Mem Cogn
DOI 10.3758/s13421-012-0252-y

Idealness and similarity in goal-derived categories: A computational examination

Wouter Voorspoels · Gert Storms · Wolf Vanpaemel

© Psychonomic Society, Inc. 2012

Abstract The finding that the typicality gradient in goal-derived categories is mainly driven by ideals rather than by exemplar similarity has stood uncontested for nearly three decades. Due to the rather rigid earlier implementations of similarity, a key question has remained—that is, whether a more flexible approach to similarity would alter the conclusions. In the present study, we evaluated whether a similarity-based approach that allows for dimensional weighting could account for findings in goal-derived categories. To this end, we compared a computational model of exemplar similarity (the generalized context model; Nosofsky, Journal of Experimental Psychology. General 115:39–57, 1986) and a computational model of ideal representation (the ideal-dimension model; Voorspoels, Vanpaemel, & Storms, Psychonomic Bulletin & Review 18:1006-114, 2011) in their accounts of exemplar typicality in ten goal-derived categories. In terms of both goodness-of-fit and generalizability, we found strong evidence for an ideal approach in nearly all categories. We conclude that focusing on a limited set of features is necessary but not sufficient to account for the observed typicality gradient. A second aspect of ideal representations—that is, that extreme rather than common, central-tendency values drive typicality—seems to be crucial.

Keywords Concepts · Category representation · Computational models of representation · Typicality · Goal-derived categories · Semantic memory

Introduction

A central topic in cognitive science is the study of the human conceptual system, containing the knowledge that people have of the world and of their culture. This knowledge is systematized in *concepts*—that is, the basic units of knowledge associated with a category or class of objects (e.g., Barsalou, Simmons, Barbey, & Wilson, 2003; Komatsu, 1992; Lakoff & Johnson, 1980; Medin, 1989; Murphy, 2002; E. E. Smith & Medin, 1981). Concepts are the mental representations that serve as the input to the wide variety of cognitive activities that people engage in (Barsalou et al., 2003), such as categorization and induction. Since concepts are often considered the building blocks of thought, understanding human cognition requires a solid theory on the nature of concepts (e.g., Pinker, 1997; Solomon, Medin, & Lynch, 1999).

Arguably the most dominant concept in concept research is *similarity* (e.g., Goldstone, 1994; Goldstone & Son, 2005; Komatsu, 1992; Markman & Gentner, 1993; Medin, Goldstone, & Gentner, 1993; Murphy, 2002; Rosch & Mervis, 1975; Sloman & Rips, 1998; L. B. Smith, 1989; Tversky, 1977; but see also Murphy & Medin, 1985). As Quine (1969, p. 114) noted, "Similarity is fundamental for learning, knowledge and thought, for only our sense of similarity allows us to order things in kinds so that these can function as stimulus meanings." Indeed, similarity is a crucial organizing principle for a large class of concepts that are *world-derived*. This type of concept originates from salient groupings of similar objects in the environment—that is to say, categories that derive from the correlational structure of the world (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). For example, the concept *raven* systematizes our knowledge on ravens and is grounded on a very distinct category of objects, in that its members are roughly similar to each other, and dissimilar to other categories, such as *aardvarks* or *arm chairs*.

For some kinds of concepts, however, similarity does not seem the underlying principle. In particular, *goal-derived concepts*, which find their origin in the goals of a cognitive agent in a particular context, are—at least at first sight—far less dependent on correlations in the environment (Medin,

W. Voorspoels (✉) · G. Storms · W. Vanpaemel
Department of Psychology, University of Leuven,
Tiensestraat 102,
3000 Leuven, Belgium
e-mail: wouter.voorspoels@psy.kuleuven.be

Published online: 13 September 2012

Lynch, & Solomon, 2000). For example, when coming home from work, someone might construct the concept *things to rescue from a burning house*. This concept could have the following extension: "letters," "iPad," "favorite blanket," "car keys," "rich aunt Terry's huge amount of extremely flammable paper money," and "my favorite cat." These objects clearly do not reflect any obvious correlational structure or salient similarity grouping, but they do form a sensible category, certainly from the perspective of the poor man who finds his house blazing. It is doubtful that he will be tempted to include "uncle Sam's huge amount of extremely flammable Monopoly money" as a member of the category, its obvious similarity to aunt Terry's money notwithstanding. In a classic study, Barsalou (1985) confirmed this intuition: He concluded on the basis of a thorough correlational analysis that *idealness*, and not similarity, is the underlying principle in goal-derived categories. Idealness can be defined as the extent to which a certain object displays an ideal quality—that is to say, a quality that is directly related to the goal at hand.

While similarity between objects indeed seems unable to capture the structure in goal-derived categories, it is at the same time obvious that members of *things you rescue from a burning house* do have in common that they are valuable items. If we focus exclusively on valuableness, it is easy to differentiate uncle Sam's Monopoly money from aunt Terry's real money, and perhaps even to relate this money to my favorite cat. Put differently, whereas global similarity might not be the underlying principle, a more focused similarity, constituted by a limited number of relevant dimensions (e.g., emotional and financial "valuableness"), may be able to account for goal-derived categories. Indeed, it has been extensively demonstrated that such a more flexible notion of similarity, which allows attention to distribute across stimulus dimensions depending on the requirements of the context, is crucial in understanding human cognitive activity, and in particular categorization (e.g., Nosofsky, 1986; L. B. Smith, 1989).

Despite the availability of a number of excellent cognitive models of category representation that allow for such a context-dependent focus on the relevant dimensions (e.g., Nosofsky, 1986; Minda & Smith, 2011), these models have never been considered in the context of goal-derived categories. The question thus remains whether the distinct notion of *ideal* is required to explain the structure of goal-derived categories, or whether a more flexible notion of similarity can provide an adequate account. In the present study, we thoroughly examined this question by applying computational cognitive models that allow for a stern test. Before describing the empirics and model-based analyses, we will first provide crucial details on the similarity and idealness accounts, in particular the aspects that differentiate the two approaches.

Typicality, similarity, and idealness

An often-used variable for tapping into the internal structure of concepts is *typicality*, referring to the extent to which members of the extension of a concept are considered good examples of the category. Typicality has been shown to be influential in a variety of category-dependent tasks (e.g., Hampton, 1993; Keller & Kellas, 1978; Malt & Smith, 1984; Rips, 1975; Rosch, 1977), and therefore is an important criterion to evaluate theories of concept representation. Both approaches under consideration provide clearly distinct accounts of typicality differences: According to a similarity-based approach, typicality derives from the similarity relations between category members. The more a category member resembles other category members, the more typical it is of the category (e.g., Hampton, 1979; Rosch & Mervis, 1975). The ideal-based approach, on the other hand, proposes that typicality is determined by "characteristics that a category member should have to best serve a goal associated with a category" (Barsalou, 1985, p. 630). Objects that display such characteristics are expected to be more typical of the category.

A first apparent difference between the ideal-based and similarity-based approaches is that ideals tend to focus on a limited set of characteristics (perhaps only one) that are tightly related to the goal of the cognitive agent. Exemplar similarity, on the other hand, generally relies on a more global (i.e., multidimensional) similarity between category members, as can be found in salient groupings following the correlational structure of the world (as in Barsalou, 1985). This, however, is not an essential aspect of similarity. As was already indicated in the introduction, more sophisticated implementations of this approach can allow attention to focus on a limited set of relevant characteristics through dimensional weighting. If this were the only difference between ideal- and similarity-based representations, the two approaches would not be distinguishable.

There is, however, a second, crucial difference. Contrary to the similarity-based representations, ideal representations are not based on the most common values (or central tendencies) of the category members. Rather, extreme values on the ideal dimension are what drive typicality. Extreme values are by definition uncommon, or even unrealistic. For example, the ideal of the category *things to eat on a diet* presumably is "zero calories," which clearly is not a common value for the category members. Put differently, the ideal approach proposes a "more-is-better" principle: The more an item has of whatever it is that constitutes the ideal of a category, the more typical it will be judged (Voorspoels, Storms, & Vanpaemel, 2011). Note that the similarity-based approach favors common, frequent, central-tendency values, due to its dependence on other category members. As such, extremity versus central tendency is

the crucial difference between the ideal approach and the similarity approach.

