Typicality and graded membership in dimensional adjectives

Steven Verheyen[a,b*] (steven.verheyen@ens.fr)

Paul Egré[a*] (paul.egre@ens.fr)

[a] Institut Jean Nicod, Département d'études cognitives [Department of Cognitive Studies], ENS, EHESS, PSL Research University, CNRS, Paris France.

[b] Laboratoire de Sciences Cognitives et Psycholinguistique [Laboratory of Cognitive Science and Psycholinguistics], Département d'études cognitives [Department of Cognitive Studies], ENS, EHESS, PSL Research University, CNRS, Paris France.

* Address for correspondence: Steven Verheyen & Paul Egré, Pavillon Jardin, 29 rue d'Ulm, 75005 Paris, France.

Keywords: conceptual spaces; gradable adjectives; antonyms; comparison class; prototypes; categorization; subjectivity; thresholds.

Word count: 16537

Abstract

This paper concerns an investigation of the manner in which typicality constrains graded membership in antonymous dimensional adjectives such as *short*/*tall* and *cheap*/*expensive* using the conceptual space framework. In this framework, items are organized in a space comprised of one or more dimensions along which they can be compared. The items' graded membership is established by their relative proximity in this space to the prototypical instances of contrasting concepts. Because dimensional adjectives can be applied to an indefinite variety of things and grammatically have no upper bound to serve as cognitive reference point, they have been argued to lack prototypes. We present the results of an empirical study showing that the conceptual space framework can nevertheless be extended successfully to dimensional adjectives by complementing them with a comparison class argument (such as *short*/*tall* for an adult man and *cheap*/*expensive* for a smartphone), allowing participants to retrieve meaningful prototypical instances, which can be used to establish membership degree. Since dimensional adjectives are subjective, we investigate how the framework can accommodate inter-individual variability in membership degree judgments. We find that the predictions of the framework significantly improve if prototypical instances themselves are assumed to come with a gradient instead of being considered equally typical, thereby providing a more fine-grained account of typicality and furthering the development of the conceptual space framework.

Typicality and graded membership in dimensional adjectives

1. Introduction

Thresholds are at the heart of linguistic, philosophical, and psychological accounts of categorization (Ashby & Gott, 1988; Bartsch & Vennemann, 1972; Egré, 2017; Fara, 2000; Hampton, 2007; Kennedy, 2007; Lassiter & Goodman, 2015; Raffman, 2014; Verheyen, Hampton, & Storms, 2010; Williamson, 1994). These accounts entertain that for a predicate like *tall* to apply to an object, the object needs to surpass a threshold along a relevant underlying dimension, such as height. The centrality of the threshold notion also becomes apparent in the methodology of empirical studies on categorization, where participants are often asked to directly specify the threshold value: "It is true to say that a man is tall if his height is greater than or equal to _____ centimeters." (Bonini, Osherson, Viale, & Williamson, 1999; see also Alxatib & Pelletier, 2011). The prevalence of thresholds in the categorization literature might give the impression that thresholds are the primary means for language users to apply a verbal label to an item. However, it contrasts with the observation that language users seldom spontaneously explicate the necessary threshold value (Dunning & McElwee, 1995; Verheyen, Dewil, & Egré, 2018).

Thresholds might merely play a secondary role and may be derived from other information language users have at their disposal, such as the typical instances of application of a predicate. This insight was voiced by Eleanor Rosch in 1978 when she wrote: "*Another way to achieve separateness and clarity of actually continuous categories is by conceiving of each category in terms of its clear cases rather than its boundaries.*" (Rosch, 1978, p. 35-36). An influential account along these lines is Gärdenfors' conceptual space (CS) framework, in which prototypical values within a continuous metric space determine the border between categories (Gärdenfors, 2000, 2014). Those values are taken to ground our representations: they come first in terms of representation and learning, and they partition conceptual space into regions of points that are closer to a given prototype than to

alternative prototypes, thereby explaining the more or less extended character of categories and their boundaries.

The CS framework has been used to deal with categories that admit degrees of membership (so-called fuzzy categories). Graded membership there is derived by sampling instances from regions consisting of *multiple* prototypical values (Decock & Douven, 2014; Douven, Decock, Dietz, & Egré, 2013). That is, the multiplicity of prototypical points for a concept generates a multiplicity of possible thresholds between categories, namely the points equidistant between prototypical values of those categories. From that multiplicity of thresholds, a notion of degree of membership intermediate between 0 and 1 can easily be defined for an item, as the proportion of thresholds the item surpasses. This CS model has been tested experimentally: Douven, Wenmackers, Jraissati, and Decock (2017) have shown that for color adjectives such as *blue* and *green*, one can find a strong correspondence between the *observed* degree of membership of an item under a color category C (revealed by the proportion of participants placing the item in C), and the *predicted* degree of membership for that item (based on a measure of the partitions of conceptual space that include the item under C; see below for details). Douven (2016) has found the same correspondence for the shape categories *vase* and *bowl* in relation to a stimulus set gradually morphing a vase to a bowl (Douven, 2016; Gärdenfors, 2000; Labov, 1973). What those studies suggest is that prototypical values constrain our verdicts of membership.

However, it remains an open issue whether this account of degree of membership applies to categories in general (i.e., any set of entities that a predicate can refer to). One class of potentially problematic cases concerns relative gradable adjectives, such as *tall*, *heavy*, or *expensive*. For such expressions, the notion of membership degree appears intuitively meaningful (they are, after all, textbook instances of fuzzy categories, see Smith, 2008) but it is disputed whether such expressions have prototypical values. Kamp and Partee (1995:176) point out that relative gradable adjectives seem to lack prototypes for two main reasons. First because of the “indefinite variety of things” to

which they can be applied: there is no single typical value for *tall* that applies to buildings and to people, for example. Secondly because "there is in general no natural upper bound on how *tall* [*heavy*/*expensive*] things can be". Indeed, unlike so-called *absolute* gradable adjectives such as *empty* and *full*, *bent* and *straight*, the meaning of relative adjectives like *tall* or *expensive* is not relative to a maximum or a minimum standard on the relevant scale of comparison. For something empty, for example, there is a natural lower bound on how empty things can be (zero). That zero value may then serve as a prototype driving membership judgments. For *tall*, on the other hand, no such value seems to stand out (Burnett, 2016; Kennedy & McNally, 2005).

Another issue with relative gradable adjectives such as *tall*, *heavy,* and *expensive* is that they show subjectivity (Egré, 2017; Kennedy, 2013; Sæbø, 2009; Verheyen, Dewill, & Egré, 2018). For *tall* and *heavy* as applied to persons, for example, one can observe reliable differences in the membership degree judgments of different participants (Verheyen, Dewill, & Egré, 2018; see also Hersh & Caramazza, 1976 on *small* and *large*). It is unclear whether an account of degrees of membership based on the notion of typicality can accommodate this inter-individual variability. If typical means socially salient, or socially robust, then any individual differences in membership degree would have to originate from another source.

We believe that none of the previous difficulties are insurmountable. Regarding the indefinite variety and unboundedness arguments, ascriptions of tallness, heaviness, and so on, are always made relative to an explicit or implicit *comparison class argument* (Kamp, 1975; Kamp & Partee, 1995; Klein, 1980; Rips & Turnbull, 1980). We picture a tall person and a tall building as being of very different heights. Kamp and Partee acknowledge that even though the adjective *tall* by itself does not have a prototype, an adjective-noun combination like *tall tree* does. However, they deny that the prototype for a combination like *tall tree* determines membership under the concept.[1] Brownell and

[1] Kamp and Partee (1995: 172 *sqq*.) in their typology describe *tall tree* as +P (comes with a prototype) but -PE (the prototype does not determine the extension). Their argument is that one can mistake a category nonmember that fits a prototypical description for a category member. For instance, someone fitting the description of a prototypical grandmother - in terms of age and appearance - may fail to have children who

Caramazza (1978) on the other hand suggest that the semantic description of the noun that is provided as a comparison class constrains the interpretation of gradable adjectives. That is, one's idea about the height of men in general, will supply a normative value for what constitutes a typically tall man. While it is correct that unlike absolute gradable adjectives (such as *empty*, *straight*), relative gradable adjectives do not select a minimum or maximum value on a scale, this does not imply that typical values will vary without limit once a comparison class argument has been specified.[2] Although a tall person and a tall building will be of very different heights, the range of actual persons' heights and building heights is within certain bounds, and some values within these bounds might stand out, for instance because they are more common or more salient. We will therefore investigate whether participants are able to retrieve prototypical values for adjective-noun combinations, and secondly whether these typical values constrain membership.

Regarding the subjectivity argument, previous studies have taken membership degrees to be real at the individual level, but they have assumed that there are no individual differences in the membership degree function, basically because typical instances are assumed to be socially shared (Douven, 2016; Douven et al., 2017). This is partly driven by the assumption that color adjectives and shape categories have focal values, stable across perceivers. But the assumption that what is typical is always socially shared is questionable. In other cases, what one considers typical may be partly subjective, as a result of distinct past and current experiences depending on the individual (Barsalou, 1989, 1993; Rosch, 1999, 2011). Because relative gradable adjectives are prone to individual differences in membership (Hersh & Caramazza, 1976; Verheyen, Dewil, & Egré, 2018), they provide an ideal test case to see whether typicality and membership are related at the individual level.[3] In

---

have children. We think the issue is different, however. That is, the issue is whether typicality can drive membership judgments for cases in which the subject has all the relevant information (rather than is uninformed as to whether certain prototypical features apply).

[2] Brownell and Caramazza (1978) also suggest that *tall* is bounded because of the existence of bordering expressions such as *very tall* and *sort of tall* and their accompanying meanings/values.

[3] Although in principle, such an investigation could also be undertaken for colors and containers. They too show some context-sensitivity (compare red for wine vs. red for cars or flowers; Anishchanka, Speelman, & Geeraerts, 2015; Hansen & Chemla, 2017) and subjectivity (older people regarding glass bottles typical, younger people considering plastic bottles typical; White, Storms, Malt, & Verheyen, 2018).

what follows, we will investigate whether we can obtain reliable indications of inter-individual differences in typicality, and we will test whether those differences induce reliable differences in membership judgments.

Our main aim in this paper is therefore to investigate whether typicality constrains membership degree in relative gradable adjectives, both at the group level, and at the individual level. To do so, we will use the CS framework, for which it was left as an open question whether it can be successfully applied to more abstract categories than shapes and colors (see Douven et al., 2013: 138). Here, we will focus on a subclass of relative gradable adjectives, namely antonymous dimensional adjectives restricted by a comparison class argument such as *short*/*tall* for a man or *cheap*/*expensive* for a smartphone. Antonymous dimensional adjectives are gradable adjectives that refer to the same scale of a given dimension, but that are ordered in opposite directions (Bierwisch, 1989: 88). Since both the underlying dimension and the pair of contrasting categories C and C' are apparent for dimensional adjectives, they constitute the class of gradable adjectives to which the CS framework can be applied most straightforwardly.

Whereas Douven et al. always asked participants to judge perceptual stimuli (color patches, container silhouettes), the use of adjectives that have a standard measurement scale along a single physical dimension (height in cm for *tall*, price in euros for *expensive*, and so on) allows us to probe participants' abstract representations of prototypes ("a man of 185cm", "an 800-euro smartphone"). Indeed, speakers are not only faced with the task of assigning perceived stimuli to verbal categories, but they also need to interpret language in the abstract: online interpretation of language must rely on an abstract level of representation, in particular when we have to interpret occurrences of a predicate in absence of direct perceptual input (compare hearing "John is tall" when John is away, with pointing to John to utter "John is tall''). The use of numerical rather than visual stimuli has the added benefit that there is no hidden multidimensionality: by design there is only one source of information in the stimuli. If we were to rely on visual stimuli instead, additional dimensions might be

unforeseen (e.g., whether a person counts as tall or not might depend not just on height, but also on the person having the right ratio of height to width, see Lang, 1989).

Before we turn to the empirical investigation, we offer more details about the CS framework in the following section.

