



# City Research Online

## City St George's, University of London

**Citation:** Hampton, J. A. & Passanisi, A. (2016). When Intensions Do not Map Onto Extensions: Individual Differences in Conceptualization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(4), pp. 505-523. doi: 10.1037/xlm0000198

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/12699/>

**Link to published version:** <https://doi.org/10.1037/xlm0000198>

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

AQ: 1

# When Intensions Do not Map Onto Extensions: Individual Differences in Conceptualization

AQ: 2

AQ: au

James A. Hampton  
City University London

Alessia Passanisi  
“Kore” University of Enna

Concepts are represented in the mind through knowledge of their extensions (the class of items to which the concept applies) and intensions (features that distinguish that class of items). A common assumption among theories of concepts is that the 2 aspects are intimately related. Hence if there is systematic individual variation in concept representation, the variation should correlate between extensional and intensional measures. A pair of individuals with similar extensional beliefs about a given concept should also share similar intensional beliefs. To test this notion, exemplars (extensions) and features (intensions) of common categories were rated for typicality and importance respectively across 2 occasions. Within-subject consistency was greater than between-subjects consensus on each task, providing evidence for systematic individual variation. Furthermore, the similarity structure between individuals for each task was stable across occasions. However, across 5 samples, similarity between individuals for extensional judgments did not map onto similarity between individuals for intensional judgments. The results challenge the assumption common to many theories of conceptual representation that intensions determine extensions and support a hybrid view of concepts where there is a disconnection between the conceptual resources that are used for the 2 tasks.

AQ: 3

*Keywords:* prototype, concept, typicality, features, individual differences

In order to communicate successfully in everyday life, different individuals in the same language community must represent the meanings of words or concepts in much the same way. Indeed, research on semantic categorization has generally found good consensus on a range of tasks. Rosch and Mervis (1975), for example, reported high degrees of reliability in judgments of typicality (representativeness) of exemplars in semantic categories. Robins were very reliably considered by U.S. students to be more typical birds than were ostriches or penguins. Aggregated group data for typicality, category membership, and other measures of conceptual representation have also proved to be reliable predictors of performance at the individual level in many other cognitive tasks (Hampton, 1997; Murphy, 2002).

A high level of consensus in conceptual judgments is, however, still compatible with the existence of systematic individual variation. Barsalou (1987) reported that while the reliability for mean

typicality ratings may be above 0.90, the average correlation between two different individuals' typicality judgments was a much more modest .3 to .6. In contrast, the test–retest correlation between judgments given by the same individual was around .8. The fact that within-individual consistency is greater than between-individual consensus provides strong evidence for systematic individual variation showing that different people have different personal versions of particular concepts.

Our aim in this paper is to use this individual variation to test a central assumption of most theories of concept representation (Hampton, 2006; Murphy, 2002). Concepts have two aspects to them, an extension and an intension. The extension is the category of items (objects, actions, situations etc.) to which a term applies, while the intension is the set of features that those items in the extension typically share. It is almost universally assumed that for every day concepts such as fruit or sport, people use intensional information as a basis for classifying items in the extension. A banana is in the extension of fruit on account of its possession of the appropriate set of features that characterize the intension of fruit. At the same time, the intension of fruit is the set of features that pick out the corresponding class. Theories of concepts differ in how the intensional information is used to categorize the items, but they all assume an intimate connection between the two aspects of a concept, as will be detailed here.

The assumption to be tested here is that there is an integrated single representation of the concept, incorporating both extensional and intensional aspects. If this assumption is correct then we should find that individual variation in extensions should map onto individual variation in intensions. That is, if two individuals are more similar than average in their judgments of item typicality, then they should also be more similar than average in their judg-

AQ: 15 James A. Hampton, Department of Psychology, City University London; Alessia Passanisi, Faculty of Human and Social Sciences, “Kore” University of Enna.

We gratefully acknowledge Bob Rehder, Nicholas Shea, Steven Verheyen, and also Nat Hansen, Åsa Wikfors, and others at the Communication in Context: Shared Understanding in a Complex World (CCOM)

AQ: 16 Project workshop in Tallinn, Estonia, in October 2014 for their helpful comments and advice.

Correspondence concerning this article should be addressed to James A. Hampton, Department of Psychology, City University London, Northampton Square, London EC1V 0HB, United Kingdom. E-mail: [hampton@city.ac.uk](mailto:hampton@city.ac.uk)

ments about the relative importance of different features for defining the concept.

In order to test this assumption, it is first necessary to establish systematic individual variation for both extensional and intensional measures that is stable over time and specific to particular concepts. Once that has been established, then we can correlate individual pairwise similarities for a given concept across the two measures to see whether extensional and intensional variation correspond in the expected way.

### Evidence of Individual Variation in Concepts

In addition to Barsalou's (1987) studies of typicality judgments, there is substantial evidence for individual variation in extensions. McCloskey and Glucksberg (1978) had students categorize lists of exemplars in semantic categories, the lists including many borderline and uncertain cases (see also Hampton, Aina, Andersson, Mirza, & Parmar, 2012). The categorization task was repeated after a period of a month, and within-individual consistency was considerably greater than between-individual consensus. In a similar vein, Bellezza (1984a) had people generate category exemplars on two occasions a week apart. Within-subject consistency was .69, while between-subjects consensus was just .44. Individual variation in intensions has been less often studied, but Bellezza (1984b) found that when generating definitions of categories on two occasions within-subject consistency (.48) again exceeded between-subjects consensus (.22).

Further evidence for variation in people's categorization behavior has been reported recently by Verheyen and Storms (2013), in a development of the threshold model for categorization proposed by Verheyen, Hampton, and Storms (2010). The threshold model predicts the probability that an individual will place an item in a category using a logistic function involving two parameters, one reflecting the degree of membership of the item and the other the breadth of the individual's category. The greater the degree of membership, and the broader an individual's category, then the more probable is a positive categorization. Verheyen and Storms (2013) analyzed individual categorization patterns in eight semantic categories and found that for five of the categories, an improved fit could be obtained if participants were divided into two or more groups, each with its own associated item parameters. Thus, each group had a different way of ordering the items for membership within the category. When the different orderings were related to category features, in some cases the underlying basis for the group differences could be seen. For example, for sports, one group emphasized individual activities played indoors, categorizing darts, chess, and billiards as sports around 80% of the time, but hiking only 40% of the time. The second group showed the reverse pattern. In another study, Zee, Storms, and Verheyen (2014) showed that similar category exemplars (such as roller skates and skateboards as vehicles) tend to show correlated categorization. People tended to count either both, or neither to be vehicles (see also Hampton, 2006).

In general, it appears then that individuals have a relatively stable personal view of conceptual structure. For a range of measures, test-retest consistency is higher than between-individual consensus. As a preliminary aim, we first wished to extend this finding to another important measure of conceptual structure, namely feature importance. Given a set of features associated with

a concept, people are able to reliably judge how central or important they are to the meaning of a concept (Hampton, 1987; Sloman, Love, & Ahn, 1998). For example, being evidence-based may be considered more important to people's concept of science than involving laboratories, even though both are considered to be characteristic of the concept. Based on the evidence of variability in extensions, we expected to find corresponding evidence of individuals also having personalized ways of judging intensions via feature importance. This evidence should appear as greater within-individual consistency across occasions than between-individual consensus on a single occasion.

### Mapping Extensional and Intensional Similarity

As described earlier, assuming that there is reliable individual variation in feature importance judgments, our primary goal was to examine the relation between individual variation in extensions and intensions. Does individual variation in intensions map onto individual variation in extensions? If, for example, Verheyen and Storms (2013) had asked their participants also to rate or rank the features of a category for their importance in defining the concept, would they have found the same groups of individuals emerging, with those who favored darts and billiards actually endorsing personal skill as more important than physical exercise?

Many current theories of concept representation would predict such a mapping between intension and extension, providing that there is some systematic individual variation in concept representations to be mapped. For example, theories based on prototypes or causal schemas (e.g., Hampton, 2006; Murphy, 1993; Rosch & Mervis, 1975; Smith, Shoben & Rips, 1974; Verheyen, Hampton & Storms, 2010) propose that categorization and judgments of typicality are based on the degree of match between an item and a summary representation of the intensional features of a category. Rosch and Mervis (1975) found clear evidence that typicality is correlated with the possession of characteristic features. Robins and doves are typical birds because they have the physical and behavioral features that are typical of birds, whereas ostriches and penguins do not. Hampton (1979) established the same finding for degrees of membership in a category. By this account typical exemplars should have typical features, and typical features should be possessed by typical exemplars. Thus, if stable individual variation is found in extensions, it is expected that it should also be seen in intensions, and the two should match up.

Other models that integrate extensional and intensional aspects of concepts are Rogers and McClelland's (2004) parallel distributed processing network model, and causal-explanatory schema accounts (Murphy & Medin, 1985; Rips, 1989), where categorization involves finding a concept that best explains the features of an exemplar. In both cases, the connection between exemplar typicality and feature centrality or importance is plain. Likewise, essentialist accounts of concepts (Gelman, 2003; Medin & Ortony, 1989) posit that people categorize on the basis of observable features, which are taken to indicate the presence of the essence. Knowledge of which features are most diagnostic should affect which exemplars are considered most typical. It is safe to say then that the majority of theories of concepts would predict a positive mapping between individual variability in extensions and intensions.

Two possible exceptions to this prediction can be considered. First, there is evidence that for some concepts, such as kinship terms and biological kinds, different features may determine category membership as opposed to typicality (a proposal originally made by Smith, Shoben, & Rips, 1974). In that case, it would be possible that judgments of feature importance might reflect involvement in category membership decisions rather than typicality, so that variability in the two measures would not match up.

AQ: 6 Hampton (1998) analyzed the relation between likelihood of categorization and typicality in data published by McCloskey and Glucksberg (1978). For most categories, the relation was extremely close, such that the more typical an item was rated, the more likely it was to be included in a category (see also Verheyen & Storms, 2011). But biological categories contained some exceptions. For example, bats were moderately typical of birds, but were very unlikely to be included in the category. Dolphins and whales had a similar relation to the category fish. In general, however, the close relation between the two variables has been confirmed many times over (Hampton, Dubois, & Yeh, 2006; Verheyen, Hampton, & Storms, 2010), so that in broad terms one would still expect to see a match between variation in typicality judgments and variation in feature importance judgments, with the possibility of some category-specific differences. People should be able to differentiate among characteristic features such as flight or song for birds in terms of their importance for the category, and their beliefs about these features would be expected to be reflected in their beliefs about the relative typicality of category members that possess those features to differing degrees.