By means of correlational methods, Barsalou (1985) evaluated the contributions of both similarity-based processes and ideal-based processes in a set of goal-derived categories. In particular, he compared correlations between the typicality scores of the members of a goal-derived category and the summed similarity to other category members with correlations between the typicality scores and the idealness of members relative to a crucial characteristic related to the goal of the category. As expected, in world-derived categories, the typicality of category members depended on their global similarity to other category members. In goal-derived categories, however, typicality was largely determined by the extent to which a member displayed the ideal characteristic. For example, for the world-derived category of *vehicles*, a member was judged typical if it was sufficiently similar to other vehicles. For the category *foods not to eat on a diet*, on the other hand, typical members were high in calories. Crucially, Barsalou (1985) used a measure of global exemplar similarity in his analyses and did not consider that similarity can, in principle, focus on a limited set of characteristics.

Goals of the present research

The present study focuses on goal-derived categories and reexamines whether a similarity-based approach can accommodate the typicality gradient when a more flexible notion of similarity is considered. We compared two computational cognitive models in their accounts of the typicality gradients of ten goal-derived categories: a similarity-based exemplar model and an ideal model. The exemplar model (the generalized context model; Nosofsky, 1986) assumes that a concept is represented by its members and that typicality is the summed similarity to other category members. Crucially, the model allows for differentially weighting the stimulus dimensions (Nosofsky, 1986). The ideal-based model (the ideal dimension model; Voorspoels, Vanpaemel, & Storms, 2011) relates the typicality of a category member to the value of the member on an ideal dimension in an underlying stimulus representation. The key idea underlying the ideal-based model is that the extent to which members display an ideal characteristic (or a combination of ideal characteristics) monotonically determines their typicality within the goal-derived category. In effect, the ideal-dimension model relates typicality to a characteristic in a "more-is-better" fashion, differentiating it from the generalized context model, which predicts that common values will drive typicality.

In what follows, we first describe the data collection, as well as a number of preliminary statistical analyses that confirmed Barsalou's (1985) findings. Then we apply the computational models of exemplar similarity and idealness to the typicality judgments, elaborating on the models and presenting the model comparison in detail. In the Discussion section, we take a step back and look at our results in the light of earlier findings regarding ideals.

Experiment

We based our analyses on a newly gathered data set. The methodology followed in both the data collection and the preliminary analyses was largely identical to that of Barsalou (1985). We ran a new experiment to implement an empirical approach to determining the nature of the ideals—instead of intuitively selecting ideals, as Barsalou (1985) did—and we applied a different technique to pairwise similarity: a sorting task. More details will be given in the corresponding sections.

Method

Ten goal-derived categories were used: *things you put in your car*, *things you rescue from a burning house*, *things not to eat/drink when on a diet*, *wedding gifts*, *things you use to bake an apple pie*, *things you take to the beach*, *means of transport between Brussels and London*, *properties and actions that make you win the election*, *weapons used for hunting*, and *tools used when gardening*.[1]

Exemplar generation

A group of 160 undergraduate students participated in the exemplar generation task for course credit. The ten goal-derived categories were divided into two groups of five categories. For each set of five categories, we created four booklets for a pen-and-paper task, containing the five categories in different orders. For each category, ten lines were supplied for answering. Each of the constructed booklets was filled in by 20 participants, leading to 80 responses for each category. The participants were asked to write down at least eight and a maximum of ten members for the categories on the open lines under the category label. The task took no more than 15 min.

For each category, 20–25 members were selected for further study. These were sampled to cover the entire generation frequency dimension. For eight of the categories, we selected 20 category members. For two categories, we selected 25 exemplars (*things not to eat when on a diet* and *wedding gifts*). Only the selected exemplars were used in the remainder of the study.

[1] In what follows, we will use shorter descriptions for the categories. For the data collection, however, these are the descriptions that were used.

Ideal generation

Contrary to Barsalou (1985), we asked participants to generate potential ideals for the ten goal-derived categories. In total, 25 undergraduate students participated for course credit. Each participant received a booklet containing a short introduction and instructions for the task. For each of the categories, they were asked to generate characteristics or qualities that members of the categories would ideally display. Each booklet contained all ten categories. For each category, the participants could write down up to seven characteristics. The number of ideals per category ranged from 5 to 13, with an average of 7. We considered ideals that were generated more than four times for the further data collection.

Goodness of example

The typicality gradient of the goal-derived categories was measured by the goodness of each example. Sixty undergraduate students participated for course credits. The selected exemplars from the generation task were used as the stimuli in this study. The ten goal-derived categories were divided into two groups of five categories each. On the basis of this division, 12 booklets were constructed for a pen-and-paper task: For each group of five categories, the categories were presented in two different orders, and the members of each category were presented in three different orders.

The participants were each given a booklet containing instructions and the five categories for which typicality judgments were asked. In the written instructions, they were asked to indicate the extent to which each of the presented members was a good example of the category on a 20-point Likert scale, with 1 denoting *a very bad example* and 20 denoting *an excellent example*. In total, the typicality for the members of each category was judged by 30 participants.

The reliability of the judgments was evaluated by means of split-half correlations corrected with the Spearman–Brown formula, which ranged from .94 to .98.[2] The judgments were averaged across participants, resulting in a single goodness-of-example score for each exemplar. The goodness-of-example scores ranged from 5.65 to 18.98, with a standard deviation of 3.89 (these values were averaged across categories).

Exemplar similarity

To arrive at a measure of pairwise similarity between category members, we opted for a sorting task,[3] an often-applied technique for arriving at a similarity measure for stimulus sets (e.g., Ameel & Storms, 2006; Van der Kloot & van Herk, 1991).

Sixty undergraduate students participated for course credit. Each participant sorted the exemplars of each category in a completely computerized task. The order of the categories was randomized. Participants were instructed to divide each set of words into groups according to whatever principle they thought was fitting, the only restriction being that there had to be more than one group and less than the number of members in a category. They were then presented with the exemplar set of a category on a computer screen and could drag the verbal stimuli into groups. Following the initial sort, they were asked either to further divide the groups that they had made into subgroups, when this was deemed possible (when the number of groups in the initial sort was less than five), or to join groups together (when the number of groups was greater than five).

The sorting procedure resulted in 120 exemplar-by-exemplar matrices per participant for every category (each of the 60 participants sorted each exemplar set twice). The matrix cells reflected whether a certain exemplar pair was in the same group for a particular sort, indicated by 1 (vs. 0 if the pair was not in the same group). The estimated reliability of the similarity data ranged from .94 to .98. The 120 matrices were summed for each category, leading to an exemplar-by-exemplar similarity matrix, with the cells entries reflecting the number of times that two exemplars were in the same group across the 120 matrices, with higher numbers reflecting higher similarity. The similarity matrices were used to calculate a measure of exemplar similarity, by summing the similarity of a member to all other members of the category. The resulting sum was large if a member was similar to many category members, and small if it was dissimilar to most category members.

Exemplar idealness

For the exemplar idealness judgments—that is, the degree to which the members of a category displayed an ideal property—216 participants volunteered. Each participant judged the idealness of an exemplar set relative to one ideal for five categories (a different ideal for each category), yielding 15 participant judgments for each ideal. The participants were instructed to indicate on a 7-point Likert scale the extent to which the presented objects (i.e., the members of the category for which the ideal was generated) possessed the quality, without mention of the category in which the objects were generated. The estimated reliability of the judgments ranged from .71 to .98, with an average of .89. The judgments were averaged across participants, resulting in a single score for each exemplar on each relevant ideal.

Correlational analyses

To evaluate whether the earlier findings would emerge from our set of goal-derived categories, we ran analyses very similar to Barsalou's (1985). In particular, we examined

[2] This method of reliability estimation is applied throughout the article.
[3] Barsalou (1985) relied on judged pairwise similarities.

the relation between exemplar similarity, idealness, and typicality by calculating the correlation across members between the different measures for each category. For both exemplar similarity and idealness, we expected a positive correlation with typicality: According to a similarity-based approach, more similarity to other members leads to a higher typicality, and according to the ideal approach, having more of whatever the ideal characteristic is leads to higher typicality. Note that, whereas Barsalou (1985) also included familiarity and the frequency of instantiation in his analyses, we focused on exemplar similarity and idealness.