2. Typicality in the CS framework

In this section, we give an overview of the CS framework and of its development. The section serves two functions. First, it explains the procedure we shall use to model the relationship between typicality and membership judgments in dimensional adjectives. Second, it supplements the framework with a treatment of typicality that we find lacking in previous developments, namely *graded typicality*.

The development of the CS framework has been concerned with the proper way to determine the prototypical values from which the thresholds or category boundaries are derived. The CS framework originated from the early formulations of prototype theory (Hampton, 2007; Osherson & Smith, 1981), in which similarity to the most representative category instance or prototype was thought to determine category membership. Gärdenfors (2000) proposed to represent instances as points in a metric space at distances inversely proportional to their similarity. He argued that categories then naturally arise as the convex regions made up of the instances that are closer to the category's prototype than to any other category's prototype. Together these regions form a so-called Voronoi tessellation in which the category boundaries are comprised of points that are equidistant between prototypes. Panel A of Fig. 1 gives an example in a 2-dimensional space.

Douven, Decock, Dietz, and Égré (2013) proposed to have prototypical *areas* instead of prototypical *points* determine category membership in the CS framework. Their proposal rests on a positive and a negative argument. The positive argument is that concepts need not have unique

prototypes, but several instances may be representative of the category instead (see also De Wilde, Vanoverberghe, Storms, & De Boeck, 2003). The negative argument goes that the category boundaries in a Voronoi tessellation are too thin to provide a satisfactory account of fuzzy categories. A fuzzy category can be characterized by the admission of borderline items (some shades are intermediate between *blue* and *green*), and borderline items can be represented as points equidistant between prototypes. In general, however, these borderline items are not isolated, but tend to be surrounded by other borderline items. Douven et al. (2013) show that overlaying the individual Voronoi tessellations obtained from the various combinations of prototypical points that make up the categories' prototypical areas, results in a so-called collated Voronoi tessellation (Panel B of Fig. 1), which not only carves out categories comprised of instances with a unique closest prototypical area, but also yields thick (rather than thin) boundary regions. Those are comprised of the instances that have no unique closest prototypical area (indeterminate cases).

From a semantic point of view, Voronoi tessellations correspond to the idea of a trichotomy for category membership: an item is either a clear member, a borderline case, or a clear nonmember. This, however, is intuitively too coarse-grained to account for the smooth transition between adjacent categories. In general, it is natural to think of membership as a more fine-grained, possibly even continuous function, in line with the original conception of prototype theory, where a monotonically decreasing relationship between membership and distance admits degrees of membership (Hampton, 2007; Osherson & Smith, 1981).

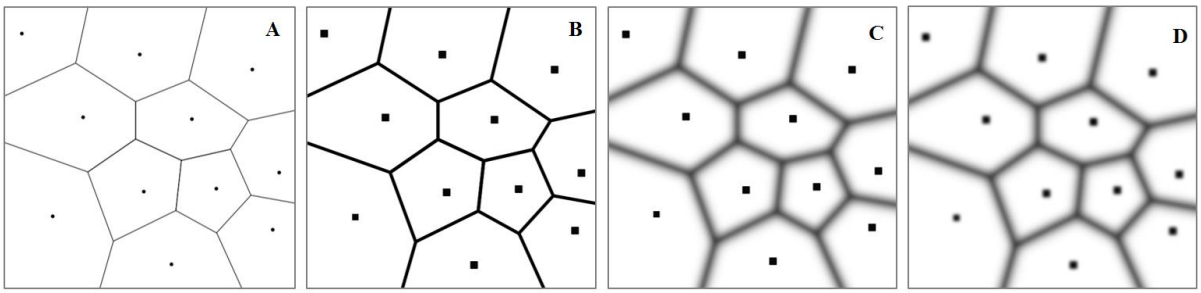

Figure 1. Borderline regions in successive developments of the CS framework. Thin (Panel A), thick (Panel B), and graded (Panel C) category borders arise depending on the assumed nature of the prototypical instances. Panel D represents our proposal to treat the typicality regions themselves as graded.

Decock and Douven (2014) note that having multiple prototypical points allows one to move beyond a classification of instances as members, borderline instances, and nonmembers to a more fine-grained notion of graded membership (possibly continuous). An instance that does not have a unique closest prototypical area, might lie relatively closer to the points from one category's prototypical area than to those from another category. Decock and Douven therefore propose to use the proportion of tessellations that include an item under a category as a measure of the degree of category membership of that item. The boundary regions then become graded (Panel C of Fig. 1). The procedure proposed by Decock and Douven can be considered an application of supervaluationism (Kamp, 1975; Kamp & Partee, 1995; Lewis, 1970) to the CS framework. On the supervaluationist approach, a fuzzy category can be represented by a set of delineations splitting up the indeterminate, borderline cases into members and nonmembers. Relative to that set, an item can be assigned a membership degree, based on the relative number of delineations that include the item under the category (see Kamp 1975; Lewis, 1970; Williams, 2011).

The most recent development of the CS framework concerns its empirical validation (Douven, 2016; Douven, Wenmackers, Jraissati, & Decock, 2017). To predict an item's membership degree, Douven and colleagues elicited typical category instances of two contrasting categories C and C' from participants. As a measure of degree of membership, they used the relative number of times that an item falls closer to some typical instance for C than to some typical instance for a contrast category C' across bisections of the underlying conceptual space resulting from sampling different typical instances of C and C'. In line with the idea that these instances constitute a prototypical area, the instances were assumed to carry *equal weight*, and thus were sampled in a *uniform manner*, without

regard of the relative elicitation frequency of the different instances. That is, in their sampling procedure Douven and colleagues treated the typical instances as *equally typical*. The resulting measure of membership degree corresponded closely to observed membership degree in color and shape categories (see also above).

While the idea of multiple typical instances of a category comes out natural, it is less clear why they should carry equal weight. Indeed, in the literature on concepts and categories, typicality is generally considered a gradient phenomenon (Barsalou, 1985; Hampton, 2007; Rosch, 1975). The gradient could itself have different sources: it might reflect the fact that people believe some instances are more representative of a concept than others (Hampton, 1979; Rosch, 1978); it might be due to the relative availability of some instances over others (Janczura & Nelson, 1999; Löbner, 2002); or it might reflect lack of knowledge, unfamiliarity, or uncertainty about what counts as typical (Lynch, Coley, & Medin, 2000; Malt & Smith 1982). These factors could operate at the individual level as well: two individuals may recognize the same items to be typical, but the first individual may identify a particular item as more typical, while the second individual believes another item is. The ramifications of a typicality gradient for the CS framework deserve investigation, regardless of its source. The work by Douven and colleagues on the empirical validation of the framework (Douven, 2016; Douven, Wenmackers, Jraissati, & Decock, 2017) allows such an investigation to be undertaken by taking the relative frequency with which language users deem instances prototypical into account in the sampling procedure.

In our application of the CS framework to the membership degree of dimensional adjectives, we are interested in whether a model based on the idea that typicality itself could have a gradient could outperform a model in which typicality is uniform. This is depicted in Panel D of Fig. 1 by blurred typicality regions. Though not apparent in Fig. 1, the assumption of gradient typicality makes different predictions regarding the shape of the membership functions, compared to the assumption of uniform typicality.

This can be seen on a representation of the relationship between typicality and graded membership for one-dimensional adjectives like *short* and *tall*. In Fig. 2 (top), the grey rectangles represent the uniform distribution of typical values for *short* and for *tall*. The middle red curve represents the membership degree function generated by sampling typical instances from both regions/distributions. Each sample corresponds to a combination of typical values determining a threshold (the point equidistant between those values). Hence, for each item along the dimension, we get a membership value of 0 or 1, depending on whether the item lies on the side of *short* or on the side of *tall*. By repeating this procedure, the proportion of times an instance is classified as *short* or *tall* corresponds to the predicted degree of membership under either category.

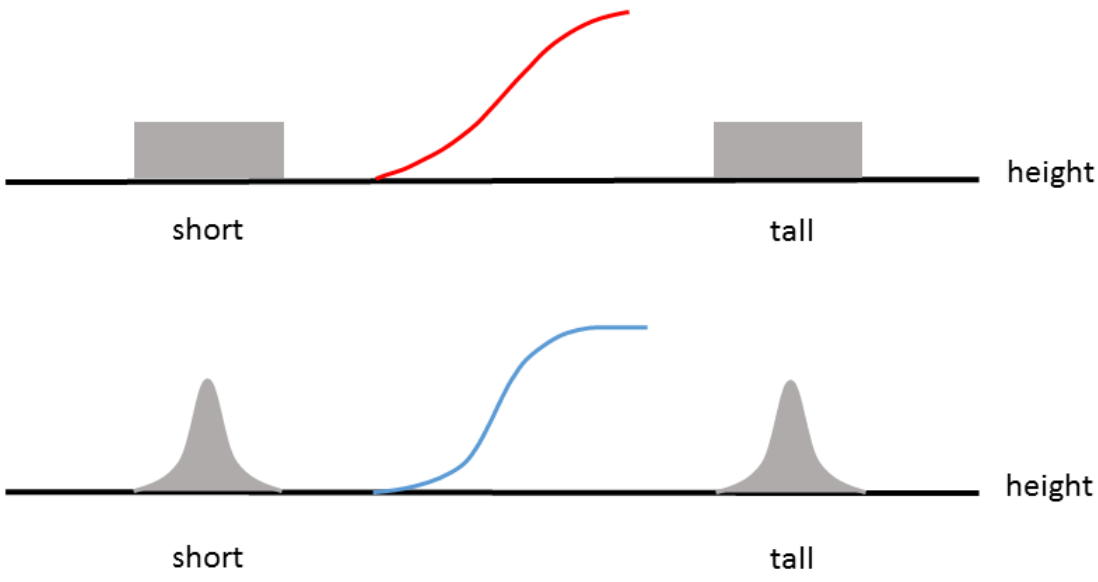

Figure 2. The effect of distributional assumptions regarding typicality on the steepness of the resulting membership degree function. Truncated normal typicality distributions (bottom) yield a steeper membership degree function (blue) than uniform typicality distributions (top) do (red).

If we repeat the above procedure using a truncated normal distribution, bounded by the minimum and maximum values of the uniform distribution instead, this will result in a steeper membership curve, as illustrated in Fig. 2 (bottom). The intuitive reason is that the truncated normal

sampling down-weights extreme typicality values. The further values are located from the sample mean, the less likely to be sampled and to influence the membership curve they become. To see this, imagine the situation where there is only one prototypical value for *tall* and one prototypical value for *short*. Together these values will determine a single threshold, located halfway between the two reference points, resulting in a discrete threshold function. That is, the narrower the prototypical regions that are sampled from, the less extensive the borderline region will be and thus the steeper the resulting membership curve will be. The reverse holds as well. If the sampling extends beyond the bounds of the uniform distribution, the resulting membership curve will be "pulled down" by the more extreme typicality values.

In the next section we will investigate the relationship between typicality and graded membership within the augmented version of the CS account we just described, comparing it with the predictions of the standard account based on the idea of uniform typicality. In Section 3, we present the design and procedure of our study. In section 4 we present the results. In section 5 we give a discussion of the main findings and draw comparisons with those of previous studies.

3. Design and Procedure

To cast light on the issues identified in the previous sections, we designed a study involving eight pairs of dimensional adjectives complemented with a comparison class argument (e.g., *short-tall* for a male adult). Participants judged abstract indications of magnitude ("a male adult of 176 cm"), rather than visually presented stimuli, in terms of prototypicality and category membership. The typicality judgments allow predictions of membership degree to be obtained, which can be compared against participants' actual performance on categorization tasks, as exemplified in the work of Douven and colleagues (Douven, 2016; Douven, Wenmackers, Jraissati, & Decock, 2017). We included tasks necessary to obtain predicted membership degrees both at the aggregate (typicality generation task) and at the individual level (typicality selection task), as well as tasks yielding

observed membership degrees at the aggregate (dichotomous categorization task) and the individual level (trichotomous and continuous categorization tasks). Using this information, we evaluated how the CS framework fares for dimensional adjectives, comparing a CS model that assumes all prototypical instances to be equally typical (uniform CS model) with a CS model that assumes a gradient across prototypical instances (normal CS model).