AQ: 7 A second possible exception to our prediction are exemplar models. Exemplar models for category learning (Hintzman, 1986; Nosofsky, 1984) typically have few resources for allowing variation in feature weights (although some, such as ALCOVE [or attention learning covering map], can learn dimensional weights for simple artificial stimuli of low dimensionality; Kruschke, 1992). They may therefore predict no connection between typicality and feature importance, simply because they do not represent feature importance. However, as a consequence, they would be unable to explain why people should systematically agree on which features are more important. They might therefore predict that consensus and stability of judgments concerning feature importance differences should be much lower than those for exemplar typicality. In any case, exemplar models have had limited application to concepts in semantic memory probably because these models have been mainly developed using category learning paradigms with highly impoverished stimulus domains and short learning histories. A notable exception is work by Storms and colleagues (Storms, 2004; Storms, De Boeck, & Ruts, 2000, 2001). These authors have provided evidence that categorization in semantic categories may proceed through determining similarity to stored subcategories rather than to a featural representation of the category itself. For example, a novel food may be categorized as a fruit rather than a vegetable because of its greater similarity to a particular known fruit, rather than because it possesses more features of fruit in general. This evidence suggests that superordinate categories may be represented in terms of a set of typical subcategories at the basic level, or possibly even memories of individual cases. The model still, however, proposes that similarity to a subcategory is likely to be based on the semantic features that participants typically generate when describing their concepts. As

such, it is quite consistent with the mapping between extensions and intensions that we predict, although there may be some way to derive a dissociation within this framework.

To summarize our theoretical position, we take it as a common assumption of theories of concept representation in psychology that extensional and intensional information are integrated into a single conceptual representation. If the data should fail to support this position, then a radically alternative theory is required in which the two types of representation are not well integrated, such that each is subject to independent individual variation. Such theories have not been developed in the psychological literature on concepts in semantic memory to date, although interestingly, there have been recent proposals in philosophy that could fit this result. Hybrid theories of concepts have been proposed in which prototypes, exemplars, and causal-explanatory schemas are represented as separate entities. Machery (2009) argued the most extreme position that these different aspects of conceptual thinking were in fact dissociated from each other to the point that the notion of “concept” ceases to have any scientific interest. Others have been more sanguine about the possibility of pluralistic or hybrid concept representations (Dove, 2009; Rice, 2015; Weiskopf, 2009). As yet, no new empirical evidence has been offered to support these positions. We will leave further discussion of hybrid theories to the final section, once the results of our studies have been presented.

AQ: 8

## The Current Studies

The procedure to be adopted in each of the studies to be reported was as follows. A sample of students was given two tasks to perform on each of a number of categories. One task was to judge exemplar typicality and the second was to judge feature importance. All analyses were performed separately for each of the categories. To test for a link between individual variation in extension and intension, we first examined the similarity structure between individuals for each task separately. Thus, for typicality judgments we generated a correlation matrix for participants showing the degree to which any two individuals gave a similar set of judgments. We then did the same for feature importance judgments, so that a second correlation matrix provided a picture of how similarly any two individuals saw the relative importance of different features. These correlation matrices will be termed “similarity matrices” to avoid confusion with other correlations that will be reported. By finally correlating these two similarity matrices, we planned to test for the existence of systematic individual differences in concept representation that are reflected in both intensional and extensional measures of concepts.

The sequence of studies to be reported goes as follows. Study 1 first established the stability of individual variation for the two tasks by measuring the similarity between individuals on each task on two occasions a week apart and correlating the similarity matrices. Having established that each task showed comparable levels of stable similarity differences, the data for the two occasions were pooled, and similarity matrices for the two tasks were then correlated to test the prediction of a relation between extensional and intensional representations. Study 2 provided two replications of the main findings of Study 1, using just a single point in time, while Study 3 was a closer replication of Study 1 with some methodological changes for greater generality. Finally a fourth study is reported, which, although not part of the current

project (it was conducted in the 1980s), provides a broader means of testing the generality of the results, as it used both different instructions and different materials.

### Study 1

As described earlier, a common finding in semantic memory is that within-subject consistency when a task is repeated after an interval is greater than between-subjects consensus. Study 1 aimed first to compare consistency and consensus for two tasks, an exemplar typicality rating task and a feature importance rating task. These will be termed the *typicality* and *importance* tasks. Once it could be established that within-subject consistency over time is greater than between-subjects consensus for each task, the analyses comparing similarity matrices for the two tasks could be performed. First, the similarity matrices for the same task across occasions were correlated to establish whether each task showed stable pairwise similarity variation. Then the data for each task were pooled across occasions, and similarity matrices for the two tasks were correlated with each other to test our main prediction of a connection between extensional and intensional individual variation.

### Method

**Participants.** Thirty students (27 females) at the “Kore” University of Enna in the Italian island of Sicily participated voluntarily by completing the two tasks in a classroom setting. Participants were aged between 23 and 43 years (mean age = 26.6;  $SD = 4.4$ ). Twenty-seven returned for the second test, and their data were initially retained for the analysis. After data screening (see Results, Data Cleaning section), the final sample size was reduced to 20.

AQ: 9

**Materials.** The study was conducted in Italian. Twelve exemplars ranging in typicality and 12 features ranging in importance were selected for six different categories from the norms provided in Verheyen and Storms (2013) and De Deyne et al. (2008). They were insects, sports, fish, tools, science, and vegetables. All six had also been used in Verheyen and Storms (2013) and were originally created by Hampton et al. (2006). Lists are provided in Appendix A. For typicality, in order to anchor the two ends with clear examples, and to provide a check on participant engagement with the task, a highly typical exemplar, and a clear nonmember were included as fillers. For importance, it was not deemed necessary to include anchors and 12 features were sampled at random from the 29 to 39 features listed in the norms. The two anchors for typicality were treated as fillers and were removed from the data prior to analysis. (In Study 3, which was a partial replication of this study, the anchor point items for typicality were incorporated in the analysis, and the list was reduced to 12 items for comparability with importance by removing two other items. Apart from a raised degree of consistency and consensus for typicality, the results were essentially the same).

In selecting the exemplars and features to use, there were two important criteria. On the one hand, the features needed to predict typicality and category membership of exemplars at a group level, as demonstrated in earlier research on prototypes (Hampton, 1979; Rosch & Mervis, 1975). On the other hand, the features and exemplars needed to be sensitive to possible individual differences

in concept representation, which means that exemplars were deliberately chosen to have a restricted range of moderate typicality, and features to have moderate importance to the concept. In order to demonstrate that the features were nonetheless predictive of typicality (and therefore good candidates for showing up individual variation in concept representations), an analysis is reported after the first three studies have been presented. The analysis shows that while possession of the features correlated with typicality in our restricted sample of 12 exemplars at a variable level across the categories used ( $.14 < r < .92$ ), when correlated with typicality for a larger more representative sample of exemplars, the correlation was uniformly high ( $.63 < r < .94$ ,  $M = .80$ ) and comparable with the population of features from which they were sampled ( $.72 < r < .88$ ,  $M = .82$ ). The features used were therefore a representative sample of the intensions of the concepts.

**Design and procedure.** Both tasks were based on a scale from 1 (*not typical, not important*) to 7 (*highly typical, highly important*). Each task was presented category by category, with the 12 (or 14) items listed within each category in alphabetical order. Half the participants did the typicality task first followed by the importance task, and half the reverse. Two weeks after the first testing session, participants repeated the same tests. The second testing occasion reversed the order in which the tasks were done, and the items within each category were ordered in reverse alphabetic order.

**Instructions.** Instructions for the tasks were based on earlier research (e.g., Hampton, 1987; Hampton & Gardiner, 1983). They included examples of ratings for a different category, not used in the main task. They were as follows (the original was in Italian):

*Typicality of membership:* In this booklet, you will find six category names, and under each one a list of words. Please give us a judgment of how representative or typical you think each word is of that category. Use a 7 to indicate something that is highly typical of the category, down to 1 for something that is unrelated to the category. Use the numbers in between for different degrees of typicality.

Examples of items and possible ratings were then given for the category “furniture,” to illustrate how the judgment should be made.

*Importance of features:* In this booklet, you will see six category names, and under each name a list of features that describe things that might be in that category. We want you to tell us how important you think each feature is for deciding whether something is in the category. Use 1 to indicate that some feature is completely unimportant, and numbers up to 7 to indicate that something is very important.

As for typicality, the instructions were supplemented by examples of features of “furniture” and possible ratings that they might be given and why.

### Results

Ratings of item typicality and feature importance for each of the two testing occasions were tabulated for each category. End anchors were removed for typicality, so that both scales were based on 12 items per category. A small number of ratings (less than 1%)

were left blank by participants and were replaced with the mean for the item. The data were analyzed category by category, so that with 6 categories and two tasks each administered on two occasions, there was a total of  $6 \times 2 \times 2 = 24$  datasets each consisting of a matrix of ratings from 27 participants for 12 items.

**Data cleaning.** For the purposes of considering the similarity between participants, it was particularly important to ensure that all participants engaged with the task and responded independently of each other, otherwise the similarity data would likely contain artifactual effects. A participant who was not attending to the tasks would end up being dissimilar from the other participants on both tasks and hence contribute to a positive correlation of the similarity matrices. Alternatively if participants had been checking each other’s responses, they would end up as a similar pair on both tasks, and likewise contribute to a positive correlation of similarities.

First, to check understanding and engagement, for each of the 24 datasets, the ratings of each participant were correlated with the average of the remainder of the group. Those with systematically poor engagement were defined as those having a correlation with the group of less than 0.2 for more than four of the 12 tests for a given task. Two participants were excluded based on this criterion, one for typicality and one for importance. Second, to test for independence, each of the 24 datasets were screened for “very high” correlations between individuals defined as  $r$  being greater than 0.95. Five participants with a large number of very high correlations (20 or more each across the 24 datasets) were excluded on the basis that we suspected some collusion between them in the classroom setting. For example one pair of participants taking the first typicality task across the six categories had one correlation of .98, three of .99, and one of 1.00. For a 7-point rating task across 12 items in each case, with mean agreement between individuals of around 0.4, this level of agreement is clearly “unexpected.” The experimenter running the study (A. P.) on recalling the occasion reported that the possibility of collusion could not be ruled out, in spite of her presence in the room. Given the clear evidence of extreme nonindependence in the data we chose to exclude the participants concerned.<sup>1</sup>

Fn1

**Reliability.** To validate our two measures, we wanted first to show that they constituted reliable scales. We expected that average ratings of typicality and importance should show reliable differences between items or features, indicating that the tasks were meaningful for participants and tapped into the same common representations (prior to looking for individual differences within these). Reliabilities for typicality and importance on each occasion were calculated. For typicality, across the six categories and two occasions, Cronbach’s alpha was between .89 and .96, with the exception of science which had values of .69 and .49 for the first and second test, respectively. Inspection revealed that this exception was the result of most exemplars receiving high typicality ratings, leading to lower variance and lower reliability. For importance, reliabilities were comparably high (.88 to .95) for all categories except this time for insects, which had alphas of .72 and .68 on the two occasions, again because of lower variance in the mean ratings. In general then, ratings were reliable (alpha around .9), except for science for typicality and insects for importance. (Recall that our rationale for selecting materials which would reflect individual differences meant that there could be restricted variance within our samples, leading to lower reliability.)