First, we considered only the most frequently generated ideal for each category. Given that these charateristics were suggested most often by participants as crucial, they are presumably the most intuitive ideals.[4] The resulting correlations with typicality are presented in Table 1, which presents, for all ten categories, the raw and partial correlations between goodness of example and either exemplar similarity or idealness. If we focus on the raw correlations, averaged across categories, the correlation between exemplar idealness, using the most frequently generated ideal, and the goodness of example is .70, which is identical to the corresponding correlation reported by Barsalou (1985, Table 1, p. 636). The correlation is positive and significant for each category separately, being above .8 for five of the ten categories and dropping below .5 for only one category (*things you use to bake an apple pie*, $r = .43$, $t(18) = 2.00$, $p_{\text{directional}} = .03$).

For exemplar similarity, on the other hand, the raw correlation with the goodness-of-example ratings, averaged across categories, is only .14 (vs. .38 in Barsalou, 1985, p. 636). Looking at the categories separately, exemplar similarity achieves a significant positive correlation with typicality in only two out of the ten categories [*diet ruiner*, $r = .50$, $t(23) = 2.75$, $p_{\text{directional}} = .003$; *election strategies*, $r = .66$, $t(18) = 3.68$, $p_{\text{directional}} < .001$]. For *things you rescue from a burning house*, we found a strong negative correlation, which makes exemplar similarity a good predictor for this category, but a negative correlation is not in accord with the predictions of a similarity-based approach [$r = –.82$, $t(18) = –6.18$, $p_{\text{directional}} > .999$].

In further analyses, we calculated first-order partial correlations between the goodness-of-example scores and either exemplar similarity or idealness, controlling for the other. In this way, the unique predictive power of the predictor under examination was assessed. These partial correlations are presented in Table 1, in the third and fifth columns. Controlling for exemplar similarity, idealness on average correlate .65 with goodness of example, reaching significance in all categories except *things you rescue from a burning house* [$r = .25$, $t(17) = 1.38$, $p_{\text{directional}} = .092$]. Controlling for idealness, the correlation between exemplar similarity and goodness of example drops to .08, averaged across categories. Only in one category does the first-order partial correlations of exemplar similarity with typicality, controlling for idealness, rise to statistically significance [*things you use to bake an apple pie*; $r = .51$, $t(17) = 2.43$, $p = .01$].

Finally, we examined two additional aspects of the ideal approach, which remained untouched in Barsalou (1985). First, the most frequently generated ideals—that is, the most *intuitive* ones—need not necessarily correspond to the ideals that drive typicality. In other words, the "real," typicality-driving ideals are not necessarily the ones that people generate most frequently. Second, there is no a priori reason that only one ideal would be involved in a category. To further examine these issues, we performed a stepwise regression analyses with goodness of example as the dependent variable and the set of ideal ratings as independent variables. The appropriate ideals were selected from the pool of available ideals (from the generation task) using forward selection and were allowed in the model if they provided a significant increase in R^2 (as quantified by an incremental F test). When none of the remaining ideals contributed significantly, the selection procedure stopped.

Table 2 presents the R^2s of the subsequent regression models for each category. I1 refers to the first ideal that was entered, I2 to the second ideal, and so forth. The rightmost column shows the verbal descriptions of the ideals that were entered, in the order that they entered the regression model (i.e., the first verbal description refers to I1, the second to I2, and so on). For example, for *car trinkets*, a regression model with the ideal "useful" as a predictor and typicality as the dependent variable results in an R^2 of .86, with "useful" being the ideal that reflects the highest correlation with typicality. Of the remaining ideals, adding "easy to store" to the regression model resulted in the largest significant contribution to R^2, raising it to .9 [$F_{\text{incremental}}(1, 17) = 6.42$, $p = .021$]. Adding more ideals to the regression model did not result in further significant rises in R^2, and therefore the stepwise procedure was stopped. Apparently, for *car trinkets*, not the most frequently generated ideal —"small"—but a combination of two ideals—"useful" and "easy to store"—provides the best account of the typicality gradient, explaining 90 % of the variability. For the remaining categories, the results show that for all but one (*gardening tools*), a combination of ideals drives the typicality gradient. Moreover, only for the category *wedding gifts* does the most frequently generated ideal also reflect the highest correlation with the typicality scores.

Conclusion

In sum, the data that we have collected for the computational model analyses show the same basic pattern reported earlier (Barsalou, 1985): Idealness gives a far better account

[4] In Barsalou (1985), the ideals were chosen intuitively for each category.

Table 1 Raw and partial correlations of exemplar similarity and idealness with typicality for each of the ten goal-derived categories

Category	Exemplar Similarity		Idealness		
	Raw *r*	Partial *r*	Raw *r*	Partial *r*	Verbal Description
Car trinkets	.02	.12	.53**	.54**	small
Burning house	–.82	–.68	.66**	.25	emotional value
Diet ruiners	.50**	.13	.59**	.39*	contains sugar
Wedding gifts	–.01	–.46	.81**	.85**	useful
Pie necessities	.33	.51*	.43*	.57**	heat-resistant
Beach trinkets	.18	.19	.61**	.62**	UV protection
Means of transport	.13	.31	.88**	.89**	fast
Election strategies	.66**	.23	.82**	.68**	reflects charisma
Hunting weapons	.13	.29	.83**	.84**	weighs little
Gardening tools	.34*	.13	.88**	.86**	useful
Averages	.14	.08	.70	.65	

The rightmost column presents verbal descriptions of the most frequently generated ideals, which constitute the idealness measure in these analyses

*$p_{\text{directional}} < .05$

**$p_{\text{directional}} < .01$

of the typicality gradient of the present categories than does exemplar similarity. Interestingly, the ideal approach could be improved even more by not restricting it to using the most intuitive ideal (that is, in our case, the most frequently generated one) and by allowing more than one ideal. Identification of the optimal (combination of) characteristics clearly is imperative in accounting for the membership structure in goal-derived categories. The similarity approach, which relied on global similarity between the category members, was not able to capture the typicality gradient.

Computational modeling results

Next, we tested whether a similarity-based model that allows for differential weighting of stimulus dimensions can provide a better account of the typicality gradient in goal-derived categories than can an ideal-based model. Differential weighting of the dimensions allows the similarity-based model to focus on the relevant characteristics. For the ideal approach, we considered a model that derives typicality from a single ideal dimension in the underlying stimulus representation, taking into account that more extreme values lead to more typicality. We then used the goodness-of-example ratings as a dependent variable to test the model predictions. The models base their predictions on underlying geometric stimulus representations, derived from the sorting data.

Before describing the results of the model evaluations, we will discuss the models in more detail. For an overview of the formal implementations, we refer the reader to Appendix A.

Models

Several similarity-based models of concept representation have been developed on the basis research on world-derived categories. For the present study, we focused on the exemplar approach to concept representation (Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1986). It has been shown that the exemplar approach is more successful than other similarity-based approaches to category representation in a broad array of conditions (Nosofsky, 1992; Vanpaemel & Storms, 2010; Voorspoels, Vanpaemel, & Storms, 2008).

According to the exemplar view, categories are mentally represented by memory traces of previously encountered members, and typicality is the summed similarity to all members of the category. In this view, the concept *dog* consists of representations of previously encountered instances of dogs, such as *poodle*, *German shepherd*, and *dalmatian.* A German shepherd is a typical dog because it is highly similar to many stored instances of *dog*. In this study, the exemplar view is implemented by the generalized context model (GCM; Nosofsky, 1986). Crucially, the GCM makes use of dimensional weights when calculating similarity. These weights, which constitute the free parameters of the model, implement the idea that, depending on the specific context, some features may increase or decrease in importance.

To implement the ideal-based approach, we made use of the ideal-dimension model (IDM; Voorspoels, Vanpaemel & Storms, 2011), which posits that an ideal dimension exists in the underlying stimulus representation and determines the typicality gradient of the category. Each exemplar of a category has a certain value along this ideal dimension. The farther the value is located out along the dimension in the ideal direction, the more typical the exemplar is. For example, the lower a food product is in calories, the more typical it is of the category *things you eat when on a diet*, and the more valuable something is, the more typical it is of *things you rescue from a burning house*.

