Typicality and graded membership in dimensional adjectives

Steven Verheyen* (steven.verheyen@ens.fr)

Paul Egré (paul.egre@ens.fr)

Institut Jean Nicod, École Normale Supérieure

*Address for correspondence: Steven Verheyen, École Normale Supérieure, PSL Research University, Institut Jean-Nicod, Pavillon Jardin, 29, rue d'Ulm, 75005 Paris, France. E-mail: steven.verheyen@ens.fr

Abstract

This paper concerns an investigation of the conceptual spaces account of graded membership in the case of gradable adjectives. Douven and collaborators have shown that the degree of membership of an item intermediate between two color categories (green vs. blue) or two shape categories (vase vs. bowl) can be derived from the categories' typical instances. An issue left open is whether the conceptual spaces approach can account for graded membership in more abstract categories. In this paper we consider dimensional adjectives such as *tall* and *expensive*, for which the notion of prototypicality is more problematic. We present the results of an empirical study showing that the account can be extended successfully to that class, taking advantage of systematic relations of antonymy in those adjectives. The approach's assumption that typical instances of a category are equally typical and its ability to account for inter-individual differences in degree membership are discussed.

Keywords: vagueness; conceptual space; gradable adjectives; antonyms; prototypes; categorization.

Word count: 15498

Typicality and graded membership in dimensional adjectives

1. Introduction

Thresholds are at the heart of linguistic, philosophical, and psychological accounts of vagueness (Bartsch and Vennemann 1972; Borel 1907; Egré, 2016; Fara 2000; Hampton, 2007; Kennedy 2007; Lassiter & Goodman, 2015; 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 vagueness, 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 preponderance of thresholds in the vagueness literature might give the impression that thresholds are the primary means for language users to apply a verbal label to an item, but it contrasts with the observation that language users seldom spontaneously explicate the necessary threshold value (Dunning & McElwee, 1995). It is also in tension with observed violations of monotonicity (e.g. failing to say that some object taller than another classified as *tall* is also *tall*) and with intra-individual application differences (Hampton & Williams, 2016; Verheyen, Dewil, & Egré, 2017).

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) approach, in which prototypical values within a continuous metric

space determine the border between categories (Gärdenfors, 2000). 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 idea has been extended to deal with vagueness, by assuming that fuzzy thresholds between categories are derived by sampling instances from regions consisting of *multiple* typical values (Decock & Douven, 2014; Douven, Decock, Dietz, & Egré, 2013). The model has been tested experimentally: Douven, Wenmackers, Jraissati, and Decock (2016) 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, defined as the proportion of participants assigning it to the category, and the *predicted* degree of membership for that item, calculated as the relative number of times that item falls closer to some typical instance for C than to some typical instance for a contrast category C' across bisections of the space resulting from sampling different typical instances of C and C' (so-called Voronoi tessellations, see Decock and Douven, 2014, for a review of the model, and Kamp and Partee, 1995, for a definition of that notion of degree in a supervaluationist framework). Douven (2016) has found the same correspondence for the nominal categories *vase* and *bowl* in relation to a Labov-style stimulus set gradually morphing a vase to a bowl (Douven, 2016; Gärdenfors, 2000; Labov, 1973). What those studies suggest is that typical values constrain our verdicts of membership. They also confirm the idea that decisions of membership of an item under a concept are a function of the distance of that item to cognitive reference points.

Some difficulties remain, however, for an account of degrees of membership based exclusively on this model. First of all, not all vague categories appear to rely on the notion of prototype. Kamp and Partee (1995:176) point out that relative gradable adjectives like *tall*, *big*, or *heavy* seem to lack prototypes for two main reasons: first because of the "indefinite variety of things" to which they can

be applied, and secondly because "there is in general no natural upper bound on how tall [big/heavy] things can be". Yet, those are categories for which the notion of degrees of membership is particularly meaningful. Relative gradable adjectives are after all the textbook examples of vague categories. A second difficulty is more specific to the methodology of the conceptual spaces framework, and concerns inter-individual variability. Color concepts like *blue* and *green* and shape concepts like *vase* and *bowl* indisputably admit focal values or paradigmatic cases. Concepts such as *tall*, *heavy* or *expensive* on the other hand have been argued to show a lot more variability, not only because they can apply to so many things, but because they appear to leave more room for subjectivity (Egré, 2016; Kennedy, 2013; Verheyen, Dewill, & Egré, 2017). It is unclear whether an account of degrees of membership based on the notion of typicality can accommodate inter-individual variability in application in the case of such adjectives. It could well be the case that for such adjectives, the notion of threshold remains psychologically fundamental, in the way implicit in standard semantic accounts (see Kennedy, 2007).