**Within-subject consistency and between-subject consensus.** The first aim was to replicate the earlier finding that within-subject

Table 1

*Correlation Between First and Second Occasions as a Measure of Within-Subject Consistency and Average Correlation Between Participants Averaged Across Occasions for Between-Subject Consensus, Study 1*

Category	Typicality		Importance	
	Within-subject consistency	Between-subject consensus	Within-subject consistency	Between-subject consensus
Insects	.57 (.25)	.33 (.30)	.41 (.32)	.10 (.32)
Sports	.69 (.24)	.47 (.27)	.52 (.37)	.36 (.30)
Fish	.64 (.25)	.35 (.28)	.67 (.32)	.30 (.33)
Tools	.68 (.24)	.46 (.24)	.58 (.27)	.32 (.32)
Science	.42 (.38)	.11 (.35)	.69 (.28)	.50 (.25)
Vegetables	.70 (.24)	.53 (.24)	.70 (.17)	.36 (.28)
<i>M (SD)</i>	.62 (.27)	.37 (.28)	.60 (.29)	.32 (.30)

consistency across occasions was greater than consensus between participants for typicality, and to discover whether the same pattern was true for importance. The columns headed “Within-subject consistency” of Table 1 show the mean and standard deviation for the within-subject correlations between the first and second test for each category averaged across participants. Typicality is shown on the left and importance on the right. Overall average within-subject consistency was 0.62, 95% confidence interval (CI) [.50, .74] for typicality and 0.60, 95% CI [.48, .72] for importance. Between-subjects consensus was calculated within the first and second sessions separately, and then averaged, and the results are shown in Table 1 in the columns headed “Between-subject consensus.” Values were lower than those for within-subject consistency in every case, with a mean of .37 (95% CI [.21, .53]) for typicality and .32 (95% CI [.19, .45]) for importance.

T1

AQ: 10

The first analysis therefore confirmed the presence of systematic individual responding for both typicality and importance tasks, and to a very similar degree in each task and for each category. Across 2 weeks, correlations between ratings for the same individual were around 0.6, whereas correlations between different individuals in the same testing session averaged around 0.35. The lower reliabilities noted above for typicality for the science category, and for importance for the insects category are reflected in Table 1, where both within- and between-subjects correlations were noticeably lower than for the other categories.

**Similarity matrices.** The second (and principal) aim of Study 1 was to test whether individual variation seen in the typicality task could be mapped on to that seen in the importance task, using the correlation between individuals as a measure of their similarity. Each of the six categories was analyzed independently as follows. Four  $20 \times 20$  participant similarity matrices were computed based on the correlation between participants, one for each of the two

<sup>1</sup> Based on an estimated correlation of 0.5 between individuals across the 12 ratings, and using Fisher’s  $Z$  transform, a correlation of 0.95 represents approximately 4  $SD$  above the mean, with an expected frequency of 3.2 per 10,000 correlations, or about 2.3 in the full dataset of  $25 \times 12 \times 24 = 7,200$  correlations. Our finding of stable similarity structure means that the correlations are not however randomly distributed, so one may expect more than this rate of high correlations in the sampled data. Figure 1 shows the observed distribution of pairwise correlations aggregated across all tasks.

tasks on each of the two occasions. Each of the resulting 24 similarity matrices thus showed the similarity between all pairs of participants in their ratings for that particular task, category and testing occasion. Each square matrix was reduced to a triangular table showing the  $(20 \times 19) / 2 = 190$  unique correlations between pairs of participants.

As a check on the statistical features of the data, before correlating the similarity matrices the distribution of similarities (the person-person correlations within each matrix) was plotted, and is included in Figure 1 for typicality and importance separately. (Figure 1, in fact, shows pooled data from Studies 1 through 3.) The distributions were quasnormal with a similar degree of negative skew ( $-.47$ ) in each case. Because a positive correlation was predicted between similarities for the two tasks, the matching direction and degree of skew means that skew would not restrict the degree of positive correlation achievable. Fisher's  $Z$  transform applied to the same set of correlations produced a distribution that replaced the negative skew with a slightly stronger positive one, but which was also quite strongly leptokurtic. We therefore proceeded with the analysis of similarities based on untransformed correlations.

First, to discover whether there were stable similarities and differences between individuals in their judgments, the 12 similarity matrices (2 tasks  $\times$  6 categories) from session one were correlated with their equivalents from session two, and the results are shown in the columns of Table 2 headed "Typicality–typicality" and "Importance–importance" with subheadings "Within-category." All 12 correlations were significantly greater than zero on a one-sample  $t$  test with  $df = 189$ , with a mean of .28 for typicality and .33 for importance. One-sample  $t$  tests across the six categories also showed that both means were clearly greater than zero,  $t(5) > 8.0$ ,  $p < .0001$ .<sup>2</sup>

From Table 2, we can conclude that there were stable and systematic similarities and differences between individuals, because on each task the similarity between participants on the first occasion could be used to predict their similarity on the second occasion. (Note that as before, typicality for science and importance for insects had rather lower values, owing to a restricted variance in the original ratings.)

Having established that there were reliable and systematic similarities and differences on each of the tasks considered separately, it then remained to compare the degree of similarity of any two individuals on one task with their degree of similarity on the other. For maximum power, the typicality ratings from the two sessions for each participant were averaged, as were the importance ratings. Similarity matrices between participants were then recalculated based on these combined ratings, one for typicality and one for importance, for each of the six categories separately. The typicality and importance similarity matrices were then correlated and the results are shown in the column of Table 2 headed "Typicality–importance" with subheading "Within-category." It is clear from the Table that the correlations of similarities across the two tasks were considerably lower ( $M = 0.07$ ) than those across sessions for the same task (.28 and .33). However, the mean was still significantly greater than zero on a one-tailed one-sample  $t$  test across the categories,  $t(5) = 2.12$ ,  $p < .05$  one-tailed, and two of the correlations were individually significantly greater than zero. So there was apparently some, albeit very weak, evidence for a correspondence between similarity structures for the two tasks.

**Testing for category specificity.** The results presented thus far showed substantial systematic individual variation on each of the two tasks (typicality and importance) as seen in the correlation of similarity matrices across occasions. Although weaker, there was also evidence for a significant relation between the two tasks, as seen in the significant positive correlation between the similarity matrices based on the typicality and importance judgments for at least two of the six categories.

However, further exploratory analysis of the data undermined this conclusion. In correlating the different matrices, we also generated a set of correlations looking at how similarity in ratings for one particular category correlated with those for other categories. Prima facie, there is no reason why, for example, the similarity of people's typicality ratings for fish should correlate with the similarity of their typicality ratings for sports. Surprisingly, these correlations, although small, were generally positive, and in some cases of comparable size to the correlations between tasks reported above. Reasons for these positive correlations may include different levels of motivation, or different ways in which the task instructions were interpreted. Participants with low motivation may respond more randomly and therefore have lower similarity to other participants on all category measures. Alternatively a group of participants who focus on one interpretation of the task (e.g., basing typicality on frequency rather than family resemblance) will show greater within-group similarity across all categories for that task.<sup>3</sup> In effect, showing that similarity structure is stable over time can only be taken as evidence for variation in actual conceptual contents, rather than more general factors, if the similarity is specific to a given category.

We therefore needed to institute a further control to test that the correlations between similarity matrices that we are concerned with were greater than the general background level of positive correlation seen between different categories. To implement this, within-category correlations were compared to between-category correlations for each of the measures. Results are shown in the "Between-category" columns of Table 2 for (a) the consistency of typicality similarity matrices across sessions (typicality–typicality), (b) the consistency of importance similarity matrices across sessions (importance–importance), and (c) the correlation between the similarity matrices for typicality and importance based on ratings averaged across sessions (typicality–importance). Independent  $t$  tests comparing the six within and 30 between category correlations for each measure, and using Fisher's  $Z$  transform showed significant differences for typicality–typicality,  $t(34) = 4.28$ ,  $p < .001$ , for importance–importance,  $t(34) = 4.25$ ,  $p < .001$ , and none for typicality–importance,  $t(34) = -0.38$ .

To summarize, the similarity matrices for the same task across occasions correlated significantly higher within than between categories. However the correlation of similarity matrices obtained from the two different tasks (.07) was no higher than expected from the baseline level of correlation observed across different categories (.09).

<sup>2</sup> Estimates of significance levels for correlations between similarity matrices should be treated with some caution given the lack of independence within the similarity matrices. Using the Fisher  $Z$  transform, and assuming  $n = 190$ , the estimated 95% CI for a correlation of .30 would correspond to the range  $r = .15$  to .42.

<sup>3</sup> We thank Bob Rehder for this suggestion.

F1

T2

Fn2

Fn3

## CATEGORIZATION DIFFERENCES

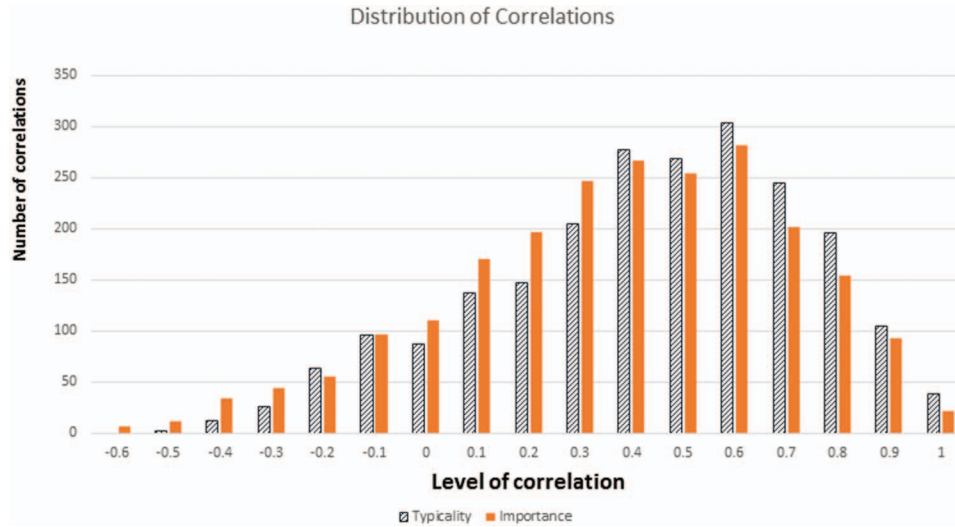


Figure 1. Distribution of correlations in the similarity matrices for the typicality and importance tasks across Studies 1 to 3. See the online article for the color version of this figure.