According to the IDM, the ideal dimension can be found in an underlying geometric stimulus representation, and typicality predictions can be derived solely on the basis of

Table 2 R^2 values in regression analyses, using a forward selection method based on the rise in R^2 when adding an ideal

Categories	Ideals Entered			Verbal Description (Corresponding to I1–I3)
	I1	I2	I3	
Car trinkets	.86	.90		useful, easy to store
Burning house	.91	.96	.97	important, emotional value, financial value
Diet ruiners	.86	.90		unhealthy, contains sugar
Wedding gifts	.67	.79		useful, personal
Pie necessities	.74	.79		appropriate shape, useful
Beach trinkets	.75	.85	.89	easy to use, protects against wind, entertaining
Means of transport	.89	.92		comfortable, cheap
Election strategies	.94	.97		reflects integrity, grabs attention
Hunting weapons	.91	.93		easy to use, effective
Gardening tools	.90			efficient

I1 refers to the first ideal that enters the model, I2 to the second ideal, and so on. Only the ideals that provide the largest contribution to R^2 in each phase, and only significantly contributing ideals, are presented. The verbal descriptions are rank-ordered in the same way—that is, for *car trinkets*, "useful" is I1 and "easy to store" is I2

this representation. The free parameters in the model allow it to find the appropriate dimension in the stimulus representation. An apparent advantage of this approach is that we do not need to determine the "correct" ideal a priori; the model is sufficiently flexible to identify the dimension that provides the best correlation with typicality. Interestingly, this implies that the ideal does not require verbalization, and that the model is not restricted to a single ideal, but can also cope with a combination of several ideals. The only restriction is that the stimulus dimension that constitutes the ideal be present in the underlying stimulus representation.

To reiterate, by applying the GCM and the IDM, both the similarity-based approach and the ideal-based approach are allowed to focus on the relevant characteristics, or combination of characteristics, to derive an account of the typicality gradient in each goal-derived category. The essential difference between the two approaches remains, however: Whereas the GCM relates typicality to common values in the category, the IDM relies on the principle of "more-is-better." That is, whereas according to the GCM, members are typical if they display dimensional values that are common among members of the category, the IDM predicts high typicality scores for members that have extreme values on the relevant dimension. Through a thorough comparison of the two models, we can identify which of these two principles is most successful in goal-derived categories. A crucial aspect of the comparison is that both models rely on the same underlying geometric stimulus representation, constructed from the same similarity data. Differences in model performance can thus be ascribed solely to the different cognitive processes implemented by the models.

Model evaluation

To evaluate and compare the computational models, we focused on two complementary methods. First, we evaluated whether the models were able to account for the observed typicality gradient in the categories through *goodness-of-fit* analyses. The model parameters were optimally set to produce model-based typicality scores that reflected the highest correlation with the observed typicality judgments. The goodness-of-fit was calculated by maximizing the correlation between the observed typicality scores and the model-based typicality scores, derived from the GCM and the IDM (with optimal parameter settings).

In a next step, we compared the models. The model comparison question cannot be solved by relying solely on goodness-of-fit, as inherent differences in complexity between the models could bias the conclusion. All things being equal, more flexible models can generally provide a better data fit than more parsimonious models. Model comparison therefore should rely on *generalizability*, which balances the inherent complexity in the models with goodness-of-fit (Myung, 2000; Pitt, Kim, & Myung, 2003). We relied on marginal likelihood as a measure of generalizability, reflecting a model's weighted average likelihood across the parameter space (Kass & Raftery, 1995). Details on both model evaluation methods can be found in Appendix B.

Results

Stimulus representation

The cognitive models applied in this study rely on an underlying geometric stimulus representation. In a geometric similarity representation, stimuli are represented by points in an M-dimensional space, and the distance between two points is inversely related to the similarity between the corresponding stimuli. The space is spanned by orthogonal axes that can be considered independent features that are relevant in the similarity structure of the stimuli.

The similarity matrices from the experiment were used as input for SAS nonmetric multidimensional scaling (MDS) analyses, evaluating solutions from Dimensionalities 2 to 5.

Stress values, measuring the badness of fit for the resulting geometric representations, showed a monotonically decreasing pattern in each category, which suggests that the algorithm did not get trapped in a local minimum. Overall, the stress values dropped below .1 from Dimensionality 3 onward. Taking into account the rules of thumb regarding stress and number of dimensions (Kruskal & Wish, 1978), we considered the MDS solutions from Dimensionalities 3 to 5 for further analyses.

Model evaluation and comparison

We first present the goodness-of-fit of the two models for the observed typicality in the ten goal-derived categories. Goodness-of-fit is measured as the optimal correlation between the model-based scores and the observed typicality scores—that is, with optimal parameter values. The model-based scores were calculated for underlying stimulus representations of Dimensionalities 3–5. The results are presented in Fig. 1.

Regarding goodness-of-fit, the pattern in the results is clear and consistent across categories: The GCM clearly struggles to attain a good data fit for the goal-derived categories. In Dimensionality 5, averaged across categories, it gives an optimal correlation of .64. Moreover, for four out of the ten categories, the correlation does not rise over .60. The IDM, on the other hand, gives an excellent data fit for all categories, on average .8, and except for the category *beach trinkets*, the correlations are well above .7. These results indicate that the similarity-based model has considerable difficulty accounting for the typicality gradient, whereas the ideal-based model produces good to excellent optimal correlations.

Model comparisons, however, should rely on generalizability. We performed additional analyses based on marginal likelihood. For each of the categories, we compared the marginal likelihood of the models. The results are presented in terms of model weights, which reflect the relative evidence for a model within the set of models (Burnham & Anderson, 2002; Lee, 2004): The evidence for a particular model is quantified by its marginal likelihood, and the relative evidence for the model is its marginal likelihood divided by the sum of the marginal likelihoods of the two models. To estimate the marginal likelihood, we relied on standard noninformative priors. The attentional weights of the GCM were uniformly distributed in the interval [0, 1], with the restriction that they summed to 1, and for the sensitivity parameter, we used the standard approximation to the Gamma(0, 0) distribution. For the IDM, the parameter priors translated the idea that all possible dimensions in the space were equally likely. The results of the generalizability analyses are presented in Fig. 2.

The pattern in the generalizability analyses confirms the earlier results. For only one category (*election strategies*) the GCM is the preferable model. In this category, both models perform very similarly in terms of goodness-of-fit, which suggests that the GCM was slightly favored in the generalizability analyses for being less flexible (see, e.g., Jang, Wixted, & Huber, 2011). For seven out of the ten categories, evidence is overwhelmingly in favor of the IDM, consistently across dimensionalities. In two categories (*hunting weapons* and *gardening tools*), the relative model performance is not consistent across dimensionalities, so it is difficult to decide in favor of any of the models. Note, however, that for gardening tools, the stimulus representations in Dimensionalities 4 and 5 provide overwhelming evidence for the IDM. The goodness-of-fit analyses for this category already suggested that the IDM could not achieve a good data fit in Dimensionality 3.[5]

Conclusion

Both goodness-of-fit and generalizability analyses provided overwhelming evidence that goal-derived categories are represented by an ideal dimension, rather than by similarity-based representation, and that membership judgments regarding an object derive from the value of this object along the ideal dimension. Interestingly, the similarity-based model had access to the same crucial dimensions in the underlying geometric stimulus representations and could, in principle, limit the similarity calculations to appropriate dimensions through dimensional weighting. Contrary to the IDM, however, the GCM relates the typicality to frequent or average values on these dimensions, rather than to ideal, extreme values. Clearly, the "more-is-better" aspect of the IDM is crucial to account for the observed typicality patterns in goal-derived categories.

General discussion

The finding that the typicality gradient in goal-derived categories is mainly driven by ideals rather than by exemplar similarity has stood uncontested for nearly three decades. While the early evidence was indeed compelling and intuitive, the finding has never been replicated using more sophisticated, computational models of representation. Therefore, due to the rather rigid earlier implementations of similarity, a key question has remained—that is, whether a more flexible approach to similarity would alter the conclusions. Indeed, since Barsalou (1985), it has been extensively argued and demonstrated that

[5] Additional analyses revealed that the most important ideal dimension for the category *gardening tools* ("efficiency") was not at all present in the three-dimensional MDS solution. It could, however, be found in the solutions of Dimensionality 4 onward. This clearly shows in the goodness-of-fit and generalizability analyses.

