories from the bottom-up, there were frequent failures as well. The above three types behaved consistently on pairwise similarity and category transparency tasks. However, Thematic categories showed a marked difference: despite low inter-item similarity (Experiment 2a), the connection among the group as a whole was highly transparent. We conjecture that the multiple examples in the current task invited participants to instantiate unifying spatiotemporal contexts.

General Discussion

Results across the two studies strongly bear out our predictions. Taxonomic categories show high intrinsic similarity and all the many advantages in terms of fluency and generativity which follow. Relational categories are markedly less similar, less transparent,

and less generative. The remaining two kinds of categories are intermediate. Goal-derived categories often pattern with Relational categories—not surprisingly, since relations link objects to goals. Thematic categories are in some sense the outlier; while they are highly fluent, they are grounded not in commonality, but in associativity.

Before discussing the implications of these data, there are some concerns to be addressed. Our choice of items represents our best effort to capture each type, but some factors were not precisely controlled. One issue is whether lower production can be attributed to smaller set size. Unfortunately, establishing the size of a relational category is not straightforward; e.g., should examples such as 'lack of education' (listed under barrier) be included? Relational categories may include more abstract or less familiar examples. These factors could play a role in generation.

To summarize, traditional explanations of real-world categories have appealed to feature overlap and the correlational structure of the environment. The emphasis suggested by the theory view of concepts on relations within and between category examples and the success of the structural alignment account of psychological similarity point toward a key role for relations underlying category coherence. The research reported here shows that extrinsic coherence can support categorical organization and points to individual signatures for different kinds of categories.

Given that relational categories appear to bring up the rear on all our measures, should we draw the implication that such categories are not psychologically real or natural? We would answer No, and Yes. Relational categories are indeed less natural than categories based on overall similarity; they do not provide a first-order basis for making sense of the world. But they provide structural organizers for understanding the world in ways that cross-cut object-based categories. We suggest that categories such as *barrier*, *operator*, and *catalyst*, though they may never be as facile as object categories, pay their way as tools of cognition.

Acknowledgments

This research was supported by an NIH-NRSA post-doctoral fellowship to Ken Kurtz and by NSF Grant SBR-95-11757 to D. Gentner. This paper was partially prepared while D. Gentner was a fellow at the Center for Advanced Study in the Behavioral Sciences with support from the William T. Grant Foundation, award #95167795. We thank Melissa Wu for contributions to the design of experiment, Adrienne Rosen, Evan Ransom, and Jessica Goethals for help in processing the data, Jeremy Cloud for software design, and Kathleen Braun for her help throughout the project, and the Similarity and Analogy group at Northwestern.

References

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, & Cognition, 11,* 629-649.

Barr, R.A, & Caplan, L.J. (1987). Category representations and their implications for category structure. *Memory & Cognition, 15,* 397-418.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52,* 45-56.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65,* 263-297.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability and inferential soundness. *Cognitive Psychology, 25,* 524-575.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition 52,* 125-157.

Goldstone, R.L. (1995). Mainstream and avant-garde similarity, *Psychologica Belgica, 35,* 145-165.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332-340.

Lynch, E. B., Coley, J. D., and Medin, D. L. (1999). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition, 28,* 41-50.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431-467.

Markman, A. B., & Wisniewski, E. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 54-70.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-278.

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

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

Wisniewski, E. J., Imai, M., & Casey, L. (1996). On the equivalence of superordinate concepts. *Cognition, 60,* 269-298.

Vallee-Tourangeau, F., Anthony, S.H., Austin, N.G. (1998). Strategies for generating multiple instances of common and ad hoc categories. *Memory, 6,* 555-592.