Our conclusion must therefore be altered. There were significant systematic similarities and differences between individuals in how they rated typicality of exemplars and importance of features, but there was no evidence that these similarities and differences corresponded between the two tasks.

**The effect of data cleaning.** A final check was run to test whether exclusion of the five participants with suspiciously high intercorrelations had an effect on the conclusions drawn. When all 25 participants were analyzed in the same manner (rather than the 20 selected earlier), the correlation between similarity structures for the two tasks within-categories did indeed increase, from .07 to .15. However so did the correlation *between* categories, from .09 to .17. Obviously, having one or more pairs of participants whose ratings on all tasks were much closer than expected increased the correlation of similarity matrices across the tasks. However, it did so in an indiscriminate way. In terms of consistency and consensus, including all 25 participants raised consensus to .39 and reduced consistency to .61.

## Discussion

The finding that there was no relationship between the systematic individual differences in typicality ratings and those in importance ratings is a challenge to theories of conceptual representation. Before discussing the implications, further studies were conducted to look for evidence that the relationship might be present, and to guard against a Type II error. Study 2 involved two samples of 30 students each, who did each task just once with the tasks separated by a period of 2 weeks. The aim was to remove any possible interference from doing both tasks in the same testing session, and to provide two replications of the test for the relation of similarity between intension and extension. Study 3 was broadly a replication of Study 1 using rankings rather than ratings and a slightly different selection of exemplars. Finally Study 4 reports some data from an earlier unpublished project with a number of important differences in method and with new materials.

## Study 2

Having established that similarity matrices for typicality and importance judgments are reasonably stable, correlating across occasions at around 0.3, the second study omitted the repeated testing element. Study 1 had participants make both ratings in the same session which raises the possibility that there would be some form of priming or interference taking place from one task to the other. Instead, for Study 2, each participant judged typicality and importance just once for each category. Half the participants did the typicality judgments on the first occasion, and importance on the second, and half did the reverse. The aim was again to see whether similarity between individuals on one task would map onto similarity on the other task.

## Method

**Participants.** Two samples of participants at the “Kore” University of Enna were run in a replication design. Thirty students (29 females) aged between 22 and 43 years ( $M = 25.3$ ;  $SD = 4.4$ ) took part in Study 2A. Study 2B involved 30 participants (23 females) aged between 21 and 41 years ( $M = 26.3$ ;  $SD = 4.7$ ) of whom one did not complete the task.

**Materials.** The materials used were the same as in Study 1.

**Design and procedure.** Participants were divided on Week 1 into two groups of 15. One group did the typicality task on Week 1, and the importance task on Week 2. The other group did the reverse. The instructions were the same as in Study 1.

## Results: Study 2A

**Data cleaning.** A small number (less than 1%) of responses were left blank, and were replaced by the mean for the item. Data were tabulated for each task and each category separately. Reliability of individual participants' ratings (alpha) ranged from .8 (insects) to .97 for importance, with five of six categories above .9,

**Table 2**  
*Correlation in Study 1 Between Similarity Matrices for the First and Second Testing Session for Each Task and Correlation Between the Similarity Matrices for Typicality and Importance Based on Ratings Aggregated Over Sessions*

	Typicality–typicality		Importance–importance		Typicality–importance	
	Within-category	Between-category	Within-category	Between-category	Within-category	Between-category
Insects	.39*	.06	.18*	.10	.20*	.08
Sports	.29*	.00	.26*	.08	.00	.07
Fish	.27*	.07	.40*	.12	.08	.05
Tools	.22*	–.01	.37*	.17*	.15*	.15*
Science	.20*	.03	.41*	.16*	.00	.09
Vegetables	.28*	.10	.34*	.12	.02	.08
<i>M (SD)</i>	.28 (.07)*	.04 (.13)	.33 (.09)*	.13 (.11)	.07 (.08)	.09 (.10)

Note. Critical value of  $r(188)$  for  $\alpha = .05$  is .143.

\*  $p < .05$ .

while for typicality, alpha was uniformly high (>.94). As in Study 1, the features for insects showed slightly lower reliability. No participants needed to be excluded on the basis of low individual-group correlations.

A second data cleaning analysis looked for the occurrence of very high correlations between participants, indicating that their responses were possibly not independent. For importance, there were only three out of 2,610 correlations over 0.95, but for typicality, there was an excess of high correlations for one participant who was excluded from the analysis. After this exclusion, there were just two high correlations in the set. The analysis was therefore performed on the remaining 29 participants.<sup>4</sup>

Fn4

**Analysis of similarity.** Because Study 2 only involved a single task on each occasion, just 12 similarity matrices were produced, one for each task for each of the six categories. There was a small but statistically significant background level of positive cross-category correlation for importance (mean  $r = .10$ ),  $t(15) = 4.50$ ,  $p < .001$ , but not for typicality (mean  $r = -.01$ ). Because of this general background level of positive correlation between similarities measured for different categories, to test for a correlation between the similarity matrices for the two tasks, we went directly to the comparison of within-category and between-category correlations. Table 3 shows the correlation between the similarity matrices for typicality and for importance, calculated within the same category (within-category) or between different

T3

**Table 3**  
*Within-Category and Between-Category Correlation of Similarity Matrices for Typicality and Importance Tasks for Studies 2A and 2B*

	Typicality–importance			
	Study 2A		Study 2B	
	Within-category	Between-category	Within-category	Between-category
Insects	–.01	.02	–.14	.03
Sports	–.10	.06	–.08	.06
Fish	.14	.05	.07	.03
Tools	.13	.06	–.13	.04
Science	.07	.02	–.01	.05
Vegetables	.07	.04	–.10	.03
<i>M (SD)</i>	.05 (.10)	.04 (.09)	–.07 (.08)	.04 (.10)

categories (between-category). The results for Study 2A are shown on the left side. Any degree of correspondence in individual differences for the two tasks should show up in greater levels of correlation within the same category than between different categories.

The first two columns of data in Table 3 show the correlations between similarity matrices for typicality and importance within each category and between categories. The six within-category correlations had almost the same mean (0.05) as the 30 between-category correlations (0.04), and both were close to zero. There was therefore no evidence for similarity between individuals on one task mapping onto similarity between individuals on the other task.

As a check, the analysis was also performed on the full sample of 30 participants. For the critical typicality–importance correlation of similarities, both within-category and between-category correlations averaged 0.04. Because the possible collusion occurred in only one task on this occasion, it had no effect on the level of correlation between similarity matrices for the two different tasks. (Recall that the tasks were done on different occasions.)

**Results: Study 2B**

**Data cleaning.** As in previous studies, a small number (less than 1%) of ratings were left blank, and were replaced by the item mean. The same two data cleaning analyses were run. Reliability of the six typicality scales ranged from .88 to .96, and for the importance scales from .91 to .97. On the basis of individual-group correlations, one participant was excluded who had correlations of less than 0.2 for four of the six categories for importance. The analysis of high pairwise correlations showed no problems with lack of independence in this sample, so no exclusions were needed on that account. The final analysis was therefore based on 28 participants.

AQ: 11

**Analysis of Similarity.** The last two columns of Table 3 show that the replication produced very similar results. The background positive correlation across categories was not significantly greater

<sup>4</sup> The occurrence of possible collusion in Study 1 was not discovered until after the running of Study 2. We had noted the unexpectedly high correlation between the similarity matrices for different categories, but it was only in the light of later studies that the source of the problem was identified.

than zero for either task alone. The correlation between similarity matrices for typicality and importance for this sample was actually lower for within-category correlations ( $-0.07$ ) than for between-category correlations ( $0.04$ ). Furthermore, there was no evidence of any systematic variation between the different categories.

## Discussion

Across two studies and three samples, no evidence has been found thus far to indicate any relation between the variety of ways in which people judge exemplar typicality and the variety of ways in which they judge feature importance. This failure cannot be explained by low reliability, as in Study 1, not only were the scales equally reliable, and the individual ratings reliably correlated across occasions (showing that the tasks were undertaken conscientiously), but the similarity matrices for each category and for each task across occasions were also significantly correlated above the background level of correlation sometimes seen across different categories.

## Study 3

In Study 3, we aimed to consolidate the results obtained to date. We returned to the design of Study 1, to confirm the systematic individual variation within each task across testing occasions 2 weeks apart, and we also introduced some small variations in the procedure. Some typicality items had several missing data points in the previous studies because of low familiarity (watercress is not commonly found in Sicily), and these items were removed. In addition, other typicality items were removed so that both exemplar and feature lists contained just 12 items per category. As a consequence, we dropped the strategy of excluding the top and bottom anchor items from the analysis. For both tasks, all 12 items were analyzed.

Another change was that instead of a rating from 1 to 7, participants were asked to provide a ranking from 1 to 12 for each list, so that similarity and difference between individuals simply in their use of the rating scale would be eliminated. For example, if some participants had given ratings across all categories with low variance, then their correlation with all others would be lower, owing to the effect of restricted variance on estimated correlation. Hence, they would be consistently less similar to others across both tasks (but also across categories). Using ranks, all participants necessarily have the same variance for their judgments across a list of 12 items.

## Method

**Participants.** Thirty participants (26 females, 87%) aged between 19 and 32 years (mean age: 21.70;  $SD = 3.63$ ) were recruited from the same population as the earlier studies. Six failed to complete the second test, and so 24 remained in the study.

**Materials.** The same lists of features for the six categories were used as previously for importance. For typicality, two items were removed from each list as shown in [Appendix A](#). The high and low anchor items were no longer treated as fillers but were included in the analysis.

**Procedure.** The procedure was similar to Study 1, except that ranks rather than ratings were given. Furthermore this time we

rewrote the instructions of the two tasks. A possible reason for the background level of positive correlations between similarity matrices for different categories may lie in some ambiguity in the task instructions. Participants may show similarity across categories because of adopting the same particular interpretation of how to do the task. In order to reduce this possibility, the instructions were made more specific. In particular we asked participants to focus on the interpretation of the scales most relevant to prototype and exemplar theory. For typicality, participants were asked to judge how similar an item was to other category members and how representative it was. For importance, we asked participants to judge the degree to which a feature was present in members of the category and absent in its contrast categories. The revised instructions (translated here from the Italian) were as follows:

*Typicality:* In this booklet, you will find six category names, and under each one a list of words. Please give us a judgment of how representative or typical you think each word is of that category. A typical example of a category is one that has a lot in common with the other members of the category; it would be a good example to represent what that category is normally like. Decide which is the most typical and place a 1 against that item. Then continue writing 2, 3, and so on, to order the items in terms of their typicality down to 12 for the least typical.