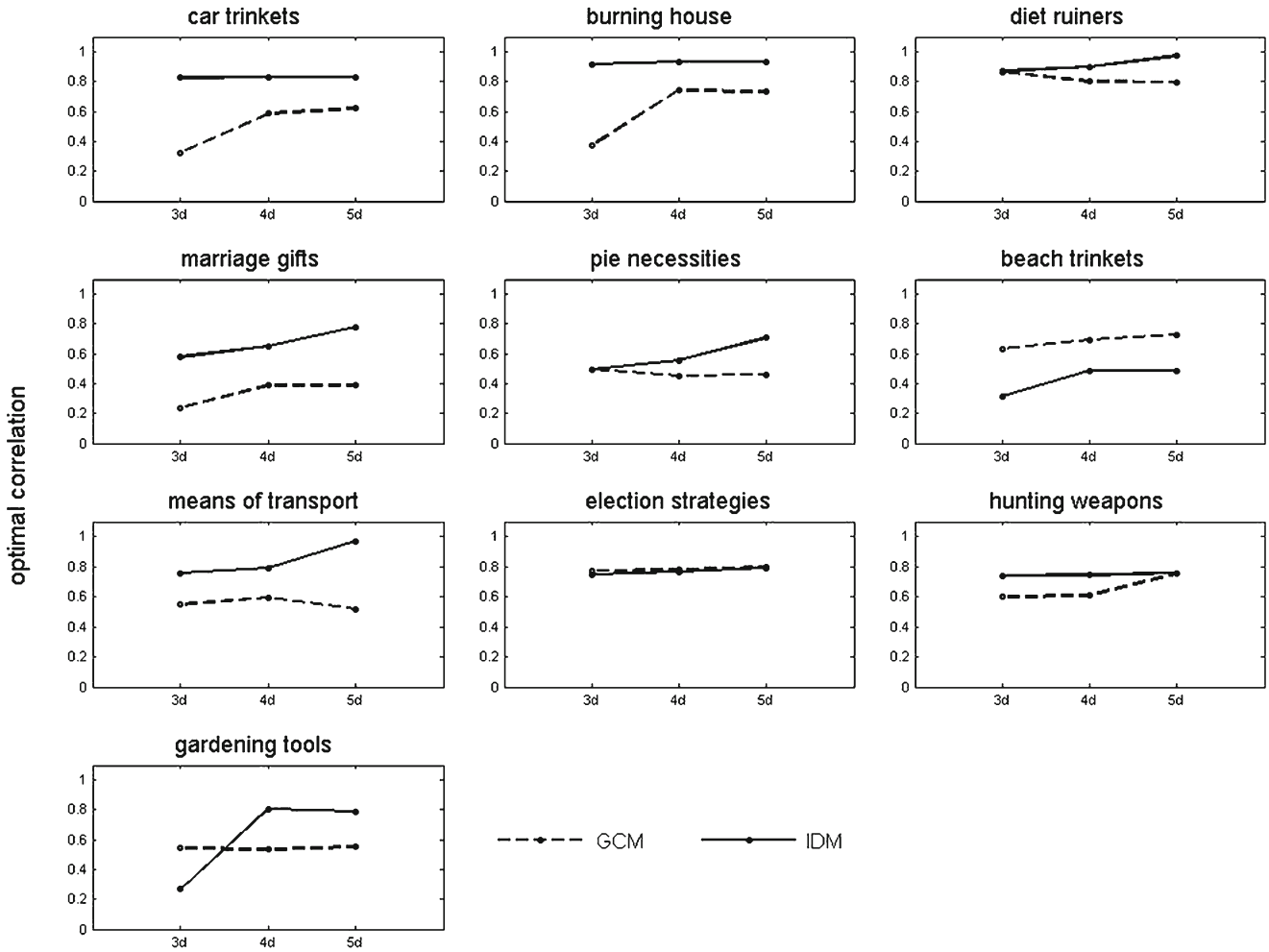

Fig. 1 Goodnesses of fit for the generalized context model (GCM) and the ideal-dimension model (IDM) for the ten goal-derived categories in Dimensionalities 3–5. For each category, the correlations between the observed typicality judgments and model-based typicality scores are presented for both models, as a function of the dimensionality of the underlying geometric stimulus representation

similarity is not invariant across contexts (e.g., Nosofsky, 1986), and that people tend to apply a differential weighting to stimulus dimensions, depending on the requirements of the context. In the present study, we tested whether a similarity-based approach that allows for such differential weighting can account for the typicality gradient of goal-derived categories, and thus, whether the theoretical notion of an ideal is indeed required, and what differentiates an ideal-based approach from a more flexible similarity-based approach.

In a first phase, we used empirical measures of idealness and exemplar similarity and applied correlational methods to relate these measures to typicality in a set of ten goal-derived categories. In all ten categories, we found solid evidence that exemplar typicality ratings were best accounted for by the extent to which the exemplars had an ideal quality, replicating the classic finding of Barsalou (1985). On the basis of these data, we compared two computational models in their accounts of the typicality gradient of the categories: the GCM (Nosofsky, 1986), a similarity-based, exemplar-based model, and the IDM (Voorspoels, Vanpaemel, & Storms, 2011). For the present study, it was crucial that the GCM allows differential attentional weighting of underlying stimulus dimensions. In terms of both goodness-of-fit and generalizability, we found overwhelming evidence in favor of the ideal approach, again for nearly all of the categories. The IDM offered an excellent quantitative account of the observations and convincingly performed better than the GCM.

In sum, both correlational methods and thorough model-based analyses lead to the same conclusion: The typicality gradient in goal-derived categories is determined by ideals rather than by exemplar similarity. Importantly, however, the computational model analyses revealed that a more flexible notion of similarity still cannot accommodate the findings in goal-derived categories. Both the GCM and the IDM were allowed to search the same stimulus representation, constructed from the same similarity data, for the most relevant dimensions. Apparently, the aspect of "more-is-