3.1. Participants

Eighty undergraduate students in psychology participated in exchange for course credit. Their age ranged from 17 to 25 (*M*=18.46; *SD*=.98). Seventy-three participants were female. All participants had Dutch as their first language. Informed consent was obtained from all participants.

3.2. Materials

All materials were in Dutch. Table 1 provides an overview of the English translations. We included eight pairs consisting of antonymous dimensional adjectives (e.g., *short*-*tall*). A comparison class was set for each pair (e.g., male adults). Each pair came with a standard measurement scale along a single physical dimension (e.g., height in cm). We refer to the member for which the associated question is unmarked in English as the *positive* adjective (e.g., *tall*, since "how tall is Mary?" does not presuppose she is tall, unlike "how short is Mary?", which presupposes she is short). The other member (*short*) will be referred to as the *negative* adjective (Bierwish, 1989; Ruytenbeek, Verheyen, & Spector, 2017; von Stechow, 1984).

In this study, the adjectives serve as targets in that participants are either asked to spontaneously generate instances they consider typical of the adjectives (*typicality generation task*),

are presented with instances to select as typical (*typicality selection task*), or must decide whether or to what degree the adjective applies to an item (*categorization tasks*).

The instances participants were presented with in the selection and categorization tasks were indications of the magnitude of instances of the comparison class along the relevant dimension (e.g., an adult male of 176 cm). In the categorization tasks, 21 such instances, equally spaced along the relevant dimension, were presented for each of the adjectival pairs. For *short* and *tall*, for instance, participants would be shown indications of the height of men, ranging from 140 cm to 200 cm in steps of 3 cm (see Table 1). In the typicality selection task, participants saw 20 additional instances, effectively doubling the range of instances used in the categorization tasks. Participants were, for instance, asked to select typical instances of short or tall adult men, from among instances ranging from 110 cm to 230 cm in steps of 3 cm. A narrower range was presented in the categorization tasks since their aim was to measure application of the adjectives in the borderline region. In the typicality selection task, by contrast, participants needed to be able to identify clear instances, generally situated at more extreme ends of the underlying dimension. For the majority of adjectival pairs, the doubling of the range was achieved by adding 10 additional instances to both ends of the narrower range, except for *slow-fast*, *cheap-expensive*, and *thin-thick*, where this procedure would have resulted in meaningless values (e.g., smartphones with negative prices). For these pairs, 20 instances were added at the positive end of the narrow range.[4]

[4]The appropriate ranges and step sizes for each adjectival pair were arrived at through two pilot studies, each with 20 participants drawn from the same student population involved in the main study. None of the pilot candidates participated in the main study. The initial values were determined by looking at the stimulus ranges in related studies and by consulting publicly available sources such as reports on growth curves (for *height* and *weight*) and life expectancy (for *age*) in Flanders, Belgium, where the study was conducted. The adjustments made following the pilot studies ensured that the broad ranges encompassed the values participants would spontaneously generate as typical instances, and the narrow ranges encompassed the borderline region along with a number of clear instances of both the negative and the positive adjectives.

Table 1: Overview of materials.

Negative Adjective	Positive Adjective	Comparison Class	Unit	Step	Min Typ	Min Cat	Max Cat	Max Typ
short	*tall*	male adult	cm	3	110	140	200	230
light	*heavy*	female adult	kg	4	0	40	120	160
young	*old*	adult	years	4	2	18	98	162
low	*high*	ceiling	cm	15	0	150	450	600
cold	*warm*	summer's day	° C	2	-20	0	40	60
slow	*fast*	cyclist	km/h	3	2	2	62	122
cheap	*expensive*	smartphone	€	40	20	20	820	1620
thin	*thick*	book	pages	30	10	10	610	1210

Note: Min and Max stand for the minimum and maximum instance values used in the typicality selection task (Typ) and the categorization tasks (Cat).

3.3. Procedure

The participants completed the study online through the survey software tool Qualtrics (http://www.qualtrics.com). On average, they spent 31 minutes on the study. The resulting data are available on the Open Science Framework (osf.io/djkdg). The study consisted of six tasks. They are described below in the order in which they were presented to the participants. In each of the tasks the eight adjective pairs were presented in a random order to participants. For each of the tasks we provide sample instructions involving the adjective pair *short/tall* and the comparison class male adult.

3.3.1. Typicality generation task

The participants were asked to generate typical values for each of the 16 adjectives. The instructions for *tall* read: "*What height (in cm) comes spontaneously to mind when you imagine a TALL male adult?*". The positive and negative adjective making up a pair were presented on a single screen, with their order counterbalanced across participants.

3.3.2. Typicality selection task

The participants were presented with 41 values for each of the 16 adjectives, from which they were invited to select the typical ones. The range and step size of the values for each of the adjectival pairs can be found in Table 1. For *tall* the instructions read: *"Which of the heights below do you find TYPICAL for a TALL male adult? Do NOT select ALL values you find TALL, only the ones you find most TYPICAL for TALL male adults"*. Heights ranging from 110 cm (Min Typ, Table 1) to 230 cm (Max Typ, Table 1) in steps of 3 cm (Step, Table 1) were then presented simultaneously in increasing order to the participants. The participants were required to select at least two values from the list. As for the typicality generation task, the positive and negative adjective of each pair were presented on a single screen, with their order counterbalanced across participants.

3.3.3. Trichotomous categorization task

For each of the adjective pairs, the participants had to decide for 21 values which of three response options applied best: the negative adjective, the positive adjective, or an option labeled "intermediate". The range of the values can be found in Table 1. It was half the range used in the typicality selection task, as the purpose of the categorization task was to determine the borderline

region, which will not include the values at the low or high end of the scale. For *short/tall* the instructions read: *"The values below represent MALE ADULTS of different heights. Indicate whether you find these men SHORT or TALL. Opt for INTERMEDIATE when you are uncertain about your response"*. Heights ranging from 140 cm (Min Cat, Table 1) to 200 cm (Max Cat, Table 1) in steps of 3 cm (Step, Table 1) were then presented simultaneously in increasing order to the participants.

3.3.4. Continuous categorization task

In this part of the study the participants were presented with the values for which they chose the intermediate response option in the trichotomous categorization task. They were asked to indicate the degree to which they are inclined to apply the negative or the positive adjective to these items. They could accomplish this by positioning a slider along a horizontal scale with 101 equally spaced positions, ranging from the negative (far left position) to the positive adjective (far right position). The instructions for a participant who would have chosen intermediate instead of *short* or *tall* for a male adult of 176 cm, would read: *"How inclined are you to call a male adult of 176 cm SHORT or TALL? Position the slider between SHORT and TALL to indicate your response"*. If for a particular adjectival pair "intermediate" was chosen for more than one value, these values were presented on a single screen in increasing order.

3.3.5. Dichotomous categorization task

In this part of the study the participants were presented with the same 21 values as in the trichotomous categorization task, but were required to apply the negative or the positive adjective to each of the values. They could no longer opt for the intermediate option. For *short/tall* the instructions read for each of the heights: *"Do you find a male adult of X cm SHORT or TALL?"*. This

question was repeated for each of the 21 values, which were presented to participants in a random order on separate screens.

3.3.6. Average and ideal rating task

For each comparison class, participants were asked to generate the value they thought was average (*"According to you what is the AVERAGE height (in cm) of male adults?")* and the value they thought was ideal (*"According to you what is the IDEAL height (in cm) of male adults?")*. The questions for different comparison classes were presented on different screens. Half of the participants always answered the question regarding the average first. The other half always answered the question regarding the ideal first. This task was included for different research purposes.[5]

4. Results

The Results section of this paper is organized in three parts:

In Section 4.1 we establish which values participants consider typical of each of the adjectives. This allows us to determine the typicality regions the CS framework needs in order to predict membership degrees.

In Section 4.2 we compare the average degrees of membership obtained in the continuous and dichotomous categorization tasks with the degrees predicted on the basis of the typicality values by the CS framework. To do so, we sample from the typicality values participants spontaneously generated in the typicality generation task. The standard CS model proposed by Douven and

[5] This task was added to confirm — at the level of individuals instead of at the aggregate level — the finding by Bear and Knobe (2017) that the categorization threshold is predicted by considerations of both average and ideal values. We were able to replicate this finding at the individual level by operationalizing an individual's categorization threshold as the point intermediate between the typicality values s/he generated for the positive and the negative adjective.

associates assumes that typical values all carry equal weight (uniform CS model). We compare it with a model in which typical values can have a gradient (normal CS model).

In Section 4.3 we repeat these analyses at the individual level by making use of the individual-specific data our procedure affords. The typicality selection task provides us with a region of typical values for every participant, while the continuous membership task provides us with a continuous membership curve for every participant. We again compare CS models with different sampling assumptions.

The analyses we present involve eight pairs of adjectives, constituting a multiple comparisons problem. In order to avoid inflating the type I error rate, we employ a more conservative significance level by applying the Bonferroni correction. This means that we only reject null hypotheses when $p < .0063$ for analyses performed at the level of pairs ($\alpha = .05/8$), and when $p < .0031$ for analyses performed at the level of individual adjectives ($\alpha = .05/16$).

4.1. Typicality

4.1.1. Results

Fig. 3 depicts the central tendency and variability of the values the participants spontaneously generated in the typicality generation task (black), and of the mean of the values they selected in the typicality selection task (grey), for the negative and positive members of each of the adjectival pairs. The unit in which the values are expressed is indicated along the vertical axis.

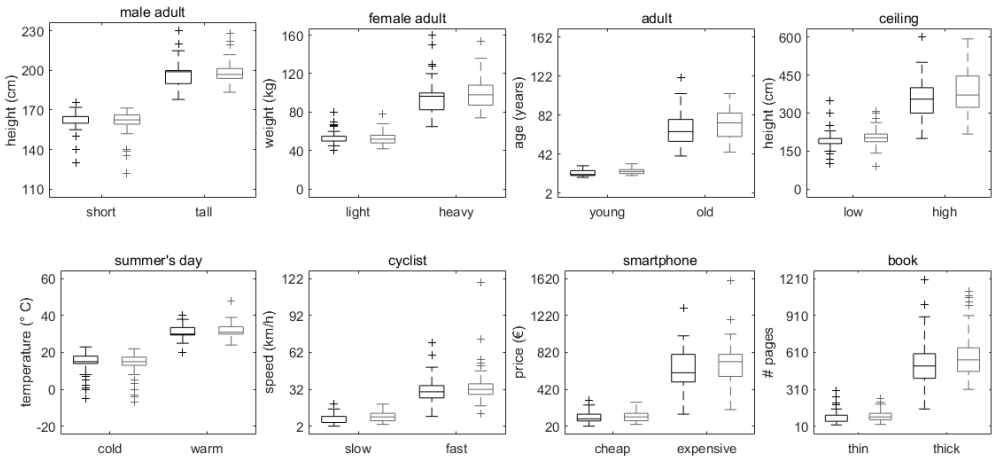

Figure 3. Boxplots of generated (black) and selected (grey) typicality values.

As expected, the typicality values are higher for the positive adjective in each pair than for its negative counterpart. Fig. 3 also indicates that there tends to be more variability at the positive end of the scale than on the negative end of the scale. We confirmed this observation by conducting Pitman-Morgan tests of the equality of the variance of the generated typicality values for negative and positive adjectives. The null hypothesis of equal variances was rejected for all adjective pairs at α=.0063 except for *short/tall* ($t(78)=-2.44$, $p=.017$). For *cold/warm* the variance difference was significant in the opposite direction ($t(78)=3.46$, $p=.0009$).

Table 2

Pitman-Morgan test comparing the variance of generated typicality values for negative and positive adjectives.

Pair	Negative SD	Positive SD	*t*
short/tall	7.59	9.95	-2.44
light/heavy	7.26	22.73	-12.42
young/old	3.53	15.15	-17.93
low/high	38.93	132.62	-15.14
cold/warm	5.42	3.72	3.46
slow/fast	3.87	10.48	-10.60
cheap/expensive	60.83	252.1	-17.41
thin/thick	51.63	207.38	-16.82

Since in the typicality selection task participants provided multiple values (instead of a single value in the typicality generation task), the data from this task allowed us to establish whether this variability difference also holds at the individual level. To do so we compared the standard deviations of the values that individual participants selected for the positive and the negative adjective in a paired samples t-test. Table 3 presents the average standard deviation across participants for each of the adjectival pairs, along with the results of the paired samples t-test. The null hypothesis that the variability for the negative adjective is greater than or equal to the variability for the positive adjective is rejected in all pairs at α=.0063 except for *short*/*tall* (p=.017). For the majority of category pairs the effect size is medium, except for the pairs *short/tall* and *cold/warm* where it is small.[6]

Table 3

Paired samples t-test results comparing individuals' standard deviation of selected typicality values for negative and positive adjectives.