Neither of those difficulties is insuperable, but both call for scrutiny. Regarding the first, Kamp and Partee point out that even though the adjective *tall* by itself may not have a prototype, an adjective-noun combination like *tall tree* does have a prototype. Since ascriptions of tallness, heaviness, and so on, are always made relative to an explicit or implicit comparison class argument, it remains meaningful to investigate the role of prototypes for decisions of membership in relation to such combinations. This partly handles the other point made by Kamp and Partee: 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. But this does not imply that typical values will vary without limit once a comparison class argument has been specified. Regarding the second difficulty, it is not necessarily the case that what is typical will be typical to the same degree across subjects. If the representation of what's typical can vary from one individual to another depending on the concept, then degrees of membership may likewise vary from one individual to another in a way that

is consistent with the model. Because of that, whether the model can accommodate such variation is worth investigating.

Based on those considerations, our main goal in this paper is an extension of the CS account of vague categories to handle the case of relative gradable adjectives, with specific emphasis on antonymous dimensional adjectives restricted by a comparison class argument. Antonymous dimensional adjectives are gradable adjectives, which refer to the same scale of a given dimension, but are ordered in opposite directions (Bierwisch, 1989, p. 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 approach can be applied most straightforwardly. Yet we propose to test whether the account can be generalized to that class of adjectives, knowing that it has been tested so far only on lexical items with different structural properties (see below). With this main goal come two subordinate goals, which concern the investigation of what counts as typical: first, the CS account so far assumes that the typical items of a category are equally typical; secondly, the CS account has been evaluated at the group level, by aggregating individual data. We propose to question both aspects in this paper. We are interested in whether an approach based on the idea that typicality itself could have a gradient could outperform a model in which typicality is uniform. And we are interested in the extent to which individual variations as to what is typical can be captured by the model. We give more details about those subordinate goals in the next section.

2. Scope and methods of our study

The scope and methods of our study differ from those of the previous studies on the CS account in the following respects.

Firstly, we focus on **relative gradable adjectives**, namely adjectives for which the notion of what counts as typical is essentially context-sensitive. Unlike absolute 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 (Burnett, 2016; Kennedy & McNally, 2005). They differ from color adjectives in that same respect: *green* and *blue* have focal values, for which there is no greener or bluer value proper. Not so for *tall* or *expensive*, for which there can always be taller or more expensive values in principle. Unlike *bowl* and *vase*, or *green* and *blue* when nominalized (as in "blue is my favorite color"), which can conjure up typical values by themselves, without the need for a modifier expression, adjectives like *tall* or *expensive* need a comparison class argument in order to ground both typicality and membership judgments (as in "*tall* man / for a man", "*expensive* smartphone / for a smartphone", see Kamp, 1975; Kamp & Partee, 1995; Klein, 1980; Rips & Turnbull, 1980). In what follows we thus examine the relation between membership and typicality judgments for adjective noun combinations.

Secondly, we focus on **dimensional adjectives** (in the sense of one-dimensional, see Bierwisch, 1989), that is adjectives for which there exists a standard measurement scale along a **single** physical dimension (height in cm for *tall*, price in euros for *expensive*, and so on). Our focus therefore differs from that of Douven et al. in earlier studies, which systematically involved multi-dimensional concepts (such as colors or shapes, each time represented along two or three dimensions).

Thirdly, we are interested i**n abstract representations of typicality in long-term memory**. That is, although physical scaling and psychophysical scaling are two distinct processes (see Luce, 1972), we postulate that a representation of numerical values for combinations of a one-dimensional adjective with a comparison class argument is accessible to naïve subjects to whom those scales are meaningful. Concretely, this means that in our study we ask participants to produce numerical values that they judge typical along standard physical scales. Our methodology here differs from the methodology followed by Douven et al. in earlier studies, which relied not just on memory but also on the presentation of actual stimuli to direct perception. Thus, Douven et al.'s participants see actual shades of blue and green among which they are asked to select the best exemplars, and

likewise they see particular silhouettes of a bowl or a vase among which to select. It would have been possible to present our participants with pictures in the same way (see Solt & Gotzner, 2012; Qing & Franke, 2014; Verheyen, Dewill, & Egré, 2017). We chose to probe participants' abstract representation of prototypes instead for two main reasons: on a practical level because it makes the preparation of stimuli easier; on a more fundamental level because speakers are not only faced with the task of assigning perceived stimuli to verbal categories, but also because they 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'').