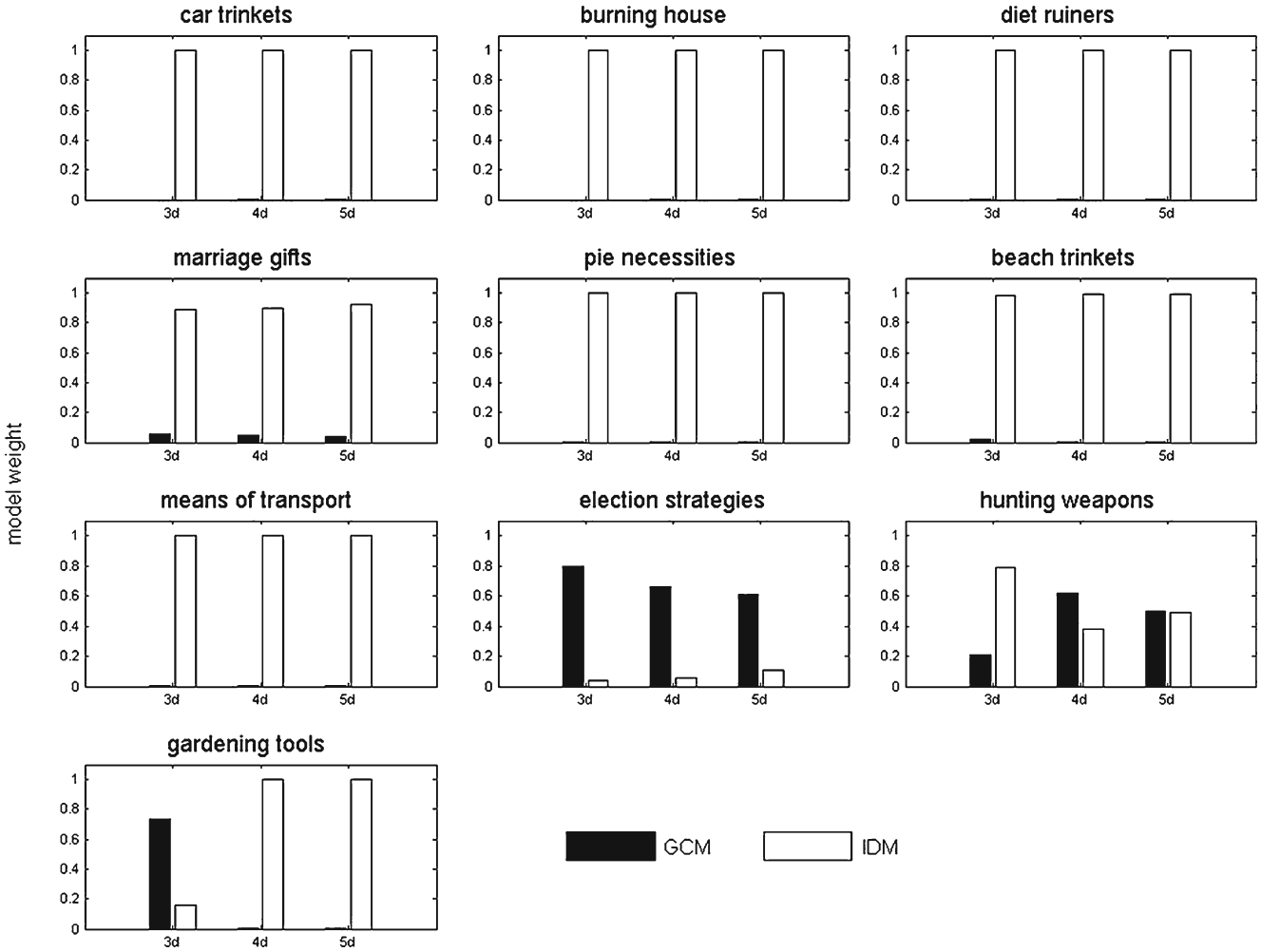

Fig. 2 Model weights for the generalized context model (GCM) and the ideal-dimension model (IDM) for the ten goal-derived categories. The model weights express the relative evidence for a particular model as compared to the other model, with higher bars referring to more evidence for the corresponding model. Evidence for a model is quantified through its marginal likelihood—that is, the data fit of the model when taking into account its inherent complexity

better" that is implemented in the IDM proved to be a crucial underlying principle in the goal-derived categories. In other words, given a focus on the relevant dimensions, extreme, ideal dimensional values seem to drive typicality rather than common values. Contrary to the IDM, the GCM relies on similarity comparisons with category members, and therefore predicts that objects displaying dimensional values that are common in the category will be more typical. While the GCM was able to focus on the relevant dimensions, this alone did not lead to an adequate account of typicality.

One could argue that the generality of this conclusion is somewhat overstated, since we have considered only one particular implementation of dimensional weighting, in which a particular dimension was given more or less weight, respectively stretching or shrinking the stimulus representation along the entire dimension. Other approaches do exist—for example, in which a dimension is stretched or shrunk in a specific interval, but left untouched in other regions (such mechanisms are thought to play a role in categorical perception; see, e.g., Goldstone & Hendrickson, 2010)—and other approaches can be conceived of two considerations are appropriate/important at play, however: First, the GCM implements a very intuitive and widespread manner of attentional weighting that has been proven successful in a broad array of conditions and experiments. Second, making similarity comparisons with other category members will favor common, central-tendency values over rare, extreme values in terms of typicality, independent of the particular derivation of similarity.[6] In light of the excellent data fit of the IDM for the majority of the categories, it seems that the extremity proposed by means of the "more-is-better" principle is exactly what captures the membership

[6] An exemplar model can account for psychological extreme points relative to a contrast category in categorization tasks (Palmeri & Nosofsky, 2001), yet in the present case there is no immediate contrast category.

structure. Consequently, we expect other approaches to dimensional weighting to experience similar difficulties in accounting for the typicality gradient.

The essential difference between world-derived and goal-derived categories seems to be the extent to which conceiving of the category can depend on its extension. In world-derived categories, one can rely on experience with the exemplars of the corresponding salient grouping in the environment. The extension is crucial in the formation of these concepts (Barsalou, 1985, 1991), and similarity comparisons with category members are essential in judgments regarding these concepts (e.g., Barsalou, 1985; Rosch & Mervis, 1975; Voorspoels et al., 2008). In goal-derived categories, on the other hand, a cognitive agent must conceive of the properties that are desirable for the potential members before these objects are classified (Barsalou, 1985, 1991). Consequently, one cannot rely on average or common values to determine membership or typicality; an extension has yet to be established. In a sense, the ideal representation does not describe the category as a similarity-based representation does. Rather, it posits the normative criteria that must be met by objects to be considered members of the category.

Ideals in a similarity space

The IDM translates the norm that people apply when conceiving of a goal-derived category into an ideal dimension in a geometric stimulus representation and relates (graded) membership in the category to the extent to which objects adhere to this norm. This approach adds a number of advantages over earlier implementations of ideals, allowing the notion of the ideal to be applied to a broad range of contexts. Most notably, the IDM does not restrict its search to the most intuitive ideal, nor is it restricted to only one ideal, nor does it require that the ideal be verbalized. The only restriction for the IDM is that the relevant dimensions be present in the underlying representation. Bearing in mind that in our empirical phase we concluded that for the majority of the categories, the most intuitive ideal was not the best predictor of typicality, and that a combination of different ideals produced the best account of typicality, the flexibility of the IDM in identifying the ideal dimension is crucial. Importantly, these advantages allow the notion of an ideal to be applied to a broad range of contexts in which little or no a priori information is available regarding which ideals should be considered.

Interestingly, while the IDM does not relate typicality to similarity, as does the GCM, the model does rely on a geometric stimulus representation that essentially positions exemplars in terms of their similarity. In other words, the similarity structure[7] in the environment does contain the information necessary to account for the graded membership structure in categories that seemingly reflect no similarity relations whatsoever. The IDM thus relates typicality to similarity. Yet, the cognitive process that produces the typicality judgments clearly deviates from a similarity-based process. The similarity-based approach is similarity-based because its proposed process is essentially a similarity comparison, whereas according to the ideal approach, typicality results from an evaluation on a single dimension. This difference in cognitive processes justifies our distinction between the two approaches. Yet, at the same time, it is clear that the present results underline the strength of similarity as an explanatory construct.

Ideals as representations of concepts

The finding that a similarity-based model cannot account for the typicality gradient in goal-derived categories supports the hypothesis that the human conceptual system is flexible and can accommodate different types of representations and different systems of categorization (e.g., Ashby & Valentin, 2005; Blair & Homa, 2003; Minda & Smith, 2001; Poldrack & Foerde, 2008; Voorspoels, Storms, & Vanpaemel, 2011). Next to similarity-based representations such as exemplar and prototype representations, idealness, too, seems to be an organizing principle of category representations.

Interestingly, since the introduction of the notion, evidence for ideals has not exclusively been found in goal-derived categories. In fact, a number of studies have provided compelling support for the existence of ideal representations in different kinds of concepts. For example, tree experts have been found to include ideal dimensions such as "height" and "weediness" in their representations of the concept *tree* (taller and less weedy trees are rated as more typical; Lynch, Coley, & Medin, 2000); for fishermen, culturally determined dimensions of desirability are aspects of the representation of *fish* (Burnett, Medin, Ross, & Blok, 2005); and ideals regarding beauty and meat tastiness determine *bird* typicality for the Itzai Maya from Guatemala (Atran, 1999). Also, ideals have been shown to determine typicality in consumer product categories (Loken & Ward, 1990) and personality attribute categories (Borkenau, 1990). Moreover, there is evidence that the level of a concept in a conceptual hierarchy influences the type of representation, even for seemingly normal, world-derived categories. For superordinate categories such as *clothes* and *mammals*, an ideal representation has been found to be more apt at explaining human typicality judgments than are either exemplar or prototype models (Voorspoels, Storms, & Vanpaemel, 2011; Voorspoels, Vanpaemel, & Storms, 2011), whereas in subordinate-level categories (*pants*, *shirts*), an exemplar representation fares better.

The general pattern in these findings is that the success of ideal representations is tightly tied to the extent to which categories are a product of either the environment or the

[7] Note that in the sorting task, the participants were not given any specific instructions as to how to group the different stimuli.

cognitive agent. Let us consider the dimension that runs from goal-derived to world-derived categories. At one end of the dimension are the goal-derived categories examined in the present study. In these categories, no imposing and salient grouping is present in the world, but rather, the grouping is entirely dependent on the categorizer's goals. These goals set criteria that an object must meet to belong to the category, invoking an ideal representation.