Pair	Negative SD	Positive SD	*t*	Cohen's *d*
short/tall	4.15	4.82	-2.17	-0.24
light/heavy	4.38	8.20	-5.36	-0.60
young/old	3.73	6.46	-6.03	-0.67
low/high	17.19	28.57	-5.67	-0.63
cold/warm	2.36	2.81	-4.22	-0.47
slow/fast	2.74	3.86	-6.11	-0.68
cheap/expensive	38.25	92.13	-9.26	-1.04
thin/thick	35.99	84.65	-7.00	-0.78

Note. All tests, hypothesis is standard deviation for negative adjective less than standard deviation for positive adjective.

The across-participant correlation of the generated and mean selected typicality value ranges from .63 (*thin*) to .84 (*expensive*) with a mean of .73 (all $p < .0031$, right-tailed). To establish whether the individual differences in typicality generation/selection are substantial, we compared each

[6] A similar result is obtained when, instead of the standard deviation of individuals' selected typicality values, the range of these values is compared for negative and positive values. The null hypothesis that the range for the negative adjective is greater than or equal to the range for the positive adjective is rejected in all pairs at α=.0063, except for *short/tall* where p=.01.

observed across-participant correlation with a reference distribution of correlations resulting from the assumption that there is nothing idiosyncratic about the values, but instead that participants share a common typicality distribution from which they sample. To obtain these reference distributions we conducted a randomization test (Edgington & Onghena, 2007): we repeatedly shuffled the generated and mean selected typicality values and each time calculated the resulting correlation across participants. The observed correlation was always higher than the maximum correlation obtained in 10,000 of these randomizations.[7]

4.1.2. Discussion

The results obtained on the generation and selection of typical values show two main findings: the first is some variability between participants in what to count as typical, and the second is a higher variability for the positive antonym than for the negative antonym.

Between-participant variability in the choice of typical values is evidenced by the standard deviations in Tables 2 and 3, and by the boxplots in Fig. 3 showing the typicality distributions for each adjectival pair. Participants appear to have different ideas about what constitute typical heights of, for instance, *tall* men. These inter-individual differences appear substantial in that participants who produce high values in the generation task also tend to choose high values in the selection task (and vice versa). This is evidenced by the significant correlations between the generated and mean selected typicality values, and the results of the randomization tests, which ascribed these correlations to idiosyncratic, not shared ideas about what is typical.

The observed variability in typicality accords with the observation in the literature on nominal concepts that whereas mean typicality ratings tend to be quite reliable, meaningful

[7] The same holds when the typicality values are not reshuffled, but repeatedly drawn with replacement from the observed values, or when values are used that are repeatedly sampled from a normal distribution based on the typicality sample mean and standard deviation.

differences in what is considered to be typical exist between individuals (Barsalou, 1987, 1989; Hampton & Passanisi, 2016). The reasons include differences in knowledge, accessibility, and personal experience (Barsalou, 1989, 1993; Rosch, 1999, 2011). The existence of substantial inter-individual typicality differences that can to a considerable degree be replicated across different tasks, is an additional incentive to investigate the performance of the CS framework at the individual level (see Section 4.3). In what follows, we remain true to our original plan of using the generated values for the aggregate analyses (Section 4.2) and the selected values for the individual analyses (Section 4.3). As we will report, none of our findings are dependent on the use of a particular type of typicality values.

The observation that in antonym pairs typicality varies more for the positive member than the negative member is as far as we can tell a new one. Being unmarked, positives adjectives might have broader semantic meanings than negative adjectives do in that the latter conjure up a subsection of the underlying scale, whereas the positive adjectives can be associated with the entire scale (at least in some linguistic environments, such as questions; "How short is Mary?" asks one to consider low heights, whereas "How tall is Mary?" does not, see Ruytenbeek, Verheyen, & Spector, 2017).[8] Furthermore, all the adjectives we consider are associated with ratio scales in which the zero point is a true zero, with the exception of the pair *warm*/*cold* where participants were given only an interval scale, namely the Celsius scale (which admits of negative values). We hypothesize that because the negative antonym always selects a region closer to the zero point on those ratio scales, there necessarily is a lower bound on the choice of those values. For the positive form of relative adjectives, on the other hand, the scale is not upper-bounded. It is therefore natural to expect more variance there.

We cannot explain all results in that way, however. Even though the degree 0 on the Celsius scale is not a minimum, the negative adjective *cold* selects a region closer to whichever value might

[8] We thank an anonymous reviewer for this suggestion.

count as a physical or psychological zero; yet on that example the variance for the negative antonym was higher than that for the positive antonym in the typicality generation task. World-knowledge about the comparison class, whether implicit or explicit, is likely to play a role in our findings. For ceilings, for example, we may expect a lot more variability for *high* than for *low* because ceilings cannot usually be lower than human size but they can have very different heights depending on the building.[9] Whether the statistical distribution of measurements for the items we consider is symmetric or not could possibly influence whether the positive and the negative form of the antonym will show equal variances in typicality. For *short* vs. *tall* in relation to human heights, heights are normally distributed. Compare this to *short* vs. *tall* in relation to buildings, where we might expect to see a similar pattern as for *low*/*high* ceilings.

The observed variance difference also signals a distinction with the cases that have been previously addressed by Douven and colleagues concerning shape and color (Douven, 2016; Douven et al., 2017), where no such systematic difference was reported. We see two differences there. First of all, Douven's stimulus set is each time finite and bounded on both ends of the spectra he considers. Secondly, for color adjectives in particular, the focal values arguably play a functional role analogous to that of a minimum or a maximum on the corresponding scale of variation. Adjectives like *blue* and *green* are more similar to absolute adjectives like *empty* and *full* in that regard than they are to gradable adjectives like *tall* and *short* (see also Hansen & Chemla, 2017).[10]

[9] While this might suggest that some of the comparison classes could have been specified more (e.g., house ceilings vs. church ceilings; recreational cyclists vs. professional cyclists), this cannot explain the variance difference since it would affect both the positive and the negative adjective.

[10] For absolute adjectives such as *empty* or *full*, one may question whether there is any variance in what counts as typically empty or typically full, and whether the CS framework could be applied. Semantically, such adjectives are standardly assumed to denote a single, context-insensitive value, even though pragmatically, they are used in relation to a variety of values (for example, a glass of water can be called full when the water level is sufficiently close to the top; see Burnett, 2016, and McNally, 2011). We set aside a further discussion of absolute adjectives in this paper.

4.2. Membership degree at the aggregate level

4.2.1. Results

In the continuous categorization task, degree of membership was quantified by awarding clicks on the scale a score in the interval [0,1] proportional to the distance from the negative end of the scale. Clicks on the far-left end of the scale (negative adjective) were thus awarded a score of zero, while clicks on the far-right end of the scale (positive adjective) were awarded a score of 1. The higher the score, the higher the membership degree. The values thus express the degree of membership toward the positive adjective. This is a convention we will use throughout this paper. Membership degree for the negative adjective is then 1 minus the membership degree for the positive adjective (see Douven, 2016, and Douven et al., 2017, for a similar operationalization of membership degree).

In the dichotomous categorization task, degree of membership was operationalized as the proportion of the participants who judged the positive adjective to apply to the item, following the convention we established above (for a similar conception of degrees, see Black, 1937; Borel, 1907; Brownell & Caramazza, 1978; Douven, 2016; Douven et al., 2017; Egré & Barberousse, 2014; Hampton, 1998, 2007; Hersh & Caramazza, 1976). In this case the membership degree for the negative adjective is 1 minus the proportion of participants who judged the positive adjective to apply.

Fig. 4 depicts in black the resulting degrees of membership, which we call *observed* degrees of membership, for each of the eight adjectival pairs. The dotted curves represent the membership degree resulting from the dichotomous categorization task. The solid curves represent the membership degree resulting from the continuous categorization task. The membership degree curves resulting from both tasks have a very similar shape. The correspondence between these curves also shows in a small sum of squared deviations (SSD; second column Table 4), compared to the average SSD of .90 for membership degree curves that pertain to different adjectival pairs. It is

also evidenced by similar points of subjective equality (PSE: the point for which the membership degree equals .50; second and third column of Table 5) and slopes (sixth and seventh column of Table 5).

We calculated degrees of membership as *predicted* by the CS framework by sampling from the typical values generated by the participants for each adjective and its antonym. For example, for an item *x* relative to the adjective *tall*, the predicted degree of membership corresponds to the proportion of times x falls closer to the prototype for *tall* than to the prototype for *short* across sampling. Each sample of a typically positive and a typically negative value establishes a threshold relative to which each item x receives a binary value. Instances that are smaller than the average of the two cognitive reference points are considered examples of the negative adjective and receive a value of 0. Instances that are greater than or equal to this threshold value are considered examples of the positive adjective and receive a value of 1. This procedure is repeated 10,000 times, each time with a new sample of reference points that produce potentially different delineations. The predicted membership curves are the averages across these 10,000 repetitions. Code samples can be found on the Open Science Framework (osf.io/rpyzq).

We discern two sampling schemes. The uniform sampling scheme draws cognitive reference points from the interval bounded by the smallest and the biggest typicality value generated by the participants. This corresponds to the procedure used by Douven and colleagues (Douven, 2016; Douven et al., 2017) and assumes equality of reference points. The normal sampling scheme is also informed by the sample of typicality values generated by the participants but draws cognitive reference points from a normal distribution with mean equal to the sample mean and standard deviation equal to the sample standard deviation.[11] That is, a gradient is assumed across the reference points. In Fig. 4 the predicted membership curves based on uniform sampling are depicted in red. The predicted membership curves based on normal sampling are depicted in blue.

[11] Discrete sampling from the generated values produces similar results, supporting the assumption of normally distributed typicality values.

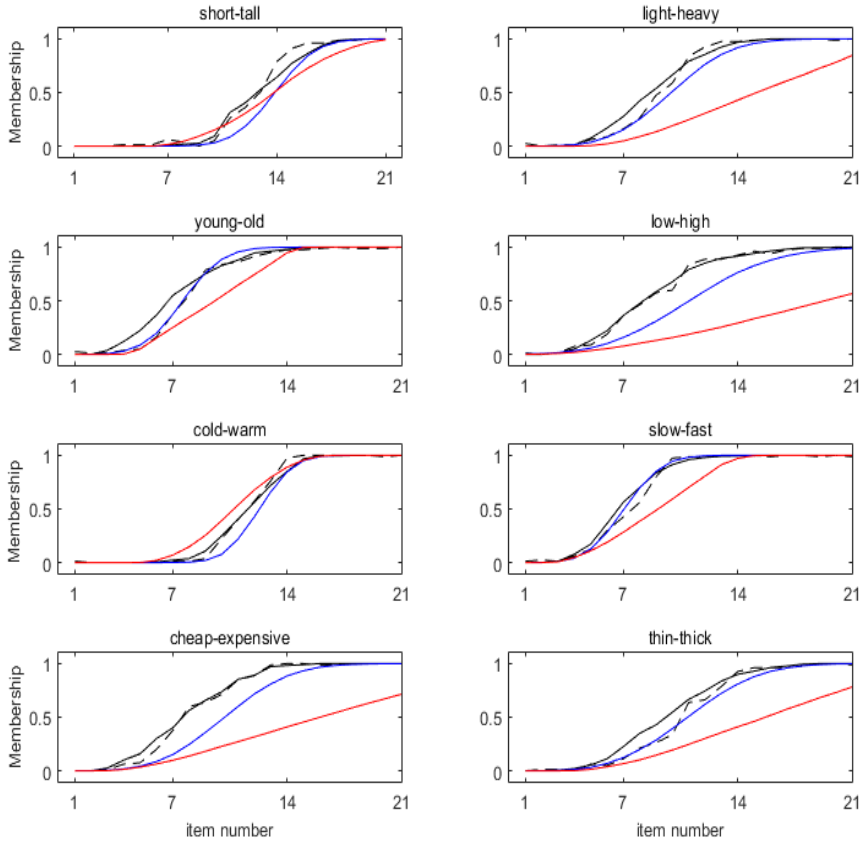

Figure 4. Observed degrees of membership (black) based on the dichotomous (dotted) and the continuous (solid) categorization task. Predicted degrees of membership based on normal (blue) and uniform (red) sampling from the generated typical values.