*Importance:* In this booklet, you will see six category names, and under each name a list of features that describe things that might be in that category. We want you to tell us how important you think each feature is for deciding whether something is in the category. An important feature is one that you find often in members of the category—most members have it, and that is also not so often found in other kinds of thing. So it is distinctive to that category. Decide which is the most important and place a 1 against that feature. Then continue writing 2, 3, and so on, to order the items in terms of their importance down to 12 for the least important.

As previously, both task instructions then included examples of exemplars or features for the category of furniture.

## Results

**Data cleaning.** There were no missing data to be replaced. Reliability of the 12 typicality scales ranged from .907 to .980, and for the 12 importance scales ranged from .800 to .948. One participant had low correlations with the group on nine of the 24 scales and was excluded. A check for independence showed no systematic pattern of pairs of participants with higher than expected correlations, so no further exclusions were needed. The final analysis was therefore based on 23 participants.

**Analysis of Similarity.** Study 1 found that for both typicality and importance, the correlation across 2 weeks for the same individual was greater than the average correlation between individuals, thus supporting the hypothesis that there are systematic individual differences in how people perform the tasks. The same analysis was applied to the data from Study 3, and the results are shown in [Table 4](#).

As in Study 1, for typicality, the level of within-subject consistency was greater ( $0.73$ , 95% CI [.63, .83]) than the level of

Table 4  
Correlations Across Occasions Within Participants and Average Correlation Between Participants Within an Occasion for Study 3

Category	Typicality		Importance	
	Within-subject consistency	Between-subject consensus	Within-subject consistency	Between-subject consensus
Insects	.66 (.32)	.52 (.28)	.51 (.27)	.16 (.33)
Sports	.82 (.15)	.70 (.18)	.70 (.17)	.45 (.25)
Fish	.70 (.31)	.34 (.29)	.59 (.20)	.29 (.31)
Tools	.77 (.12)	.59 (.21)	.38 (.40)	.26 (.31)
Science	.65 (.22)	.37 (.30)	.56 (.27)	.36 (.25)
Vegetables	.79 (.15)	.64 (.23)	.65 (.19)	.41 (.28)
<i>M (SD)</i>	.73 (.23)	.52 (.25)	.57 (.26)	.32 (.29)

between-subjects consensus (0.52, 95% CI [.37, .67]). The same was true for importance (within-subject consistency of 0.57, 95% CI [.45, .69] vs. between-subjects consensus of 0.32, 95% CI [.21, .43]). This pattern was found in all categories. Note that the increase in levels of correlation for typicality, compared to Study 1, can be explained by the inclusion in the analysis this time of the high and low anchor items, on which almost everyone agreed. The levels seen here are close to those reported by Barsalou (1987) for typicality ratings of .8 for consistency and .6 for consensus.

We next analyzed the stability of the similarity structure across individuals for each task, by correlating the similarity matrices for each task across occasions. As previously, correlations within categories were compared with correlations across different categories as a control for general response factors. The results are shown in the columns of Table 5 for the typicality task across occasions (typicality–typicality) and for the importance task across occasions (importance–importance).

The similarity structure proved reliable across occasions for both typicality ( $r = .44$  within categories vs.  $.10$  between categories),  $t(34) = 7.45, p < .001$ , and importance ( $r = .33$  within and  $.04$  between),  $t(34) = 6.78, p < .001$ . The similarity structure for each task was therefore reliably repeated on the second occasion. The final columns of Table 5 (“Typicality–importance”) then show the critical correlation between similarities on one task (averaged across occasions) and similarities on the other task. Although the

correlation between the similarity matrices across tasks for the same category (.15) was positive and greater than that for different categories (.05), this difference was not significant on an independent  $t$  test applied to the Fisher  $Z$ -transformed correlations,  $t(34) = 1.87, p = .07$ . Four of the six categories showed a positive difference, and two a negative difference. While insects, tools and vegetables showed a positive effect, the other three categories showed none.

**Proximity analysis.** As an additional check on the data and the method of analysis used, SPSS-Proxscal was used to compute proximities between participants for each task and for each category in two dimensional spaces. (multidimensional scaling aims to remove noise in the data by constraining similarities to map onto proximity in a space.) Proximities for the two tasks within the same category correlated on average at 0.11 ( $SD = .12$  across the six categories). Across different categories, the same two measures correlated on average at 0.04 ( $SD = .10$ ). This difference again failed to reach significance,  $t(34) = 1.47, p = .15$ . In contrast, comparing proximities for typicality judgments between the first and second test, correlations across the six categories averaged 0.32 within the same category and only .11 between different categories, while the same analysis for importance judgments gave average correlations of .27 within and .07 between. Both differences were significant on a  $t$  test,  $p \leq .005$ . A second method of analysis using proximities therefore confirmed the conclusions drawn from our previous analysis method.

Discussion

There were few differences in results when rankings were used rather than ratings. Once again the results supported the conclusion of no connection between the similarity in how individuals judge exemplar typicality and the similarity in how they judge feature importance. Both tasks showed a strong difference between within-subject consistency and between-subjects consensus at the level of simple correlations. Looking at the similarity matrices for each task, there were stable differences over time in how similar participants were to each other on both tasks considered separately. But the similarity structure for one task did not significantly map onto that for the other. Unlike the earlier studies there were some small signs of a positive effect, particularly for three of the six categories, but it was not of sufficient strength to justify rejecting

T5

Table 5  
Correlation in Study 3 Between Similarity Matrices for the First and Second Testing Session for Each Task (Typicality–Typicality and Importance–Importance) and Correlation Between the Similarity Matrices for Typicality and Importance Based on Ratings Aggregated Over Sessions

	Typicality–typicality		Importance–importance		Typicality–importance	
	Within-category	Between-category	Within-category	Between-category	Within-category	Between-category
Insects	.30*	.11	.25*	.11	.28*	.08
Sports	.43*	-.02	.36*	.03	.03	.08
Fish	.47*	.12	.48*	.01	-.04	.02
Tools	.51*	.13	.32*	.07	.30*	.06
Science	.42*	.12	.25*	.04	.07	.01
Vegetables	.51*	.14	.34*	.02	.27*	.05
<i>M (SD)</i>	.44 (.08)*	.10 (.11)	.33 (.09)*	.05 (.10)	.15 (.15)	.05 (.11)

\*  $p < .05$ .

AQ: 12

the null hypothesis for which the earlier studies had provided ample evidence. (The likelihood under the null hypothesis of at least one of the four analyses reaching the observed  $p$  level of .07 is  $[1 - .93^4] = .25$ .)

Note also that (a) of the three categories for which an effect was observed only one (insects) showed any similar effect in Study 1, (b) the three categories of insects, tools and vegetables do not represent any theoretically interesting subset in terms of their ontological status, and (c) there was no correspondence in the effect sizes seen for the six categories across studies (mean  $r = .07$ ; e.g., insects was positive in Study 3 but showed a negative effect in both Studies 2A and 2B).

### Analysis of Exemplar by Feature Matrices

The interpretation of the results from Studies 1 through 3 depends crucially on the assumption that the features we selected for each category were those involved in determining typicality differences among exemplars in the category as a whole. To test this assumption, we used the exemplar by feature applicability matrices generated by Verheyen and Storms (2013) and De Deyne et al. (2008). These matrices provide normative judgments of the degree to which each feature of a concept is true of each exemplar—for example, within the category of bird, whether “flies” is true of “robin.” In their study, five participants completed matrices containing 24 exemplars sampled from each category and its near neighbors and between 29 and 39 features. Each person judged yes or no whether each feature applied to each exemplar, and the applicability score for each exemplar/feature combination was the number (out of 5) who responded yes. By summing these scores across a set of features, a feature score can be calculated for each exemplar representing its featural similarity to the category prototype. We correlated these feature scores with exemplar typicality in a number of different ways (see Table 6).

The first two columns of Table 6 compare the 12 features used in Studies 1–3 (second column) with the full set of features to be found in the norms (first column). Feature scores for these two sets were compared in their ability to predict typicality in the category, as evidenced by their correlation with the rated typicality of the 24

exemplars listed in the norms for each category. In every case, the level of correlation was comparable, with the mean levels being .82 for all features and .80 for our selection of 12 features. The selection of 12 features was therefore representative of the full set of features in this important regard. The third and fourth columns in the Table show the effect of reducing the set of exemplars considered to just the 12 exemplars in Study 3 (which included top and bottom anchors) and the 12 used in Studies 1 and 2 (where anchors were removed). As would be expected from the restricted variance and smaller sample size, the correlations are somewhat lower overall and become much more variable, ranging from around .15 to .90 across categories. To understand the lower correlations, consider science where for Study 1 all the exemplars fell within the narrow typicality range of 4.6 (archaeology) to 6.6 (astronomy) on a 7-point scale. Hence, the correlation of the 12 features with the list of 12 exemplars was only .19, whereas for the 24 exemplars, with twice the standard deviation for typicality, the correlation was .80.

To conclude, the sample of features used in the three studies has been shown to be representative of the general set of features that people use to describe the concepts. In some cases, the 12 features in our sample did not predict typicality differences in the 12 exemplars in our sample. However, they all predicted typicality across a wider sample, and that is the key evidence that is needed to show that they were representative of the features involved in the concept prototype.

### Study 4

As a final test of our hypothesis, we report an unpublished study conducted by the first author some years ago (around 1985) using a different set of categories and materials, and with some other important differences. It is reported here, because although not conducted as part of the current project, the differences in method and content lend a useful test of generality to the results reported so far. The study was conducted on a different sample of students and in English rather than Italian. It also used 10 categories, only four of which were used in Studies 1–3. Most importantly the

Table 6  
*Comparison of the Correlation of Feature Score With Typicality for the Sample of 12 Features Used in Studies 1–3 and for the Full Set of Norms*

	Correlation of feature score with typicality			
	All features × 24 exemplars from norms	12 Features × 24 exemplars from norms	12 Features × 12 exemplars from Study 3	12 Features × 12 exemplars from Studies 1 and 2
Insects	.84	.83	.70	.37
Sports	.88	.80	.92	.72
Fish	.80	.77	.77	.90
Tools	.84	.94	.89	.80
Science	.84	.80	.65	.19
Vegetables	.72	.63	.15	.14
<i>M</i>	.82	.80	.68	.52

*Note.* Four analyses are shown, correlating respectively: (a) feature scores based on all the 29 to 39 features in the norms with the typicality of all 24 exemplars in the norms, (b) feature scores based on the selected 12 features with the typicality of all 24 exemplars in the norms, (c) feature scores based on the 12 selected features with reverse ranked typicality of the 12 exemplars used in Study 3, and (d) feature scores based on the 12 selected features with rated typicality of the 12 exemplars in Studies 1 and 2.

instructions for both tasks emphasized not only typicality judgments but also the degree of category membership shown.