Toward the world-derived extreme of the dimension, one can encounter basic-level categories that are maximally informative—that is, in these categories the within-category similarity and between-category dissimilarity is optimized (Rosch et al., 1976). As a consequence, basic-level categories are saliently present in the environment and do not require an active categorizer to identify the category. These categories are mainly driven by the obvious similarity between the exemplars, and exemplar similarity can be expected to be important in the representation. Empirical evidence clearly supports this claim (e.g., Barsalou, 1985; Hampton, 1979; Rosch & Mervis, 1975; Voorspoels et al., 2008).

In between these extremes, there is ample room for different concepts. For example, one can encounter categories that are relatively high in a conceptual hierarchy. Whereas subordinate categories reflect a structure of covarying (perceptual) properties in the world (Rosch et al., 1976), superordinate categories are necessarily more diverse (in that they include a number of lower-level categories). The representation of superordinate categories therefore might depend less on the inherent structure of the world, allowing for more influence of cultural utility—for example, through cultural goals or implicit norms. Malt (1995) noted that superordinate categories display more cross-cultural diversity, which points to a greater dependence on culture-specific rather than environmental factors. Again, there is evidence for ideal representations in higher-level categories (Voorspoels, Storms, & Vanpaemel, 2011; Voorspoels, Vanpaemel, & Storms, 2011). More subtle cases emerge when culture-specific, or even profession-specific, desires and norms bear upon representations that seemingly derive from salient world groupings. This is the case, for example, with tree experts (e.g., landscapers) organizing the structure of the *tree* category around ideals concerning weediness and height—dimensions that are particularly relevant for their professional activity (Lynch et al., 2000). This is not to say that similarity is irrelevant in these categories, but it is not the only structuring principle at work. The cultural norms apparently are equally present in the fishermen mentioned earlier (Burnett et al., 2005), and perhaps, too, in consumer (Loken & Ward, 1990) and personality attribute (Borkenau, 1990) categories.

In sum, the degree to which a categorizer is an active agent in the formation of a category seems to be an important factor in the emergence of different types of category representations. For clearly salient and obvious groupings in the environment—that is, purely world-derived categories—similarity-based approaches provide an appropriate concept representation. However, when a cultural cognitive agent plays a more active role of in the construction of a concept, such that the concept is not "merely" a product of a salient structure in the world, ideal representations gain importance as underlying organization principles.

Author Notes The Research in this article is part of research project G.0281.06 sponsored by the Belgian National Science Foundation – Flanders, given to the second author. We thank Steven Verheyen and Sander Vanhaute for help with collecting data and for fruitful input, and Tim Bosschaerts for programming the sorting task.

Appendix A: Models

Generalized context model

The GCM (Nosofsky, 1984, 1986) assumes that categorization decisions are based on similarity comparisons with individually stored category exemplars. Originally, the model was developed to account for categorization decisions, but it has successfully been adapted for typicality judgments (Nosofsky, 1988, 1991; Voorspoels et al., 2008). Formally, the typicality of an exemplar i for category A with n members is given by

$$t_{iA} = \sum_{j=1}^{n} S_{ij}, \tag{1}$$

where s_{ij} is the similarity of exemplar i to exemplar j, with j belonging to category A.

The similarity between two exemplars is a function of the distance of the exemplars in the M-dimensional psychological space. Formally, the scaled psychological distance is given by

$$d_{ij} = \left(\sum_{k=1}^{M} w_k \left| x_{ik} - x_{jk} \right|^r \right)^{1/r}, \tag{2}$$

where x_{ik} and x_{jk} are the coordinates of exemplars i and j on dimension k, and w_k a free parameter reflecting the attention weight for dimension k. Since Euclidean distances are generally accepted to be more appropriate for integral dimensions (Shepard, 1987), r was fixed at 2.

Similarity between stimulus i and stimulus j is related to psychological distance, as follows:

$$s_{ij} = \exp\left(-c d_{ij}\right), \tag{3}$$

where d_{ij} is the psychological distance between stimuli i and j from Eq. 2, and c is a sensitivity parameter.

The free parameters in the model consist of $M - 1$ dimension weights and a scaling parameter c. A uniform prior over

the range 0 to 1 was used for the dimension weights, adding the restriction that the dimensional weights had to sum to 1. For the sensitivity parameter, we used the standard approximation to the Gamma(0, 0) distribution (see Jaynes, 2003, chap. 12; Lunn, Thomas, Best, & Spiegelhalter, 2000).

Ideal-dimension model

Inspired by Barsalou's (1985) notion of the ideal, the IDM (Voorspoels, Vanpaemel, & Storms, 2011) assumes that an ideal dimension exists in the underlying similarity space. Each exemplar of a category has a certain value along the ideal dimension, obtained by an orthogonal projection on this dimension. The farther that this value is located along the dimension in the ideal direction, the more typical the exemplar is. It is useful to think of the ideal dimension as a specific combination of (unarticulated) features: The more that a member has of this combination of features, the more typical it is for the category.

Formally, the IDM assumes that judging the typicality of an exemplar i for a category A comes down to evaluating the value of i on a certain dimension V_A. In an M-dimensional space, the typicality of item i for category A is then given by

$$t_{iA} = \frac{\sum_{k=1}^{M} x_{ik} x_{Ak}}{\left(\sum_{k=1}^{M} (x_{Ak})^2 \right)^{1/2}}, \quad (4)$$

where x_{Ak} determines the coordinates spanning the ideal dimension V_A, x_{ik} determines the coordinates of exemplar i, and M is the number of dimensions. We restrict x_A to be at a fixed distance from the origin. This does not pose a restriction for the ideal dimension.

In Eq. 4, the point representing object i in the representational space is projected orthogonally onto the dimension V_A that is defined by the origin and point A, determined by the coordinates x_{Ak}. This dimension can be considered a feature (or combination of features) that is relevant in the similarity structure of the category. A dimensional value is then derived relative to the origin (point zero), measured in units of length $|ox_{Ak}|$, where o is the origin. The farther the projection is in the ideal direction, as determined by V_A, the higher is the value. This is illustrated in Fig. 3, which presents the (fictive) two-dimensional spatial representation of a set of eight members of the category *things you rescue from a burning house*. The ideal dimension is projected in the representation (the arrow going from the lower left to the upper right corner), and all the members are orthogonally projected onto the ideal dimension, receiving values along the ideal dimension that predict the typicality of each member for *things you rescue from a burning house*. The farther a member's projection is in the ideal direction (i.e., the upper right corner), the more typical the member is for the category. The corresponding typicality prediction according to the IDM is given in the right bar diagram.

The free parameters in the model are the coordinates of x_A. A noninformative uniform prior across all possible positions of x_A at a fixed distance from the origin was used in the model comparison. More specifically, the prior was derived using polar coordinates, with a fixed radial coordinate (i.e., the distance from the origin) and a uniform distribution in the interval [0, 2] for the angular coordinate (i.e., the counterclockwise angle toward the x-axis in a two-dimensional

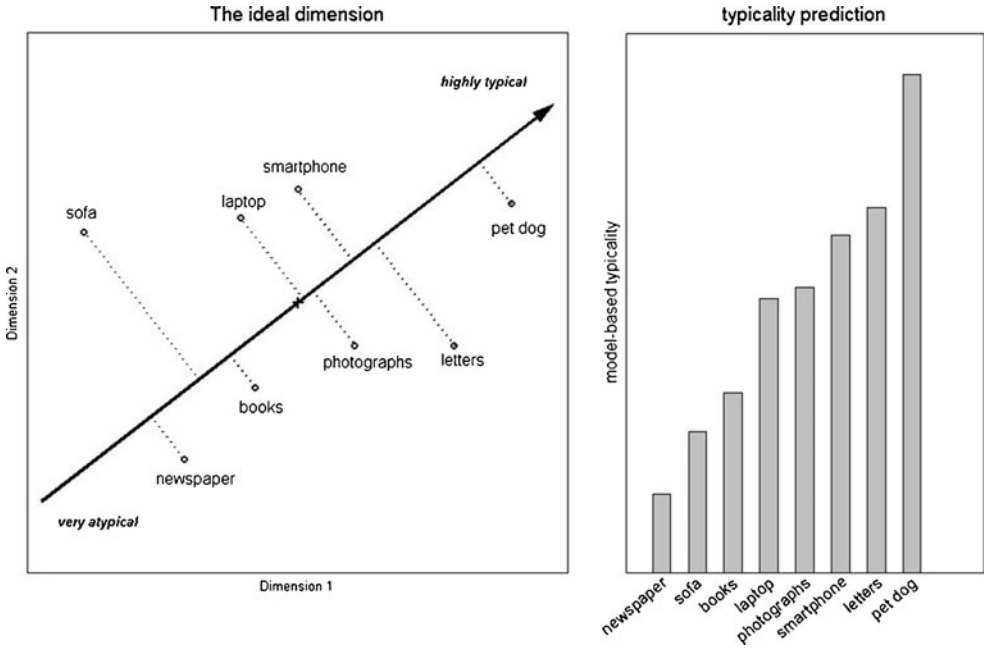

Fig. 3 Illustration of the ideal-dimension model (IDM) in a (fictive) two-dimensional geometric representation of a set of *things you rescue from a burning house*, along with the typicality prediction according to the IDM

space; in an M-dimensional space, $M - 1$ angular coordinates are needed). This results in a uniform prior over all points at a fixed distance from the origin. Together with the origin, these points determine all possible ideal dimensions that produce a unique typicality prediction.