Fig. 4 indicates that the normal predictions correspond better to the observed membership degrees than the uniform predictions do. The uniform predictions are consistently flatter than the normal predictions and the observed degree curves are. This is also reflected in the SSDs between the observed and the predicted membership degrees (Table 4). For dichotomous categorization, the SSD for normal sampling is lower than the SSD for uniform sampling in all eight adjectival pairs. The same holds for continuous categorization, except for the pairs *short/tall* and *cold/warm*. The magnitude of the SSDs for uniform sampling compares less well than the SSDs for normal sampling to

the SSDs between the observed membership curves (second column Table 4), which provide a natural benchmark for assessing the quality of the fit. Both in Fig. 4 and in Table 4 the adjectives *low/high* and *cheap/expensive* stand out as two pairs for which the CS framework fares less well.

Table 4

Sums of squared deviations (SSDs) *between the observed membership degrees (continuous vs. dichotomous categorization) and between the observed and the predicted membership degrees at the aggregate level. Predictions result from normal or uniform sampling of generated typicality values and are made with respect to either the dichotomous or continuous categorization data.*

Pair	Categorization data	Dichotomous categorization		Continuous categorization	
		Normal	Uniform	Normal	Uniform
short/tall	0.06	0.23	0.28	0.17	0.13
light/heavy	0.07	0.10	2.58	0.14	2.64
young/old	0.12	0.03	0.35	0.11	0.56
low/high	0.01	0.52	4.29	0.53	4.32
cold/warm	0.03	0.11	0.14	0.10	0.09
slow/fast	0.05	0.03	0.43	0.02	0.54
cheap/expensive	0.02	0.47	3.38	0.52	3.45
thin/thick	0.12	0.05	1.72	0.15	2.08

We fitted logistic curves to the observed and to the predicted degrees of membership for each pair of adjectives in order to approximately determine the Point of Subjective Equality (PSE) and the Slope of the membership curves (see also Hampton & Williams, 2016; Douven et al., 2017). The PSE is the numerical value *x* for which the membership degree equals .50. From the formula of the logistic function in Equation (1) it follows that its PSE is x = -a/b. The slope of this logistic function at the PSE is b/4. The higher the slope, the steeper the membership curve is.

$$M(x) = \frac{1}{1+e^{-(a+bx)}} \quad (1)$$

The approximated PSEs and Slopes are presented in Table 5. For convenience and comparability, the results are not expressed in the original units, but assume the units 1:21.

The results in Table 5 confirm the above observations. The approximated slopes (eighth and ninth column) indicate that the uniform predictions are consistently flatter than the normal predictions are. As such, the slopes of the normal predictions correspond better to the approximated slopes of the observed membership curves than the slopes of the uniform predictions do. The approximated PSE is higher for both predictions, but this tendency is more pronounced for uniform predictions.

Table 5

Approximated Point of Subjective Equality (PSE) and Slope for the observed and the predicted membership degrees.

Pair	PSE				Slope			
	Observed		Predicted		Observed		Predicted	
	Dichotomous	Continuous	Normal	Uniform	Dichotomous	Continuous	Normal	Uniform
short/tall	12.59	12.87	13.91	13.83	.20	.18	.22	.13
light/heavy	9.12	8.81	9.97	15.37	.19	.17	.16	.08
young/old	8.00	7.19	7.62	9.53	.17	.16	.24	.15
low/high	8.71	8.63	11.13	18.75	.13	.13	.11	.05
cold/warm	11.51	11.58	12.30	10.77	.26	.21	.28	.18
slow/fast	7.27	6.94	7.07	8.99	.21	.21	.25	.15
cheap/expensive	8.10	7.85	10.21	16.18	.17	.15	.14	.06
thin/thick	10.59	9.72	10.99	15.82	.15	.14	.13	.07

4.2.2. Discussion

Both dichotomous and continuous categorization tasks have been used to obtain aggregate membership degree curves. We found a very good correspondence between the membership curves resulting from the two tasks. The correspondence is evidenced by the similar visual appearance of the curves and the small SSDs between them. Since no one task is arguably the better task for establishing aggregate membership degree, we consider them both to be decent approximations of the "true" membership degree. In that sense, the extent to which these two measures of the same underlying construct differ, provides a natural benchmark against which to assess the absolute fit of the CS predictions. The SSDs between the observations and the predictions of the CS model with normal sampling of reference points, indicate a decent fit provided we leave out the two pairs the CS framework poorly accounts for (we defer a discussion of why the CS framework fares less well for *high/low* and *cheap/expensive* to the General Discussion). The average SSD is .09 for dichotomous categorization and .12 for continuous categorization, compared to an average SSD of .08 between dichotomous and continuous categorization.

The predictions of the CS framework with normal sampling outperform the predictions of the CS framework with uniform sampling, both in terms of SSD, PSE, and Slope. The uniform membership functions tend to fall out too flat in comparison with the observed and normal membership curves because extreme typicality values are weighted more heavily in the uniform sampling scheme. It is an intrinsic property of the CS framework that the broader the prototypical regions that are sampled from, the more extensive the borderline region will be, and thus the flatter the resulting membership curve will be too (see section 2).

In the previous section, we established that there is more variability among the typicality values for the positive adjective in antonymous pairs. This is taken into account in the CS framework in that the sampling (be it uniform or normal) occurs from a broader region for the positive than for the negative adjective in a pair. The greater spread around the mean of the positive adjective is,

however, not a requirement to produce the above results. When one uses the pooled variance instead of the separate variances to sample from the positive and negative distribution, a similar result is obtained. This is also the case when the variances from the positive and negative distribution are reversed. In the CS framework, the uncertainty around the prototypes is accumulated to determine the overall shape of the membership curve. One can observe this additive property in our findings: The order of the approximated slopes of the predicted degree functions corresponds perfectly to the order of the variances pooled across the positive and negative generated instances (after they were normalized). This provides us with a means of assessing the specificity of the predictions produced by the CS framework: the pooled typicality variance should be a good predictor of the slopes of the observed membership curves.

This proves to be the case for the eight adjectival pairs in our study. Across the eight pairs, the Spearman rank correlation between the pooled typicality variance and the approximated slope of the dichotomous membership curve measures -.86 (p=.0107). For the continuous membership curves this correlation measures -.93 (p=.0022). This finding is all the more important in light of the high stability of the slope rank order across categorization tasks (ρ=.98, p=.0004), signaling that slope is a stable and distinguishing characteristic of the membership curves for the antonymous gradable adjectives in our study.

4.3. Membership degree at the individual level

4.3.1. Results

So far the CS framework has only been evaluated at the aggregate level. However, we introduced the continuous categorization task and the typicality selection task in order to evaluate the framework at the individual level as well. The former task provides us with a continuous membership curve for

every individual. Our goal is to predict it with the CS framework using the values from the typicality selection task.

Here too, we compare a CS model that assumes normal sampling from the selected typicality values with a CS model that assumes uniform sampling from the selected typicality values. For the normal sampling, we set the mean and standard deviation respectively equal to the mean and standard deviation of the values selected by a participant.[12] For the uniform sampling we employed the minimum and maximum values selected by participants as interval bounds. Except for the fact that analyses are performed at the individual level instead of the aggregate level, the employed procedures are the same as the ones described in the previous section.

For each participant, we computed the SSD between the observed continuous membership curve and the predicted membership curves resulting from the normal and uniform CS models. The third and fourth column of Table 6 list the median SSD across participants for each of the adjective pairs. According to paired samples t-tests the SSD for the normal CS model is significantly lower than the SSD for the uniform CS model at α=.0063 for all adjective pairs. For the majority of category pairs the effect size is very large, except for the pairs *light/heavy*, *cheap/expensive*, and *thin/thick* where it is large (Table 6, column 6).[13]

[12] We also considered an alternative model in which the mean was set to the value the participant spontaneously generated in the typicality generation task and the standard deviation was equated to the standard deviation of the values the participant selected in the typicality selection task. This led to comparable results. We also mimicked both normal sampling procedures by employing a binomial distribution with the number of trials set to the range of the selected values and the expected mean to either the center of the selected typicality range or the generated value, and adjusting the range of the sampled values accordingly. The results of these binomial sampling procedures are comparable to the results of the normal sampling procedures.

[13] We identified participants who might have changed strategy in the course of the experiment or who might just not have been putting in any effort, by calculating the SSD between individuals' dichotomous and continuous empirical membership curves. We then discarded the data from the 10% worst participants. This did not affect the finding that the normal CS model outperformed the uniform CS model.

Table 6

Median sums of squared deviations (standard deviation) between the observed membership degrees (continuous vs. dichotomous categorization data) *and between the observed (continuous) and the predicted membership degrees at the individual level. Paired t-test results comparing the SSD of normal and uniform CS models' predictions of individual continuous membership curves.*

Pair	Categorization data	Individual predictions			
		Normal	Uniform	t	Cohen's d
short/tall	1.02 (0.80)	0.70 (1.14)	0.83 (1.16)	-14.67	-1.64
light/heavy	0.99 (0.67)	0.71 (0.98)	0.78 (1.00)	-7.41	-0.83
young/old	1.07 (1.14)	0.84 (1.18)	0.94 (1.22)	-10.92	-1.22
low/high	1.21 (1.45)	0.99 (1.84)	1.14 (1.87)	-10.94	-1.22
cold/warm	0.72 (0.52)	0.49 (0.59)	0.63 (0.62)	-12.45	-1.39
slow/fast	0.85 (0.74)	0.38 (1.28)	0.49 (1.30)	-12.25	-1.37
cheap/expensive	0.92 (0.81)	1.58 (1.75)	1.79 (1.78)	-8.48	-0.95
thin/thick	1.32 (1.06)	0.84 (2.00)	0.99 (2.12)	-7.95	-0.89

To obtain an indication of the absolute fit of the normal CS model to the individual continuous membership curves, we computed the SSD between the dichotomous and the continuous membership curve for every participant. Both curves are intended to measure the same construct but clearly an individual's dichotomous curve only provides a coarse indication of membership degree. As such, it gives a baseline against which to assess the fit of the normal CS model's predictions. The second column of Table 6 includes the median SSD between individuals' dichotomous and continuous membership curves across all participants. While the median SSD for the normal CS model (column 3) is lower than the median SSD for the dichotomous curve (column 2) in all adjective pairs except *cheap/expensive*, the difference is not significant according to paired samples t-tests (not shown). These results indicate that the normal CS model yields a rather coarse indication of membership degree at the individual level. For *cheap/expensive*, the model's absolute

fit is not good: The SSD between the continuous and normal CS membership degree is significantly different from the SSD between the observed membership curves (t=3.36; p=.001; Cohen's d=0.38).

We fitted logistic curves to each of the individual observed and predicted degrees of membership to obtain an approximation of the Point of Subjective Equality (PSE) and the Slope of these membership curves. This was done in the same manner used for the aggregate curves in section 4.2.1. Table 7 lists the median and standard deviation of the approximated PSEs and Slopes for each of the adjective pairs.[14] These results too indicate that the absolute fit of the predicted membership curves can be improved. In terms of approximated PSE, the normal and the uniform CS model overshoot the values obtained for the observed membership curves to the same extent. Paired samples t-tests (not shown) indicate that there is no significant difference in the absolute deviation from the approximated PSEs for the observed curves of the approximated PSEs resulting from the normal CS model and the uniform CS model for any of the adjective pairs. The absolute deviation from the approximated Slopes for the observed curves, on the other hand, is significantly greater for the uniform CS predictions than for the normal CS predictions for each of the adjective pairs, according to paired samples t-tests (not shown). Both models predict steeper membership curves than are observed, though.