Finally, we propose to refine two aspects of the methodology followed in the previous studies so far in the **treatment of typical values**. One is the assumption that typical values of a concept carry equal weight, and thus can be sampled from a uniform distribution. We are interested in checking whether sampling from a normal distribution might produce better results. The idea that the region of typical values itself could have more or less central instances follows the observation that typicality is generally considered a gradient phenomenon (Barsalou, 1985; Hampton, 2007; Rosch, 1975b). That gradient could moreover 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); finally it might reflect lack of knowledge or uncertainty about what counts as typical (Lynch, Coley, & Medin, 2000; Malt & Smith 1982;).

Another issue concerns the fact that in previous studies degrees of membership are calculated from aggregated data, that is, from the average of individual judgments. Relative gradable adjectives are known to exemplify subjective differences, however, to a greater extent at least than expressions that show less context-sensitivity (Kennedy, 2013; Sæbø, 2009; Verheyen, Dewill, & Egré, 2017). If

the CS model is right, it would seem that individual differences in degrees of membership ought to be predictable from individual differences in typicality judgments. Whether this prediction is right is not obvious, however. It could very well be the case that typicality judgments -- if typical means socially salient or socially robust -- account only for the commonality in individual membership responses, but that individual differences originate from another source.

To cast light on these issues, we present the results of an experiment on a sample of eight adjectival-noun combinations in the next sections. In Section 3, we present the design and methods of our study. Section 4 presents 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

3.1. Participants

Eighty undergraduate students in psychology at the University of Leuven 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 L1. Informed consent was obtained from all participants in accordance with the university's guidelines for experimentation with human subjects.

3.2. Materials

All materials were in Dutch. Table 1 provides an overview of the English translations. We included eight antonym pairs consisting of contrary adjectives (e.g., *short-tall*). Since all the adjectives were relative gradable adjectives, 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). The pairs thus comprise of antonymous dimensional adjectives, which refer to the same scale of a given

dimension, but differ in the ordering on the scale (Bierwisch, 1989). 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*, 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, which tend to be situated at more extreme ends of the spectrum. 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.[1]

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).

[1]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 experiment. 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.

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 (Verheyen & Egré, 2017). 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 adjectives that make 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 (in cm) 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. Polytomous 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 labelled "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 polytomous categorization task. They were asked to indicate the degree to which they are inclined to apply the negative or the positive adjective. They could accomplish this by positioning a slider along a horizontal scale ranging from the negative to the positive adjective. The instructions for a participant who would have chosen intermediate instead of *short* or *tall* for a male adult of 170 cm, would read: *"How inclined are you to call a male adult of 170 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 polytomous 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.[2]

4. Results

The Results section of this paper is organized in three parts according to three distinct types of analyses that were conducted:

[2] This task was added to confirm — at the level of individuals instead of at the aggregated level — the finding by Bear and Knobe (2017) that the categorization threshold is predicted by considerations of both average and ideal values. We found it to replicate 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.

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 approach 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 approach. 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 associates assumes that typical values all carry equal weight (uniform sampling). We use the same model but compare it with a model in which typical values can have a gradient (normal sampling).

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. As was the case for the second part of the paper, the focus is on the comparison of uniform and normal sampling.

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

Figure 1 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.