Note that while in Eq. 4 the origin is used as a second point to define the ideal dimension, this choice is arbitrary. What matters is the slope parameter derived from the relative positions of x_A and the arbitrary point (in this case, the origin). All possible slope parameters can be defined, no matter the exact location of the arbitrary point. As a consequence, whatever the choice of the fixed point, the sets of possible unique typicality predictions that one can derive from the IDM are identical.

Appendix B: Model evaluation

The model evaluation analyses took two different, but complementary, perspectives. We investigated whether the models could adequately account for the typicality gradient. This was done by evaluating the models on their goodness-of-fit. The second perspective concerned the comparison of the models. Model comparisons should rely on a measure of generalizability (Pitt et al., 2003) that balances the inherent complexity in the models with their ability to fit the empirical data. We relied on the marginal likelihood as a measure of generalizability. In what follows, we will describe both procedures.

Goodness-of-fit

We first tested whether the models could account for the typicality gradients of the categories by evaluating the optimal data fits of the models. For each category, we optimized Pearson's product–moment correlation coefficient between the model-based typicality predictions for each exemplar of the category and the observed typicality ratings of the exemplars:

$$r_{t\hat{t}} = \frac{\sum_{i=1}^{N} (t_i - \bar{t})\left(\hat{t}_i - \bar{\hat{t}}\right)}{N s_t s_{\hat{t}}},$$

where N is the number of exemplars in a category, t_i is the observed typicality rating for exemplar i, and $\hat{t}_i$ is the model-based prediction of the typicality of exemplar i. The values s_t and $s_{\hat{t}}$ are the standard deviations (across exemplars) of, respectively, the observed typicality ratings and the model-based typicality predictions.

Marginal likelihood

In order to compare the models in their accounts of the observed typicality ratings, we relied on the generalizability of the models as measured through marginal likelihood. The marginal likelihood considers the probability of the data under all possible parameter values of the model.

We assume that each of the observed typicality ratings t_i comes from a Gaussian distribution with variance s_j^2. A probabilistic measure of fit for a model M and a particular set of parameter values, θ, is then given by

$$p(t|M,\theta) = \prod_{i}^{N} \frac{1}{s_i\sqrt{2\pi}} \exp\left(-\frac{\left(t_i - \hat{t}_i\right)^2}{2s_i^2}\right),$$

where N is the number of exemplars in a category and $\hat{t}_i$ is the model-based typicality for exemplar i. The value s_i measures the precision of the data for exemplar i, estimated by the standard deviation across participant judgments for exemplar i (Lee, 2004).

The marginal likelihood considers all possible parameter values of the model. This is done by integrating the probability of the data for all possible parameter combinations, weighted by the prior probability of the parameter combinations $p(\theta \mid M)$:

$$p(t|M) = \int p(t|M,\theta)p(\theta|M)d\theta.$$

A useful measure when considering a set of models of particular interest is the model weight for a model (Lee, 2004),

$$w_l = \frac{P(t|M_l)}{\sum_j P(t|M_j)},$$

reflecting the posterior probability of model M relative to the set of competing models. We used a straightforward numerical method to approximate the integral that defines the marginal likelihood. By sampling a large number of parameter values according to their prior distributions, the marginal likelihood can be estimated by

$$p(t|M) \approx \frac{1}{R}\sum_{r=1}^{R} p(t|M,\theta_r).$$

References

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language, 55,* 402–421. doi:10.1016/j.jml.2006.05.005

Ashby, F. G., & Valentin, V. V. (2005). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 547–572). New York, NY: Elsevier.

Atran, S. (1999). Itzay Maya folkbiological taxonomy. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 119–204). Cambridge, MA: MIT Press.

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(4), 629–654.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 1–64). San Diego, CA: Academic press.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7,* 84–91.

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5–4 categories and the category advantage. *Memory & Cognition, 31,* 1293–1301. doi:10.3758/BF03195812

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology, 58,* 381–396.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical, information-theoretic approach*. New York, NY: Springer.

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

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 69–78. doi:10.1002/wcs.26

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 13–36). Cambridge, MA: Cambridge University Press.

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

Hampton, J. A. (1993). Prototype models of concept representation. In I. Van Mechelen, J. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67–95). New York, NY: Academic Press.

Jang, Y., Wixted, J. T., & Huber, D. E. (2011). The diagnosticity of individual data for model selection: Comparing signal-detection models of recognition memory. *Psychonomic Bulletin & Review, 18,* 751–757. doi:10.3758/s13423-011-0096-7

Jaynes, E. T. (2003). In G. L. Bretthorst (Ed.), *Probability theory: The logic of science*. Cambridge, U.K: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Keller, D., & Kellas, G. (1978). Typicality as a dimension of encoding. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 78–85.

Komatsu, L. K. (1992). Recent views on conceptual structure. *Psychological Bulletin, 112,* 500–526.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognition, 4,* 195–208.

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology, 48,* 310–321. doi:10.1016/j.jmp.2004.06.002

Loken, B., & Ward, J. (1990). Alternative approaches to understanding the determinants of typicality. *Journal of Consumer Research, 17,* 111–126.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337.

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

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology, 29,* 85–148.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32,* 517–535.

Medin, D. L. (1989). Concepts and conceptual structure. *The American Psychologist, 44,* 1469–1481.

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

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology, 51,* 121–147.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 775–799.

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 40–64). Cambridge, U.K: Cambridge University 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. doi:10.1037/0033-295X.92.3.289

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. doi:10.1006/jmps.1999.1283

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology. General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

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

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes: Vol. 1. From learning theory to connectionist theory* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology, 54A,* 197–235.

Pinker, S. (1997). *How the mind works*. New York, NY: Norton.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10,* 29–44. doi:10.3758/BF03196467

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews, 32,* 197–205.

Quine, W. V. (1969). *Ontological relativity and other essays*. New York, NY: Columbia University Press.

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

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 1, pp. 1–49). London, U.K.: Academic Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605. doi:10.1016/0010-0285(75)90024-9

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439. doi:10.1016/0010-0285(76)90013-X

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323. doi:10.1126/science.3629243

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition, 65,* 87–101.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, L. B. (1989). From global similarity to kinds of similarity: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 146–179). Cambridge, U.K.: Cambridge University Press.

Solomon, K. O., Medin, D. L., & Lynch, E. L. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences, 3,* 99–105.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352. doi:10.1037/0033-295X.84.4.327

Van der Kloot, W. A., & van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research, 26,* 563–581.

Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods, 42,* 421–437. doi:10.3758/BRM.42.2.421

Voorspoels, W., Storms, G., & Vanpaemel, W. (2011a). Representation at different levels in a conceptual hierarchy. *Acta Psychologica, 138,* 11–18. doi:10.1016/j.actpsy.2011.04.007

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality based evaluation. *Psychonomic Bulletin & Review, 15,* 630–637. doi:10.3758/PBR.15.3.630

Voorspoels, W., Vanpaemel, W., & Storms, G. (2011b). A formal ideal-based account of typicality. *Psychonomic Bulletin & Review, 18,* 1006–1114. doi:10.3758/s13423-011-0122-9