[14] The results of one participant for *slow/fast* and one participant for *cheap/expensive* are not included because their uniform CS model prediction took the form of a threshold curve for which the slope is not defined.

Table 7

Median (SD) approximated PSE and Slope across individual continuous (observed) and predicted (normal CS, uniform CS) membership degrees.

Pair	PSE			Slope		
	Observed	Normal	Uniform	Observed	Normal	Uniform
short/tall	12.78 (1.36)	14.38 (1.77)	14.36 (1.78)	.22 (1.63)	.50 (.21)	.74 (2.98)
light/heavy	8.88 (1.69)	10.00 (2.17)	10.00 (2.15)	.25 (1.38)	.46 (.24)	.71 (3.39)
young/old	7.12 (1.71)	8.99 (1.83)	8.98 (1.83)	.24 (1.35)	.56 (.24)	.87 (3.63)
low/high	8.43 (2.50)	10.11 (3.53)	10.13 (3.53)	.23 (2.06)	.50 (.25)	.81 (3.69)
cold/warm	11.55 (1.35)	12.74 (1.38)	12.76 (1.39)	.31 (2.15)	.56 (.20)	.88 (3.18)
slow/fast	6.89 (1.67)	7.25 (1.88)	7.25 (1.88)	.33 (1.20)	.62 (.21)	.99 (3.57)
cheap/expensive	7.74 (2.34)	11.00 (2.90)	11.00 (2.91)	.26 (2.01)	.38 (.24)	.63 (3.19)
thin/thick	9.63 (2.16)	11.49 (3.32)	11.50 (3.34)	.20 (1.90)	.39 (.25)	.57 (3.22)

While the results in Table 7 indicate that the CS models overestimate the PSE, the models do manage to capture some of the inter-individual differences in PSE. The values in the second and third column of Table 8, which represent the correlations between the approximated PSEs for the observed and the predicted curves are all significant at α=.0063, except for *short/tall* (p =.03) and *young/old* (p=.01). None of the correlations between the approximated empirical and predicted Slopes is significant, however, which suggests that the main problem with the model's predictions lies with the variance employed for the typicality regions.

Table 8

Correlation between observed and predicted (normal CS, uniform CS) PSE and Slope (approximated) across individuals.

Pair	PSE		Slope	
	Normal	Uniform	Normal	Uniform
short/tall	0.25	0.25	-0.23	-0.14
light/heavy	0.49	0.49	-0.06	-0.10
young/old	0.29	0.28	-0.13	-0.12
low/high	0.58	0.58	0.03	-0.04
cold/warm	0.48	0.48	0.02	0.07
slow/fast	0.41	0.40	-0.25	-0.20
cheap/expensive	0.49	0.49	-0.05	-0.07
thin/thick	0.41	0.41	-0.07	-.11

As we already explained in section 2 and again in section 4.2.2 for the aggregate data, in the CS model there is a straightforward negative relation between the width of the employed typicality regions, and the slope of the resulting membership curve. The approximated slopes of the predicted individual degree functions correlate -.99 or higher with the variances pooled across the positive and negative selected instances (after they were normalized) except for the pair *slow/fast* where the Spearman rank correlation is -.96 due to one participant whose selected typical values for *fast* were so extreme that it resulted in an almost flat predicted membership curve. In striking contrast to this relationship, the Spearman rank correlations between the pooled variances and the approximated slopes of the observed continuous membership curves varied between .10 for *slow/fast* and -.06 for *cold/warm*.

4.3.2. Discussion

The aggregate-level finding whereby the CS model assuming a normally distributed typicality region outperforms the CS model that assumes a uniformly distributed typicality region is replicated at the individual level. The normal CS model fits the individual continuous membership curves better than

the uniform CS model does. While the two CS models yield similar predictions of the PSEs of the individual membership curves, the normal CS model yields membership curves whose slopes better correspond to the observed membership curves. Contrary to the aggregate results, this is due to the normal CS model broadening (rather than tightening) the range of typical values that are being sampled compared to the uniform CS model, resulting in relatively flatter predicted membership curves.

While the normal CS model fits the empirical data relatively better than the uniform CS model, its correspondence to the observed membership curves is rather poor in an absolute sense. The resulting SSD is comparable to that between the continuous and the dichotomous membership curves, where one can consider the latter a crude rendition of the former. The normal CS model also overestimates the PSE and slope of the membership degree function. The model accounts for 18% of the variance in the approximated empirical PSEs, but does not systematically account for the variance in the approximated empirical slopes.

What can we conclude regarding the relation between typicality and membership from the fact that the CS model fares less well at the individual level than it does at the aggregate level? We believe that the most severe interpretation - that the CS framework does not hold at the individual level and as such is not a viable account of the manner in which individuals arrive at graded membership responses – would entail too harsh a verdict for the CS framework, as our procedure might have failed to produce accurate data at the individual level. The difficulty of the CS framework to account for the observed membership degree at the individual level may be due to an insufficiently precise assessment of the typicality region through the typicality selection task. Since in the CS framework the thresholds and thus the boundary regions are derived from cognitive reference points, precise specification of these typical instances is a requirement for the CS framework to deliver meaningful predictions. It would appear that the procedure we employed made participants indicate a rather restricted range of typical instances, resulting in a predicted membership curve with

too steep a slope when uniform sampling was employed. The normal CS model might then correspond better to the empirical membership curves than the uniform CS model because it allowed for the sampling of typical instances beyond the minimal and maximal values specified by participants in the typicality selection task (much like the normal CS model outperformed the uniform CS model at the aggregate level by providing a natural way of down-weighting outlying generated typicality values; see section 4.2.2). The CS framework's account of individual membership curves could potentially be improved, however, if one could specify the typicality distributions in a more precise way.

In section 2 we indicated that the development of the CS framework has been concerned with the proper way to determine the prototypical values from which the thresholds or category boundaries are derived. The above results bring this further into focus: in the CS framework, the challenge of determining the boundaries (degrees) of antonymous adjectives appears to be replaced by the challenge of determining the boundaries (degrees) of the typical instances of these adjectives. In order to establish where application of one adjective stops and application of another begins, the CS framework requires one to establish the boundaries of the typicality regions of each of the adjectives. Or put differently, to establish membership degree, the CS framework requires one to establish typicality degree first. While this observation does not imply any kind of vicious regress for the framework (degrees of typicality themselves need not derive from the distance to further typical values) it does involve a transition into a different topic of investigation and debate regarding the nature of typicality differences (representativeness, availability, uncertainty, lack of knowledge or familiarity; see Hampton, 1979; Janczura & Nelson, 1999; Löbner, 2002; Lynch, Coley, & Medin, 2000; Malt & Smith, 1982; Rosch, 1978).

5. General Discussion

5.1. Main findings

We set ourselves three goals in this paper. Our first and main goal was to see whether prototypical values constrain verdicts of graded membership for relative gradable adjectives. Making use of the CS framework, this relationship had so far only been confirmed for color adjectives and shape categories (Douven, 2016; Douven et al., 2017). Our answer to this first question is generally positive. For 6 of the 8 pairs of antonymous dimensional adjectives (a subclass of relative gradable adjectives; Bierwisch, 1989) that we tested in combination with a comparison class argument, we saw that observed degrees of membership are well predicted from the typicality values generated across participants (see Section 5.2 for a discussion of the two other pairs).

This confirms that even though relative gradable adjectives can apply to an indefinite variety of things and *grammatically* the positive form of these adjectives has no upper bound to serve as cognitive reference point (Kamp & Partee, 1995), such adjectives do conjure typical values when combined with noun phrases fixing the comparison class (such as *short/tall* for a male adult), as suggested by Brownell and Caramazza (1978). Consequently, what we see is that those prototypical values can indeed determine the extent to which an item is likely to be placed under the category.

Furthermore, whereas so far the studies by Douven and associates always asked participants to categorize *visually perceived* stimuli, we see that the account generalizes to more abstract semantic categories, consisting of adjective-noun combinations for which participants did not see actual exemplars, but were asked to imagine values relative to abstract physical scales (see also section 5.3). This finding is worth highlighting, for Douven and associates (see Douven et al. 2013: 138) left as an open question whether the CS framework could be extended to abstract semantic categories.[15]

[15] This is not to say that this is the final account of these data and that models that assume thresholds rather than prototypes the primary means for establishing graded membership might not account (better) for the

The second goal we had was to investigate the connection between typicality and graded membership at the individual level. Because relative gradable adjectives are subjective (Egré, 2017; Kennedy, 2013; Sæbø, 2009), we assessed the CS framework at the individual level by investigating dimensional adjectives for which individual differences in graded membership have already been established (Hersh & Caramazza, 1976; Verheyen, Dewil, & Egré, 2018). We found considerable individual differences in typicality that replicated over different tasks. Our findings were less successful when it came to relating these typicality differences to the observed individual differences in membership degree. We did capture some of the inter-individual variability in the observed point of subjective equality, but not in that of the slopes. Overall, the fit between observed and predicted curves did not prove as satisfactory as at the aggregate level. This finding should not be taken as conclusive against the CS framework, however, since our method for determining individuals' prototypical values might not have been sufficiently precise. More work is therefore required to clarify the CS framework's ability to account for inter-individual differences in membership degree (more on this below).

Our third goal was concerned with the question of whether equipping prototypical instances with a gradient themselves (Barsalou, 1985; Hampton, 2007; Rosch, 1975) would not yield superior predictions of graded membership. We have undertaken a systematic investigation of that question by comparing the predicted degrees of membership based on uniform sampling of typical instances with those based on normal sampling, and we have observed a superiority of the predictions based on normal sampling throughout our analyses. We find this result satisfactory, because intuitively not all typical items need be equally typical or carry equal weight. In the case of relative gradable adjectives some instances may be considered more representative or are more available than others. More will be said about the choice of normal vs. uniform sampling in the following section.

---

data. An anonymous reviewer suggested that a decision bound model with criterial noise (Ashby & Maddox, 1993) would be a good contender.

5.2. Challenges for the CS framework

The CS framework accounts for the graded membership of fuzzy categories not by looking at the categories' boundaries, but at their typical cases. It posits that the membership degree of an item under a category C relative to a category C' can be calculated as the number of times the item is positioned closer to the prototype of C across bisections of the space between prototypical instances of C and C'. As such the framework puts tremendous emphasis on the specification of these typical instances and at least in its current form suggests typicality is all there is to deriving graded membership. One reading of our results is that the former is not readily achieved. We hypothesized that the unsatisfactory performance of the CS framework at the *individual* level could be due to a misspecification of the individual typicality distributions. The finding that at the *aggregate* level the CS model assuming normal sampling of typical instances outperformed a CS model assuming uniform sampling of typical instances, could also be interpreted along these lines: normal sampling could provide for a natural means of dealing with outliers (rather than constitute evidence for a true typicality gradient). This way, the influence of participants who find it difficult to report sensible abstract values (e.g., due to lack of knowledge) is limited.

In principle, the modeling procedure we used could be abandoned in favor of an alternative one. Instead of fixing the parameters of the hypothesized typicality distributions based on summary statistics from observed behavior (typicality generation or typicality selection) and subsequently deriving membership degrees from these distributions, the free parameter values of the distributions that allow for the best prediction of observed membership degree could first be estimated (yielding potentially better fits than the predictions we have reported). One could then test whether optimal normal distributions account better for observed membership degree than optimal uniform distributions do. Whether these optimal distributions would reflect typicality rather than another source of information would still remain to be shown, however, for instance by relating the optimally predictive distributions to empirically observed typicality data. Since in the current study we wanted to investigate whether typicality constrains membership in dimensional adjectives, we decided to use

participant generated typicality data to fix the parameters of the distributions from which membership degree is derived. This approach is in line with the modeling procedure that is customary in the CS framework (see Douven, 2016; Douven et al., 2017) and has the advantage that predicted membership degree results solely from typicality.

Throughout this paper the results for *low/high* and *cheap/expensive* have been worse than the results for the other six adjectival pairs. An explanation of why this is the case is necessarily *post hoc*, but nevertheless also identifies the sharp specification and sole reliance on typicality distributions as challenges for the CS framework. These explanations of course need to be tested in future research.