**Method**

**Participants.** Twenty-five students (both males and females) at City University London participated on a voluntary basis.

**Materials.** Ten categories were used (see Appendix B), including four of those used in Studies 1–3. For each category, a list of 10 features and a list of 20 exemplars were used. The list of exemplars included a number of borderline cases and some non-members.

**Procedure.** The first section of the experimental booklet contained instructions for ranking the importance of the features of each category, including the following:

Your task will be to judge which is the most important feature in deciding whether any object is a representative member of the category or not. For example, all *birds* have feathers, and no other creatures do, so this you might consider the most important. Alternatively, you might consider *flying* most important as it distinguishes typical birds from the less typical ones.

The numbers 1–10 were used to rank importance of the list of 10 features for each of the 10 categories. In Section 2 of the booklet, instructions asked participants

to decide the extent to which each word does or does not belong to the category. First, decide whether the word belongs to the category, then if it does, decide how typical a member it is, and if it does not, decide how closely related to the category it is.

These instructions follow Hampton (1979) as a means of combining category membership and typicality into a single scale of graded membership.

Judgments of typicality were recorded using an analogue scale. Opposite each exemplar was printed a solid horizontal line 8 cm long. The left end was labeled *unrelated*, and the right end *highly typical*, while the midpoint was marked and labeled *boundary*. The instructions continued,

The three marks already on the line represent (from the left): the *completely unrelated* end of the scale; the category boundary at which it is impossible to decide if the answer is yes or no; and the *completely prototypical category member* end of the scale. Please read the whole list first to get a feel for the range of items and then work down the list carefully making a mark against each word. If you do not know the meaning of any of the words, please indicate this.

The data were then scored using a ruler to create a rank order for the typicality/membership of the exemplars for each participant. For both feature importance and typicality, an example of *birds* was used to illustrate the method.

**Results**

Six data points were missing because participants did not know the item and were replaced with the average rank. Reliability of the scales was high for all 20 scales, (mean .94 for importance, .96 for typicality), and no participants needed to be excluded for consistently low correlations with the group. The distribution of correlations was similar to the previous studies (see Figure 1) with a

negative skew. There was no evidence of pairs of participants with consistently high correlations across categories.

Similarities between participants on each of the 20 scales were computed as before using correlation matrices. The average correlation within a scale (indicating the degree of consensus in the group) was .45 for importance and .53 for typicality, with all scales showing mean values significantly above zero,  $t(299) > 10.0, p < .001$  (see Table 7).

T7

These similarity matrices were then correlated to test the prediction of a correspondence between interperson similarities on each of the two tasks. The correlation between the similarity matrices for importance and typicality within-categories is shown in the right column of Table 7. The mean was very close to zero ( $M = -0.007$ ). (There were two correlations that differed significantly from zero, but one was positive and the other negative.) Between different categories, the correlation of importance and typicality (not shown in the Table) was 0.023.

**Discussion**

In terms of materials and procedure, Study 4 differed from the previous studies in numerous ways, but the results were directly comparable, both in the level of consensus on each task and in the lack of correspondence between similarity matrices. The language used was English rather than Italian and the participants were from a large metropolitan city rather than a small provincial one. Six additional categories were used, and with a different selection of features and exemplars for the four that were repeated. Exemplar lists included borderline and nonmembers of the categories. The procedure used just 10 features to be ranked, and 20 exemplars whose ranking was derived from an analogue scale. Most importantly the instructions for each task made explicit reference to the other. Thus, when ranking importance, people were told to think about what would make an exemplar both a clear member and a highly typical category member, and when judging typicality, they considered both category membership and typicality at the same time. There was therefore good reason to suppose that there would be a strong link between the two tasks. The results, however, very clearly supported the previous findings of a near zero correlation between interindividual similarities on the two tasks. The fact that many differences in materials and procedure led to the same result

Table 7  
Results of Study 4 Showing Significant Consensus on Each Task and Zero Correlation Between Similarities Across Tasks

Category	Consensus		Correlation of similarities for importance and typicality
	Importance	Typicality	
Sports	.49	.43	.04
Fish	.68	.47	-.18
Tools	.46	.55	.04
Vegetables	.57	.58	.00
Furniture	.33	.50	.16
Vehicles	.38	.58	.01
Weapons	.57	.44	-.09
Clothing	.40	.62	-.09
Kitchen utensils	.39	.42	-.05
Fruit	.23	.73	.08
<i>M (SD)</i>	.45 (.13)	.53 (.10)	-.01 (.10)

lends weight to the conclusion that there is a genuine lack of relation between the two tasks that requires a theoretical explanation.

### General Discussion

Across four studies with five samples of participants a common conclusion was reached. Judgments of exemplar typicality and of feature importance were made reliably by most participants. Group consensus and test-retest consistency were both substantial indicating that the tasks were tapping a common understanding of concepts. Both tasks also showed a greater degree of correlation within individuals than between individuals, supporting the idea that there are consistent individual differences in representation. This study is the first to confirm that this pattern, previously reported for typicality and other measures of conceptual structure is also to be found for judgments of feature importance.

The present research took the analysis a step further by asking whether the similarity between people in the sample could be shown to form a stable structure. Studies 1 and 3 confirmed that the matrix of pairwise similarities between individuals was stable over time for both the typicality and the importance tasks. Yet in none of the five samples was there any significant evidence that the similarity structure for typicality judgments could be mapped onto that for importance. Studies 1, 2A, 2B, and 4 found the relationship to be close to zero, once the level of general positive correlation seen between the matrices for different categories was taken into account. Study 3 obtained a small trend in the right direction, but the correlation was not significantly greater than the background level of correlation, in spite of the stable similarity differences shown for each task. Studies 1, 3, and 4 even had the two tasks performed on the same occasion one after the other. This procedure might be expected to prime a relation between the two tasks. Having rated hiking as a typical sport and darts as atypical, one might expect the participant to go on to judge physical fitness as a more important feature of sports than specific skill. (Recall that Verheyen & Storms, 2013, identified reliable group differences in their sample in how darts vs. hiking were categorized.) But there was still no connection seen between the two tasks in terms of participant similarities.

Because we were comparing similarities across tasks, care was needed to ensure that artifactual effects were excluded. If a participant found both tasks hard to follow, then their similarity to others would be low on both tasks, and hence contribute to a positive correlation between similarity matrices. To this end, we used the fact of a participant having a low set of correlations with the group mean across different categories as an exclusion criterion. At the other end of the scale, we excluded five participants in Study 1 and one in Study 2A who had an unusually high number of very high correlations with other participants. A pair of participants who were artificially similar on both tasks would clearly contribute to an apparent match across the tasks. In the end, this exclusion had no effect on the conclusions that were drawn, because we were able to introduce a control that took account of individual differences on factors *other than* differences in representation. By comparing the correlation of the similarity matrices between tasks for the *same* category with that for *different* categories, any systematic individual similarities introduced by differences in motivation, or differences in interpretation of instructions

could be controlled for. Similarities based on motivation or interpretation of the task would show up equally in the correlations between different categories as within the same category.

One question that arises is whether the similarity structure seen in each task separately was sufficiently strong for the correlation between them to emerge. When the data from the two testing occasions were combined, the Spearman-Brown reliability of the two similarity matrices averaged .43 and .49 for typicality and importance in Study 1 and .61 and .49, respectively, for Study 3. These are relatively high values for reliability, and given that there were six to 10 categories and five samples, there was ample opportunity for even a small effect to appear. Interestingly, if one correlates the size of the correlation between typicality and importance similarity matrices with the joint reliability of the two measures, there was a negative correlation across categories of  $-.29$  for Study 1 and  $-.39$  for Study 3. Thus when the two similarity measures were more reliable, their intercorrelation was actually lower.

### Meta Power Analysis

In making a strong claim for the absence of an effect, it is useful to provide some statistical support from a consideration of power. The results of the four critical  $t$  tests from Studies 1 through 3, comparing the correlations of similarity matrices between the two tasks for within-category versus between-category comparisons were used in a meta-analysis to estimate the power of detecting a difference in the predicted direction. We took the level of between-category correlation as given in each case, and estimated the power to detect an increase in the within-category correlation above that level. To find a difference in mean correlations of as great as 0.10 (as seen in Study 3), each study had a power independently to detect such a difference of between .55 and .76. The estimated likelihood of a Type II error in each sample was therefore between .45 and .24, and multiplying these together, the likelihood of obtaining a Type II error in each of the four samples was estimated to be 0.008. In other words, the power to detect a significant result of this magnitude, with  $\alpha = .05$ , in at least one of our four samples was greater than 99%. In fact the set of studies had a combined power of 90% to detect an increase in the correlation for within-category comparisons of as little as 0.07 in at least one of the samples. Our observed effect size based on a weighted mean across studies was actually 0.007 with  $SE = .046$ .

If, to these calculations, the null result of Study 4 is added, where the power to detect a mean correlation of .10 across the 10 categories was estimated at 90%, the likelihood of there being a theoretically interesting level of correlation between the two similarities, using our methods, is vanishingly small.

### Sources of Individual Variation

A positive result from these studies was the evidence for stable levels of individual difference in concept representation as seen both in extensional and intensional measures. Differences between individuals in performing these tasks may be related to various aspects of their knowledge. In fact, Connell and Lynott (2014) argue that every time a concept is instantiated in memory it will have a different representation. First, it is possible that people represent the conceptual categories differently. For example, Ver-

heyen and Storms (2013) have shown that sports may be more about physical effort for some people, and more skill-based for others. Second, knowledge and views of the individual exemplars being judged for typicality may also vary. Some people may be very familiar with a particular sport and so rate it highly, while others may know little about it. Third, there may be differences in how people understand the features being rated for importance, or in their beliefs about which exemplars possess them. Any of these three sources of variation would give rise to individual differences in concept representation. The conundrum is to understand why there is a disconnect between the individual differences and similarities that are seen in extensional and intensional behavioral measures.

It is important to understand that the evidence presented here is indirect, being based on the (lack of) correspondence between similarity structures among individuals, rather than a direct measure of each individual's links between exemplars and features. Judgments of typicality and importance each no doubt have multiple influences, some of which will be specific to one or other task and may lead to a reduction in the correlation between similarity structures. However a central assumption of concept theories—that intensions determine extensions—predicts that there will be at least some, possibly low, but significant level of correlation, and this is what has failed to appear. Suppose two individuals are both keen runners, but have no interest in archery. The assumption is that their ideas about the features which are important for counting something as a typical sport will be similar as a consequence. Personal learning histories and individual interests should have an influence in both tasks. Intuitions of typicality should be based on what features are considered important and whether the exemplar possesses them, while intuitions of which features are important should be based on the exemplars that a person brings to mind as being typical.