In the case of *low/high* the explanation might lie in the multimodality of the distribution of ceiling heights. To the question of what constitutes a low ceiling, the majority of the participants (92.5%) in the typicality generation task responded with a multiple of 0.50 meter. The resulting distribution had two modes: 300 cm (30%) and 400 cm (31%). The bad fit we observed for *low/high* might thus be the result of mis-specifying the typicality distribution as being centered around a single value (normal CS model) or having no mode whatsoever (uniform CS model). Multimodality might apply to an individual's typicality distribution as well, so this explanation is not invalidated by the fact that the CS framework performed badly for *low/high* at both the aggregate *and* the individual level. The multimodality of the distribution of heights that are typical for low ceilings is another illustration of the fact that for the CS framework to be able to provide a good indication of graded membership, one has to first get a good measure of the typicality distribution. As mentioned in the previous paragraph, the empirical challenges of obtaining (particularly individual) typicality distributions, might prompt one to attempt an alternative approach whereby one first estimates the distributions that are most predictive of the observed membership degrees and then validates these distributions.

In the case of *cheap*/*expensive* smartphones, the explanation for the oddity might lie in the greater subjectivity of this particular assessment (Kamp & Sassoon, 2014; Kennedy, 2013). What counts as a *cheap* or an *expensive* smartphone depends on one's income, one's expenses, one's savings, one's smartphone needs, and these factors might even impact the distribution of

smartphone prices one considers. Compare this to some of the other materials in our study, such as *short/tall* men, for which (i) it is intuitively more difficult to come up with idiosyncratic reasons why one might or might not consider a particular man *tall*, and (ii) the height distribution of men one encounters is not directly under one's influence.[16]

For the increased subjectivity argument to hold, however, one would have to entertain the possibility that it influences individual application, but not typicality selection, which would then reflect some kind of normative judgment (Barner & Snedeker, 2008; Bear & Knobe, 2017). For if one were to generate/select subjectively typical values, one would assume to see better fits at the individual level compared to the aggregate level, but *cheap/expensive* was not well accounted for at either level. Existing research on subjectivity suggests that in addition to typicality, other factors also contribute to the application of terms (e.g., egocentric reference, see Verheyen, Dewil, & Egré, 2018; practical interests, see Fara, 2000, 2008; or a representation of what counts as ideal, see Bear & Knobe, 2017). Whether the CS framework should embrace idiosyncratic typicality distributions or rather assume a shared typicality distribution, but allow for additional influences on graded membership deserves further investigation.

5.3. Scaling

We would like to conclude this paper with a note on psychological scaling. Douven and colleagues evaluated the CS framework for shapes (pictures of containers in Douven, 2016) and colors (blue and green in Douven et al., 2017). Both are perceptual concepts that can be represented in a geometric space of low dimensionality. For shapes, such a space was not readily available. Douven therefore

[16] That is not to say that *tall* as applied to people has no subjective meaning whatsoever. See Dunning and Cohen (1992) and Verheyen, Dewil, and Egré (2018) for evidence that the threshold for application of *tall* to other people is influenced by one's own height. The comparison class presumably also affects the extent to which a particular relative gradable adjective shows subjectivity. Due to increased experience and explicit instruction, we have a better mutual understanding of the heights of people than we do of the heights of buildings, for instance.

applied multidimensional scaling (MDS) to human similarity judgments of pairs of containers. MDS positions the different containers in a low dimensional space such that the distances between containers are inversely related to their mean similarity (Borg & Groenen, 2005). The resulting space is a psychological one in that it is based on similarities as perceived by human judges. For colors, such a space was already available. The so-called CIELUV space is composed in such a manner that pairs of color stimuli that human observers tend to perceive as equally different are mapped to pairs of points at equal distance in the space (Malacara, 2002).

This raises the question of whether we were justified to use objective, abstract representations of magnitude in our study instead of psychological ones. The question is one of considerable debate in the literature on number processing (e.g., Brannon, Wusthoff, Gallistel, & Gibbon, 2001; Dehaene, 2001). According to one account, numbers are represented in a linear fashion and the variability of the representation increases with number. In our study, we too assumed a linear representation of number. Hence, the observation that variability of typicality is higher at the positive end of the scale might be seen as a manifestation of the accumulation of error with magnitude. According to the second account, as numbers grow they are increasingly positioned closer to each other with a constant variance.

To investigate the impact of entertaining this type of representation for our findings regarding graded membership at the aggregate level (Section 4.2), we have performed a logarithmic transformation of our numerical stimuli. These results are documented in the Appendix. There are two main findings to take home from the additional analyses. One is that the difference between normal sampling and uniform sampling is more pronounced assuming a linear representation than assuming a logarithmic transformation. Another is that the results are dependent on the adjectival pair under consideration. For some adjectival pairs, a linear representation appears more appropriate, while for other pairs a logarithmic transformation appears to account for the data better, without a clear indication of when which is the case. Multiple factors may play a role in the

exact organization of the mental number line that underlies the participants' judgments such as gender, or familiarity with the unit of measurement (Dehaene & Marques, 2002).

This implies that we might want to obtain a separate underlying representation for every participant. We did not undertake this investigation in the current study, for language users seem to be able to rely on an abstract level of representation (e.g., when they interpret occurrences of a predicate in absence of direct perceptual input), but also for practical reasons, such as the additional requirement of obtaining 8 x 41 x 20 judgments of pairwise similarity from each of the participants. Because of the practical burden, idiosyncratic scaling solutions are generally absent in the literature (for two notable exceptions see Coltheart & Evans, 1981, and Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014). Moreover, when we do undertake the task of determining individual psychological representations for concepts like *tall*, we will likely have to turn to perceptual indications of height as in Solt and Gotzner (2012) or Qing and Franke (2014), because similarity judgments of abstract numerical information do not always favor geometric representations (e.g., Lee, 2002; Navarro & Griffiths, 2008; Tenenbaum, 1996).

Acknowledgments

We thank James Hampton and three anonymous reviewers for helpful comments on an earlier version of this paper. We are very grateful to Leon Bergen, Heather Burnett, Igor Douven, Josh Knobe, Dan Lassiter, Peter Pagin, Sophia Sklaviadis, and Robert Williams for comments and suggestions, as well as to audiences in Paris (Workshop "New Advances in Formal Pragmatics"), Stockholm (5th Graduate Philosophy Conference), Toulouse (ESSLLI 2017), and Munich (ECAP9). We also thank Eva Bollen for help with illustrations. SV was funded by ANR project TriLogMean (ANR-14-CE30-0010) and by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 313610. PE was funded by ANR project TriLogMean (ANR-14-CE30-0010). SV and PE also acknowledge grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL* for research conducted at the Department of Cognitive Studies of ENS in Paris. This paper is dedicated to Anemoon Verheyen.

References

Alxatib, S., & Pelletier, F. J. (2011) The psychology of vagueness: Borderline cases and contradictions. *Mind & Language, 26*, 287-326. *https://doi.org/10.1111/j.1468-0017.2011.01419.x*

Anishchanka, A., Speelman, D., & Geeraerts, D. (2015). Usage-related variation in the referential range of blue in marketing context. *Functions of Language, 22*, 20-43. *https://doi.org/10.1075/fol.22.1.02ani*

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 33-53. *https://doi.org/10.1037/0278-7393.14.1.33*

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372-400. *https://doi.org/10.1006/jmps.1993.1023*

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development, 79*, 594-608. *https://doi.org/10.1111/j.1467-8624.2008.01145.x*

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. *https://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: Cambridge University Press.

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76-121). Cambridge: Cambridge University Press.

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. C. Collins, S. E. Gathercole, & M. A. Conway (Eds.), *Theories of memory* (pp. 29-101). London: Lawrence Erlbaum Associates.

Bartsch, R., & Vennemann, T. (1972). The grammar of relative adjectives and comparison. *Linguistische Berichte, 20*, 19-32.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167,* 25-37. *https://doi.org/10.1016/j.cognition.2016.10.024*

Bierwisch, M. (1989). The semantics of gradation. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives: Grammatical structure and conceptual interpretation* (pp. 71–261). Berlin: Springer-Verlag.

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science, 4*, 427-455. *https://doi.org/10.1086/286476*

Bonini, N., Osherson, D., Viale, R., & Williamson, T. (1999). On the psychology of vague predicates. *Mind & Language, 14*, 377-393. *https://doi.org/10.1111/1468-0017.00117*

Borel, E. (1907). Un paradoxe économique: le sophisme du tas de blé et les vérités statistiques. *La Revue du Mois, 4*, 688–699 (English translation by P. Égré & E. Gray, An economic paradox: The sophism of the heap of wheat and statistical truths. *Erkenntnis, 79*, 1081–1088). *https://doi.org/10.1007/s10670-014-9615-z*

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. New York, NY: Springer.

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science, 12*, 238-243. *https://doi.org/10.1111/1467-9280.00342*

Brownell, H. H., & Caramazza, A. (1978). Categorizing with overlapping categories. *Memory & Cognition, 6*, 481-490. *https://doi.org/10.3758/bf03198235*

Burnett, H. (2016). *Gradability in natural language: Logical and grammatical foundations*. Oxford, UK: Oxford University Press.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences, 111*, 14565-14570. *https://doi.org/10.1073/pnas.1402594111*

Coltheart, V., & Evans, J. St. B. T. (1981). An investigation of semantic memory in individuals. *Memory & Cognition, 9*, 524-532. *https://doi.org/10.3758/bf03202346*

Decock, L., & Douven, I. (2014). What is graded membership? *Noûs, 48*, 653-682. *https://doi.org/10.1111/nous.12003*

Dehaene, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychological Science, 12*, 244-246. *https://doi.org/10.1111/1467-9280.00343*

Dehaene, S., & Marques, J. F. (2002). Cognitive euroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 55*, 705-731. *https://doi.org/10.1080/02724980244000044*

De Wilde, E., Vanoverberghe, V., Storms, G., & De Boeck, P. (2003). The instantiation principle re-evaluated. *Memory, 11*, 533-548. *https://doi.org/10.1080/09658210244000126*

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition, 151,* 80-95. *https://doi.org/10.1016/j.cognition.2016.03.007*

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic, 42*, 137–160. *https://doi.org/10.1007/s10992-011-9216-0*

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science*, *41*, 686-722. *https://doi.org/10.1111/cogs.12359*

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology, 63*, 341-355. *https://doi.org/10.1037//0022-3514.63.3.341*

Dunning, D., & McElwee, R. O. (1995). Idiosyncratic trait definitions: Implications for self-description and social judgment. *Journal of Personality and Social Psychology, 68*, 936-946. *https://doi.org/10.1037//0022-3514.68.5.936*

Edgington, E., & Onghena, P. (2007). *Randomization tests*. Boca Raton: Chapman & Hall/CRC.

Egré, P. (2017). Vague judgment: A probabilistic account. *Synthese*, *194*, 3837-3865. *https://doi.org/10.1007/s11229-016-1092-2*

Egré, P., & Barberousse, A. (2014). Borel on the heap. *Erkenntnis, 79*, 1043-1079. *https://doi.org/10.1007/s10670-013-9596-3*

Fara, D. G. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics, 28*, 45-81. Originally published under the name 'Delia Graff'. *https://doi.org/10.5840/philtopics20002816*

Fara, D. G. (2008). Profiling interest relativity. *Analysis, 68*, 326-335. *https://doi.org/10.1111/j.1467-8284.2008.00761.x*

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*, 441-461. *https://doi.org/10.1016/s0022-5371(79)90246-9*

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137-165. *https://doi.org/10.1016/s0010-0277(97)00042-5*

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*, 355-384. *https://doi.org/10.1080/15326900701326402*

Hampton, J. A., & Passanisi, A. (2016). When intensions don't map onto extensions: Individual differences in conceptualisation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 505-523. *https://doi.org/10.1037/xlm0000198*

Hampton, J. A., & Williams, S.-K. (2016). *When asking for clarity leads to greater vagueness.* Presentation at the 57th Annual Meeting of the Psychonomic Society, Boston, MA.

Hansen, N., & Chemla, E. (2017). Color adjectives, standards, and thresholds: An experimental investigation. *Linguistics and Philosophy*, 40, 239-278. *https://doi.org/10.1007/s10988-016-9202-7*

Hersh, H. M., & Caramazza, A. (1976). A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General, 105*, 254-276. *https://doi.org/10.1037//0096-3445.105.3.254*

Janczura, G. A., & Nelson, D. L. (1999). Concept accessibility as the determinant of typicality judgments. *The American Journal of Psychology, 112*, 1-19. *https://doi.org/10.2307/1423622*

Kamp, J. A. W. (1975). Two theories of adjectives. In E. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge: Cambridge University Press.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*, 129–191. *https://doi.org/10.1016/0010-0277(94)00659-9*

Kamp, H., & Sassoon, G. (2014). Vagueness. In P. Dekker & M. Aloni (Eds.), *The Cambridge handbook of formal semantics* (pp. 389-441). Cambridge, MA: Cambridge University Press.

Kennedy, C. (2007). Vagueness and grammar: The study of relative and absolute gradable predicates. *Linguistics and Philosophy, 30*, 1–45. *https://doi.org/10.1007/s10988-006-9008-0*

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry, 56*, 258-277. *https://doi.org/10.1080/0020174x.2013.784483*

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language, 81*, 345–381. *https://doi.org/10.1353/lan.2005.0071*

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy, 4*, 1-45. *https://doi.org/10.1007/bf00351812*

Labov, W. (1973). The boundaries of words and their meanings. In C.-J. Bailey & R. Shuy (Eds.), *New Ways of Analyzing Variation in English* (pp. 340–373). Washington DC: Georgetown University Press.

Lang, E. (1989). The semantics of dimensional designation of spatial objects. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives: Grammatical structure and conceptual interpretation* (pp. 263-417), Berlin: Springer-Verlag.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 1-36. *https://doi.org/10.1007/s11229-015-0786-1*

Lee, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification, 19*, 69–85. *https://doi.org/10.1007/s00357-001-0033-y*

Lewis, D. K. (1970). General semantics. *Synthese, 22*, 18–67. *https://doi.org/10.1007/bf00413598*

Löbner, S. (2002). *Understanding semantics*. London, UK: Arnold Publishers.

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. *https://doi.org/10.3758/bf03211575*

Malacara, D. (2002). *Color vision and colorimetry: Theory and applications*. Bellingham, WA: SPIE Press.

Malt, B. C., & Smith, E. E. (1982). The role of familiarity in determining typicality. *Memory & Cognition, 10*, 69-75. *https://doi.org/10.3758/bf03197627*

McNally, L. (2011). The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In R. Nouwen, R. van Rooij, H-C. Schmitz, U. Sauerland (Eds.), *Vagueness in Communication* (pp. 151-168), Springer: Berlin.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation, 20*, 2597-2628. *https://doi.org/10.1162/neco.2008.04-07-504*

Osherson, D., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 9*, 35-58. *https://doi.org/10.1016/0010-0277(81)90013-5*

Qing, C., & Franke, M. (2014). Meaning and use of gradable adjectives: Formal modeling meets empirical data. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1204-1209). Austin, TX: Cognitive Science Society.

Raffman, D. (2014). *Unruly words*. Oxford, UK: Oxford University Press.

Rips, L. J., & Turnbull, W. (1980). How big is big? Relative and absolute properties in memory. *Cognition, 8*, 145-174. *https://doi.org/10.1016/0010-0277(80)90010-4*

Rosch, E. H. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233. *https://doi.org/10.1037//0096-3445.104.3.192*

Rosch, E. H. (1978). Principles of categorization. In E. H. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.

Rosch, E. H. (1999). Reclaiming concepts. *Journal of Consciousness Studies, 6*, 61-77.

Rosch, E. H. (2011)."Slow lettuce": Categories, concepts, fuzzy sets, and logical deduction. In R. Belohlavek, & G. J., Klir (Eds.), *Concepts and Fuzzy Logic* (pp. 89-120). Cambridge, MA: The MIT Press.

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: A Journal of General Linguistics, 2(1)*: 92, 1-27. *https://doi.org/10.5334/gjgl.151*

Sæbø, K. J. (2009). Judgment ascriptions. *Linguistics and Philosophy, 32*, 327–352. *https://doi.org/10.1007/s10988-009-9063-4*

Smith, N. J. J. (2008). *Vagueness and degrees of truth*. Oxford, UK: Oxford University Press.

Solt, S., & Gotzner, N. (2012). Experimenting with degree. In A. Chereches, N. Ashton, & D. Lutz (Eds.), *Semantics and Linguistic Theory (SALT) 22* (pp. 166–187). Ithaca, NY: CLC.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 3–9). Cambridge, MA: MIT Press.

Verheyen, S., Dewil, S., & Egré, P. (2018). Subjectivity in gradable adjectives: The case of *tall* and *heavy*. *Mind & Language,* 1-20*. https://doi.org/10.1111/mila.12184*

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. *https://doi.org/10.1016/j.actpsy.2010.07.002*

von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics, 3*, 1–77. *https://doi.org/10.1093/jos/3.1-2.1*

White, A., Malt, B. C., Storms, G., & Verheyen, S. (2018). Mind the generation gap: Differences between young and old in the representation of everyday lexical categories. *Journal of Memory and Language, 98*, 12-25. *https://doi.org/10.1016/j.jml.2017.09.001*

Williams, J. R. G. (2011). Degree supervaluational logic. *The Review of Symbolic Logic, 4*, 130-149. *https://doi.org/10.1017/s1755020310000237*

Williamson, T. (1994). *Vagueness*. London, UK: Routledge.

Appendix to Typicality and graded membership in dimensional adjectives

This section documents the information that can be found in Fig. 4 and Tables 4-5 of Section 4.2 (Membership degree at the aggregate level) in the main article. Instead of a linear representation of the numerical stimuli, a logarithmic transformation is assumed.

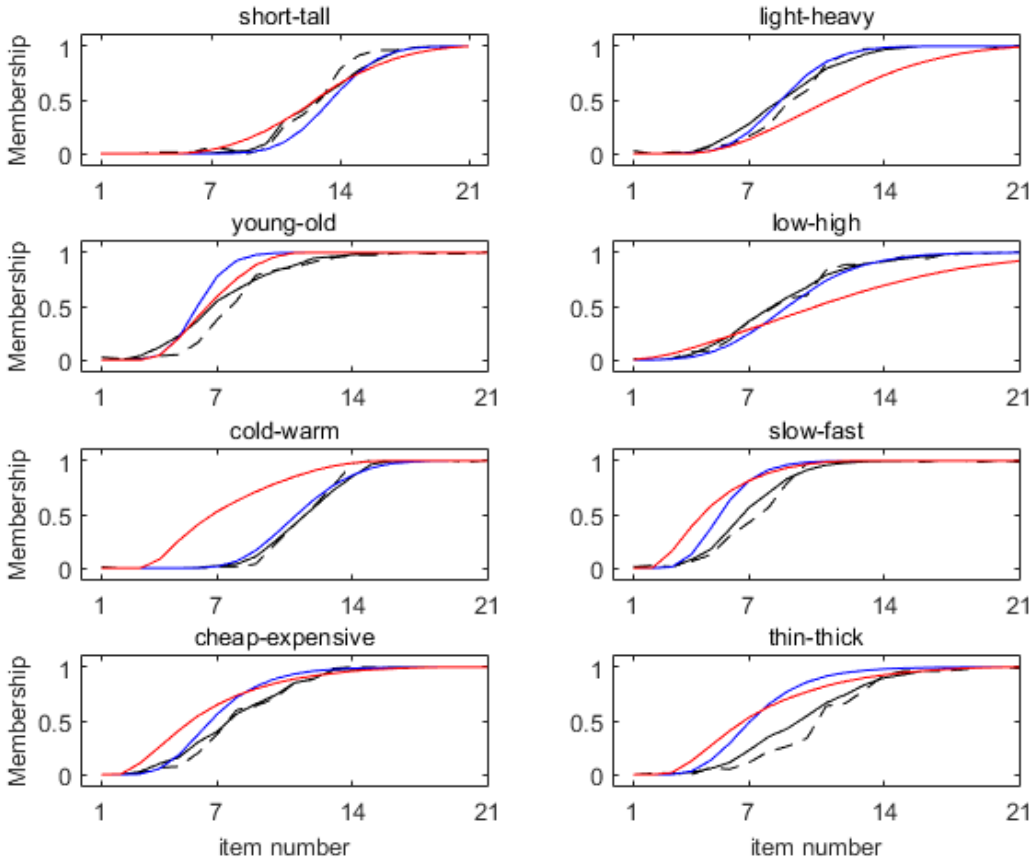

Figure A1. Observed degrees of membership (black) based on the dichotomous (dotted) and the continuous (solid) categorization task. Predicted degrees of membership based on normal (blue) and uniform (red) sampling from the log-transformed generated typical values.

Table A1 indicates that normal sampling yields a better fit than uniform sampling does, except for *short-tall*, *young-old*, and *thin-thick*. The difference between normal sampling and uniform

sampling is more pronounced assuming a linear representation. On average, the linear representation leads to a better prediction of the dichotomous categorization data than the log-transformed representation does (*M*=.19 vs. *M*=.33). With regards to the prediction of the continuous categorization data, the log-transformed representation on average does slightly better than the linear transformation does (*M*=.17 vs. *M*=.22).

Table A1

Sums of squared deviations (SSDs) *between the observed membership degrees (continuous vs. dichotomous categorization) and between the observed and the predicted membership degrees at the aggregate level. Predictions result from normal or uniform sampling of generated typicality values and are made with respect to either the dichotomous or continuous categorization data.*

Pair	Categorization data	Dichotomous categorization		Continuous categorization	
		Normal	Uniform	Normal	Uniform
short/tall	0.06	0.13	0.13	0.10	0.04
light/heavy	0.07	0.05	0.57	0.03	0.60
young/old	0.12	0.54	0.22	0.25	0.07
low/high	0.01	0.05	0.52	0.05	0.50
cold/warm	0.03	0.05	2.00	0.02	1.87
slow/fast	0.05	0.48	0.78	0.24	0.51
cheap/expensive	0.02	0.17	0.42	0.12	0.28
thin/thick	0.12	1.13	1.01	0.55	0.49

Table A2 indicates that normal sampling tends to yield membership degree curves that better represent the observed membership degree curves than uniform sampling does. A few notable exceptions are *young/old* and *thick/thin* for which PSE and Slope are better predicted from uniform sampling. For *short/tall*, normal sampling better predicts the slope, while uniform sampling predicts the PSE better. For *slow/fast*, the reverse holds.

Table A2

Approximated Point of Subjective Equality (PSE) and Slope for the observed and the predicted membership degrees.

Pair	PSE				Slope			
	Observed		Predicted		Observed		Predicted	
	Dichtomous	Continuous	Normal	Uniform	Dichtomous	Continuous	Normal	Uniform
short/tall	12.59	12.87	13.56	12.77	.20	.18	.22	.14
light/heavy	9.12	8.81	8.76	11.55	.19	.17	.22	.11
young/old	8.00	7.19	6.05	6.68	.17	.16	.36	.26
low/high	8.71	8.63	9.22	11.10	.13	.13	.15	.07
cold/warm	11.51	11.58	11.31	7.38	.26	.21	.20	.15
slow/fast	7.27	6.94	5.68	5.02	.21	.21	.29	.21
cheap/expensive	8.10	7.85	7.04	6.38	.17	.15	.20	.13
thin/thick	10.59	9.72	7.46	7.57	.15	.14	.19	.12

On average, the slopes of the observed membership degree curves are better predicted assuming a linear representation of the numerical stimuli than assuming a log-transformed representation (expressed as mean absolute deviation or MAD: MAD=0.03 vs. MAD=0.06 for dichotomous categorization and MAD=0.04 vs. MAD=0.06 for continuous categorization). For dichotomous categorization PSE is better predicted assuming a linear representation (MAD=1.06 vs MAD=1.22), while the reverse holds for continuous categorization (MAD=1.20 vs MAD=0.88).