It is also likely that some of the similarity between individuals' typicality judgments in our data was driven by features that happened to be excluded from our sample. However, given random sampling of features across multiple categories, and given that the features chosen proved to be just as predictive of typicality as those not selected, it is highly improbable that the results could just be explained by inadequate sampling. The replication of the null result with a new sample of categories and features in Study 4 supports this claim.

### Implications for Theories of Concepts

These results present a serious challenge to a number of theories of concept representation. The most seriously affected theory is the prototype theory, according to which an exemplar's typicality depends on its possession of important features, and a feature's importance depends on its association with typical exemplars (Hampton, 2006). However, even theories of concepts that deny the role of prototypes as concepts still assume that most common concepts *have* prototypes (Connolly, Fodor, Gleitman, & Gleitman, 2007). According to most accounts of prototypes and typicality effects, judgments of typicality and judgments of which features are important for categorization should rely on a single semantic representation. It is therefore difficult to explain the lack of relation between the stable individual differences in intensional and extensional task performance.

As discussed in the introduction, there is plenty of evidence to point to the close connection between exemplar typicality and shared features. Barsalou (1985) found a partial correlation between ratings of central tendency (i.e., similarity to category members) and goodness of example (or typicality) of .71, substantially above other predictors. Consequently almost all accounts of typicality differences within categories make reference to similarity among exemplars (see, e.g., Zee et al., 2014), and most accounts of similarity among rich concepts such as these make reference to shared and distinctive features (Tversky, 1977). Had our results come out with the expected match of similarities, it would have clearly been strong support for this general account of typicality effects. The failure to do so must therefore be of considerable interest to supporters and critics of prototype theory alike.

One explanation of our results that can be ruled out is to argue that people have no metacognitive access to information about features and their role in concept representations. For example, some exemplar models (e.g., Hintzman, 1986) would suggest that features are equally weighted, and so there would be no basis for making such a judgment. In a similar vein, studies in social psychology have often found that people's justification and explanation of their actions are often dissociated from their actual reasons for acting (Nisbett & Wilson, 1977; Johansson, Hall, Sikström, & Olsson, 2005; Johansson, Hall, Sikström, Tärning, & Lind 2006). Against this view, we found that (a) the degree of reliability (alpha) for feature importance judgments in Study 1 was just as high as for typicality judgments, (b) the consistency and consensus levels were comparable, and (c) the stability of the similarity matrices were also equivalent. (In Study 3, values for typicality were relatively enhanced because of the inclusion of the anchor items.) It cannot therefore be argued that the importance task was not tapping a valid and reliable source of semantic information. If there was simply a lack of metacognitive access to intensional information, then one cannot explain why intensional judgments should be reliable, and why the similarity structure for intensions should be stable across occasions. Yet it appears that at the level of individual differences the information involved in intensional judgments is disconnected from that involved in typicality judgments. People with more similar views about which features are more important are not more similar in their views about which exemplars are more typical.

The lack of a relationship between the similarity structures for the extensional and intensional tasks strongly suggests that the aspects of semantic memory tapped by the two tasks are independent of each other. This conclusion lends support to so-called hybrid theories of concepts in which there are multiple representations of the same concept within the mind. In philosophy, a number of authors have proposed that exemplars, prototypes and theories may exist independently in semantic memory. Machery (2009) is the foremost proponent of this view. He argues that "concept" is a theoretically empty term, and that cognitive science will need to replace it with theories that individuate the exemplar, prototype and theory aspects of conceptual representations. Further discussions of pluralism for concepts can be found in Dove (2009), Rice (2015), and Weiskopf (2009).

If we accept Machery's (2009) proposal, then extensional judgments about typicality and graded membership could be based on a memory store of exemplars, structured by similarity. As described in the introduction, Storms (2004) describes evidence that

superordinate concepts like *tool* may be represented as a collection of basic level concepts, such as hammer, saw, and chisel, so that typicality in the superordinate category reflects similarity to the closest basic level concept, rather than similarity to the prototype features of the superordinate itself. Individual variation in extensional judgments would then reflect which exemplars are prominently stored in a person's semantic memory, and which dimensions of similarity are used to structure the store. On the other hand, intensional judgments about the importance of a feature would rely on the "theory" aspect of concept representation, reflecting a higher more abstract level of thinking. For example, Sloman, Love, and Ahn (1998) measured the importance of a feature for a category by asking people to estimate the likelihood that an item which lacked that feature could belong in the category. Importance may therefore depend on understanding of how features are interrelated with causal and functional connections at the category level (e.g., that a bird without wings would not fly), and so similarity in task performance would reflect the degree of shared beliefs about the network of causal relations among the concepts features.

In sum, a possible explanation of our result combines an exemplar-based account of extensions with a causal schema based account of intensions. Such a hybrid model lacks the coherence and simplicity of prototype theory, but the evidence presented here is inconsistent with the latter model. It can be noted that there is also evidence for two systems of category learning in the related field of perceptual categorization (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; for a review, see Wills, 2013). Ashby et al.'s (1998) competition between verbal and implicit systems (i.e., AQ: 13 COVIS) model proposed that there is an implicit exemplar-based categorization learning system and an explicit rule-based learning system. Because rules are clearly intensional, referring to features of the stimuli, there may be a parallel with the dissociation that we have discovered with concepts in long-term semantic memory. Another parallel is with the "two systems" theory of reasoning (Evans, 2010; Kahneman, 2012; Sloman, 1996), where one could see extensional decisions as using fast heuristic methods and intensional decisions as requiring more deliberate processing (see also Hampton, 1984). Such parallels remain speculative.

## Conclusions

This article presents the first detailed exploration of the consistency and degree of consensus in people's judgments about the importance of different features in the makeup of a concept prototype. We found that these importance judgments can be just as reliable as the more commonly studied typicality judgments. Just as with typicality judgments and category membership judgments, importance judgments showed individual consistency over time that was greater than the level of between-individual consensus. In addition, the similarity matrices showing the similarity of one individual's judgments to another also showed stability over time for both extensional and intensional judgments. Both typicality and importance judgments reveal stable individual variation in concept representation. However the similarity structure for one task (typicality) did not map onto that for the other (importance), suggesting that concepts' intensions and extensions are not represented in a single integrated form, but may result from a pluralistic or hybrid form of concept representation.

## References

- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481. <http://dx.doi.org/10.1037/0033-295X.105.3.442>
- 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. <http://dx.doi.org/10.1037/0278-7393.11.1-4.629>
- Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge, United Kingdom: Cambridge University Press.
- Bellezza, F. (1984a). Reliability of retrieval from semantic memory: Common categories. *Bulletin of the Psychonomic Society*, *22*, 324–326. <http://dx.doi.org/10.3758/BF03333832>
- Bellezza, F. (1984b). Reliability of retrieval from semantic memory: Noun meanings. *Bulletin of the Psychonomic Society*, *22*, 377–380. <http://dx.doi.org/10.3758/BF03333850>
- Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, *6*, 390–406. <http://dx.doi.org/10.1111/tops.12097>
- Connolly, A. C., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition*, *103*, 1–22. <http://dx.doi.org/10.1016/j.cognition.2006.02.005>
- De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030–1048. <http://dx.doi.org/10.3758/BRM.40.4.1030>
- Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*, 412–431. <http://dx.doi.org/10.1016/j.cognition.2008.11.016>
- Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford, United Kingdom: Oxford University Press.
- Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford, United Kingdom: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195154061.001.0001>
- Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, *18*, 441–461. [http://dx.doi.org/10.1016/S0022-5371\(79\)90246-9](http://dx.doi.org/10.1016/S0022-5371(79)90246-9)
- Hampton, J. A. (1984). The verification of category and property statements. *Memory & Cognition*, *12*, 345–354. <http://dx.doi.org/10.3758/BF03198294>
- Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, *15*, 55–71. <http://dx.doi.org/10.3758/BF03197712>
- Hampton, J. A. (1997). Associative and similarity-based processes in categorization decisions. *Memory & Cognition*, *25*, 625–640. <http://dx.doi.org/10.3758/BF03211304>
- Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, *65*, 137–165.
- Hampton, J. A. (2006). Concepts as prototypes. *Psychology of Learning and Motivation*, *46*, 79–113. [http://dx.doi.org/10.1016/S0079-7421\(06\)46003-5](http://dx.doi.org/10.1016/S0079-7421(06)46003-5)
- Hampton, J. A., Aina, B., Andersson, J. M., Mirza, H. Z., & Parmar, S. (2012). The Rumsfeld effect: The unknown unknown. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 340–355. <http://dx.doi.org/10.1037/a0025376>
- Hampton, J. A., Dubois, D., & Yeh, W. (2006). The effects of pragmatic context on classification in natural categories. *Memory & Cognition*, *34*, 1431–1443. <http://dx.doi.org/10.3758/BF03195908>
- Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of*

- Psychology*, 74, 491–516. <http://dx.doi.org/10.1111/j.2044-8295.1983.tb01882.x>
- Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411–428. <http://dx.doi.org/10.1037/0033-295X.93.4.411>
- Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310, 116–119. <http://dx.doi.org/10.1126/science.1111709>
- Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition: An International Journal*, 15, 673–692. <http://dx.doi.org/10.1016/j.concog.2006.09.004>
- Kahneman, D. (2012). *Thinking fast and slow*. London, United Kingdom: Penguin.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44. <http://dx.doi.org/10.1037/0033-295X.99.1.22>
- Machery, E. (2009). *Doing without concepts*. New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195306880.001.0001>
- McCloskey, M., & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition*, 6, 462–472. <http://dx.doi.org/10.3758/BF03197480>
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, United Kingdom: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511529863.009>
- Murphy, G. L. (1993). Theories and concept formation. In I. van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173–200). London, United Kingdom: Academic Press.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316. <http://dx.doi.org/10.1037/0033-295X.92.3.289>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259. <http://dx.doi.org/10.1037/0033-295X.84.3.231>
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104–114. <http://dx.doi.org/10.1037/0278-7393.10.1.104>
- Rice, C. (2014). Concepts as pluralistic hybrids. *Philosophy and Phenomenological Research*. Advance online publication. <http://dx.doi.org/10.1111/phpr.12128>
- Rips, L. J. (1989). Similarity, typicality and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, United Kingdom: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511529863.004>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Rosch, E. R., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605. [http://dx.doi.org/10.1016/0010-0285\(75\)90024-9](http://dx.doi.org/10.1016/0010-0285(75)90024-9)
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22. <http://dx.doi.org/10.1037/0033-2909.119.1.3>
- Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189–228. [http://dx.doi.org/10.1207/s15516709cog2202\\_2](http://dx.doi.org/10.1207/s15516709cog2202_2)
- Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81, 214–241. <http://dx.doi.org/10.1037/h0036351>
- Storms, G. (2004). Exemplar models in the study of natural language concepts. *Psychology of Learning and Motivation-Advances in Research and Theory*, 45, 1–39. [http://dx.doi.org/10.1016/S0079-7421\(03\)45001-9](http://dx.doi.org/10.1016/S0079-7421(03)45001-9)
- Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar based information in natural language categories. *Journal of Memory and Language*, 42, 51–73. <http://dx.doi.org/10.1006/jmla.1999.2669>
- Storms, G., De Boeck, P., & Ruts, W. (2001). Categorization of novel stimuli in well-known natural concepts: A case study. *Psychonomic Bulletin & Review*, 8, 377–384. <http://dx.doi.org/10.3758/BF03196176>
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352. <http://dx.doi.org/10.1037/0033-295X.84.4.327>
- Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica*, 135, 216–225. <http://dx.doi.org/10.1016/j.actpsy.2010.07.002>
- Verheyen, S., & Storms, G. (2011). Does a single dimension govern categorization in natural language categories? In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), *European perspectives on cognitive science* (Paper #226). Sofia, Bulgaria: New Bulgarian University Press.
- Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PLoS ONE*, 8, e63507. <http://dx.doi.org/10.1371/journal.pone.0063507>
- Weiskopf, D. (2009). The plurality of concepts. *Synthese*, 169, 145–173. <http://dx.doi.org/10.1007/s11229-008-9340-8>
- Wills, A. J. (2013). Models of categorization. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 346–357). Oxford, United Kingdom: Oxford University Press.
- Zee, J., Storms, G., & Verheyen, S. (2014). Violations of the local independence assumption in categorization. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1670–1675). Austin, TX: Cognitive Science Society.

(Appendices follow)

CATEGORIZATION DIFFERENCES

Appendix A

Table A1  
Feature Lists for the Importance Judgments for Studies 1–3

AQ: 18	Insect	Sport	Fish
	Is more often encountered during summer	Can be practiced in competition	Has fins
	Flies or crawls	Has some rules of play	Swims in aquariums
	Is eaten by other animals	Supporters encourage the players	Doesn't live on land
	Doesn't live long	Requires concentration/mental effort	Can swim
	Often bites	Is practiced in a club	Has eyes
	Lays eggs	Requires specific skills	Breathes through gills
	Carries diseases	Is good for one's fitness	Can have different colors
	Has six legs	Is healthy	Is tasty
	Is irritating	Requires special clothing	Is healthy
	Lives in colonies	Is practiced outdoors	Is slippery
	Is vermin	Produces sweat	Is sometimes eaten by man
	There are lots of these	Is tiring	Lays eggs
	Tool	Science	Vegetable
	Is primarily used by men (not women)	Has brought modern comfort	Grows in the ground
	Was used in early ages	Allows one to make predictions	Is nutritious
	Made of wood	Is interesting	Comes in different colors
	Used in construction	People expect a lot from it	Exists in different kinds
	Is meant for people who are handy	Is responsible for longer life expectancy	Is boiled
	Is big	Explains all kinds of phenomena	Can be eaten raw
	Makes work easier	Is based on empirical data	Can have different tastes
	Can be dangerous	Is pursued at universities	Is eaten warm
	Is held in the hand	Provides insight into common things	Contains vitamins
	Is used to work with	Is held in high regard	Is not sweet
	Is an aid	One should be critical of it	Grows in the garden
	Comes in very handy/useful	Is boring	Green color

Table A2  
Exemplar Lists Used for Typicality Judgments for Studies 1–3

Insects	Sports	Fish	Tools	Science	Vegetables
Amoeba <sup>c</sup>	Aerobics <sup>c</sup>	Alligator <sup>c</sup>	Axe	Advertising <sup>b</sup>	Apple
Bacterium	Ballroom Dancing	Catfish	Calculator <sup>c</sup>	Archaeology	Artichoke
Caterpillar	Billiards	Clam	Dictionary	Architecture	Bread <sup>b</sup>
Centipede	Bullfighting	Eel	Hammer <sup>a</sup>	Astrology	Cereal
Dust Mite	Chess	Jellyfish	Key	Astronomy	Chili Pepper
Earthworm	Conversation <sup>b</sup>	Oyster <sup>c</sup>	Pen	Chemistry	Garlic
Hamster <sup>b</sup>	Croquet <sup>c</sup>	Plankton	Photograph <sup>b</sup>	Dentistry <sup>c</sup>	Parsley
Head Lice	Darts	Sardine	Scalpel <sup>c</sup>	Economics	Pineapple <sup>c</sup>
Leech	Hiking	Seagull <sup>b</sup>	Scissors	Geometry	Potato
Mosquito <sup>a</sup>	Jogging	Shark	Screw	Mathematics	Sage
Scorpion	Kite Flying	Shrimp	Sewing Needle	Medicine <sup>a</sup>	Seaweed
Silkworm	Surfing	Tadpole	Stone	Pharmacy	Spinach <sup>a</sup>
Snail	Swimming <sup>a</sup>	Trout <sup>a</sup>	Umbrella	Psychology <sup>c</sup>	Turnip
Termite <sup>c</sup>	Wrestling	Whale	Varnish	Sociology	Watercress <sup>c</sup>

<sup>a</sup> Item treated as high typicality anchor filler in Studies 1 and 2. <sup>b</sup> Item treated as low typicality anchor filler in Studies 1 and 2. <sup>c</sup> Items omitted in Study 3 to reduce the list length to 12.

(Appndices continue)

## Appendix B

Table B1  
*Feature Lists Used for Study 4*

Sports	Fish	Tools	Vegetables	Furniture
Involve skill	Have tails	Have handles	Are always cooked	Is made of wood
Involve spectators	Are catchable	Are used in industry	Are grown commercially	Is moveable
Have rules	Have fins	Are hard	Grow close to the ground	Is found in the home
Are competitive	Swim	Are for constructing things	Are green	Is for keeping things in
Have teams	Have gills	Are hand-held, manipulable	Are eaten with meat	Is functional
Involve physical exertion	Cannot survive on land	Are sharp	Are the roots of plants	Has legs
Are done for exercise	Are slimy	Have a specialized function	Have leaves and stalks	Is for sitting on
Are enjoyable to do	Have eyes	Are made of metal	Are not sweet	Is decorative
Involve special equipment	Have scales	Are strong	Are tasty	Is expensive
Take place out of doors	Are edible	Are used to repair things	Are crunchy when raw	Is comfortable
Vehicles	Weapons	Clothing	Kitchen utensils	Fruit
Made of metal	Are used to injure people	Is for protection against the weather	Made of metal in part	Contains seeds
Have wheels	Are made of metal	Is soft	Held in the hand	Is juicy
Carry people	Have a sharp edge	Is for preserving one's modesty	Made of plastic in part	Is soft
Are steered	Are used in wars	Gives an impression of one's character	Have handles	Is eaten raw
Have seats	Are explosive	Is worn on the body	Used for cooking with	Is healthy for you to eat
Carry things	Are used to destroy things	Is made of woven material	Made of wood in part	Has a peel
Have an engine	Are used to hurt people	Is pliable	Have a sharp edge	Grows on trees
Are expensive	Have restricted availability	Provides fashion	Are containers	Is colorful
Move faster than a man on his own	Shoot some projectile	Is attractive	Are heat resistant	Is sweet
Have a driver	Are dangerous	Is long-lasting	Are only found in kitchens	Is round

(Appendices continue)

## CATEGORIZATION DIFFERENCES

Table B2  
*Exemplar Lists Used for Study 4*

Sports	Fish	Tools	Vegetables	Furniture
Dueling	Whale	Excavator	Carrot	Transistor radio
Swimming	Cod	Saw	Mushroom	Card table
Croquet	Octopus	Plumb line	Rice	Wastepaper basket
Deep-sea diving	Otter	Clothes iron	Turnip	Ashtray
Football	Squid	Cooker	Olive	Chair
Skipping	Tadpole	Spirit level	Gourd	Fridge
Writing	Jellyfish	Typewriter	Avocado	Mirror
Billiards	Clam	Shovel	Garlic	Telephone
Sack race	Dolphin	Potato peeler	Rhubarb	Curtains
Roller skating	Eel	Potter's wheel	Tomato	Cushion
Bar-football	Oyster	Ruler	Celery	Picture
Fishing	Seahorse	Paint brush	Dandelion	Television
Roulette	Starfish	Pencil	Seaweed	Clock
Fox-hunting	Stingray	Mug	Cucumber	Cupboard
Chess	Lobster	Pocket calculator	Asparagus	Wall-shelf
Horse riding	Shark	Cigarette lighter	Potato	Bed
Darts	Shrimp	Blow torch	Sweetcorn	Mantelpiece
Archery	Crab	Suitcase	Radish	Hi-fi set
Surfing	Plaice	Scalpel	Beetroot	Carpet
Pot-holing	Walrus	Clothes brush	Ginger	Car seat
Vehicles	Weapons	Clothing	Kitchen utensils	Fruit
Surfboard	Gas	Wallet	Strainer	Orange
Bus	Knife	Shirt	Fork	Rhubarb
Pram	Stick	Wristwatch	Dishwasher	Sweet potato
Parachute	Words	Briefcase	Meat thermometer	Watermelon
Pavement	Gun	Dress	Knife	Carrot
Conveyor belt	Foot	Band-Aid	Radio	Pumpkin
Lawnmower	Scissors	Fingernail varnish	Sponge	Beetroot
Canoe	Poison	Umbrella	Sink	Walnut
Taxi	Rope	Shoes	Clock	Olive
Roller skates	Brick	Hearing aid	Table	Pomegranate
Feet	Car	Necklace	Scissors	Coconut
Wheelchair	Fist	Makeup	Electric blender	Apple
Bicycle	Glass	Bracelet	Corkscrew	Raisin
Wheelbarrow	Ice pick	Handkerchief	Gas cooker	Mango
Horse	Axe	Cufflinks	Plate	Fig
Train	Cannon	Buttons	Broom	Tomato
Cable car	Razor blade	Dentures	Fridge	Cucumber
Lift	Stone	Handbag	Toaster	Gherkin
Merry-go-round	Laser beam	Top hat	Mop	Lemon
Submarine	Nuclear bomb	Gloves	Glass	Peanut

Received August 18, 2014  
Revision received August 14, 2015  
Accepted August 17, 2015 ■