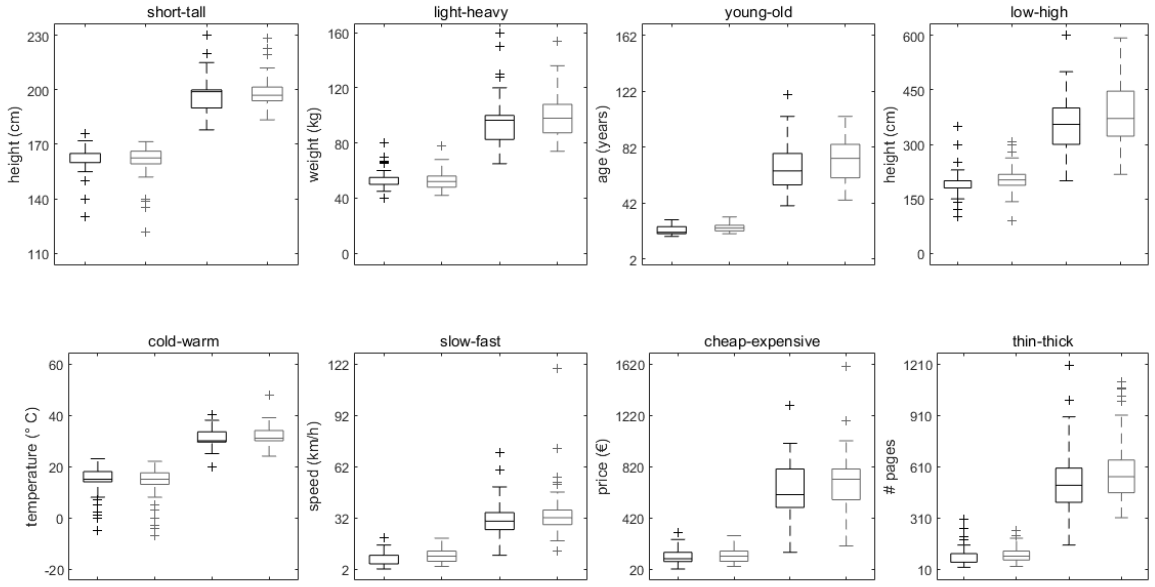

Figure 1. Boxplots of generated (black) and selected (grey) typicality values.

As expected, the typicality values are higher for the positive adjective in each pair than for its negative counterpart. Figure 1 also indicates that the selected typicality values tend to be somewhat higher than the generated typicality values. According to a one-tailed paired samples t-test, the single value generated by participants is significantly smaller ($p < .0031$) than the average of the values they selected for 4 out of 8 positive adjectives: *old* (t=-5.36, d=-.60), *warm* (t=-3.02, d=-.34), *fast* (t=-4.32, d=-.48), *thick* (t=-3.66, d=-.41), and 2 out of 8 negative adjectives: *young* (t=-7.42, d=-.83), *cheap* (t=-3.26, d=-.36). The corresponding effect sizes are small to medium, except for *young* where Cohen's *d* is large (Cohen, 1988).

Figure 1 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 Levene's test 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* (*F*(1,79)=7.11, *p*=.008) and *cold/warm* (*F*(1,79)=2.65, *p*=.11).

Table 2

Levene's test comparing the variance of generated typicality values for negative and positive adjectives.

Pair	Negative SD	Positive SD	*F*
short/tall	7,59	9,95	7,11
light/heavy	7,26	22,73	25,88
young/old	3,53	15,15	77,11
low/high	38,93	132,62	27,91
cold/warm	5,42	3,72	2,65
slow/fast	3,87	10,48	26,60
cheap/expensive	60,83	252,1	32,77
thin/thick	51,63	207,38	37,06

Since in the typicality selection task participants selected multiple values (instead of a single value in the typicality generation task), the data from this task allow us to establish whether this variability difference also holds at the individual level. To do so we compare 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.[3]

[3] 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 eight pairs at *α*=.0063, except for *short/tall* where *p*=.01.

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 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.[4]

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

4.1.2. Discussion

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

Between-subject variability in the choice of typical values is evidenced by the standard deviations in Tables 2 and 3, and by the boxplots in Figure 1 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. Within subject-variability is shown in the absolute and relative differences observed between the generated typicality values and the (mean) selected typicality values. For 6 out of the 16 adjectives, participants selected higher typicality values than they produced.

The observed variability in typicality accords with the observation in the literature on nominal concepts that while mean typicality ratings tend to be quite reliable, meaningful differences in what is considered to be typical exist between and within individuals (Barsalou, 1987, 1989; Hampton & Passanisi, 2016). The reasons for the difference are considered to be varied and include differences in knowledge and accessibility (Barsalou, 1989, 1993). The existence of substantial inter-individual typicality differences that can to a considerable degree be replicated across different tasks, is an incentive to continue with the analysis plan and investigate the performance of the CS account at the individual level as well (see Section 4.3). While the research on typicality of nominal concepts suggests that within-participant differences are to be expected, we should acknowledge that the context of both typicality tasks is quite different. In the generation task, participants are to

spontaneously produce a single value, whereas in the selection task they are to select multiple values from a presented range. The presence of a range of values the participant might interpret as "natural" according to the experimenter could make participants adjust their selection in order to agree better with their idea of the experimenter's expectations.

In this light, it is interesting to observe that all the pairs for which the range of the selection task was constructed by adding solely high values (*slow-fast*, *cheap-expensive*, and *thin-thick*) show a significant difference between generation and typicality. On the other hand, the typicality values for the majority of adjectives (10/16) are not significantly affected by the context difference between the two typicality tasks. In what follows, we remain true to our original plan of using the generated values for the aggregated 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. Importantly, 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 count as a physical or psychological zero; yet on that example the variance for the negative and the positive antonym does not differ significantly in the typicality generation task. World-knowledge is also 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.[5] For *short* vs. *tall* in relation to human heights, it is world-knowledge that heights are normally distributed. 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.

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., 2016), 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).[6]

4.2. Membership degree at the aggregated 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

[5] 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.

[6] 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 approach 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.

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 towards the positive adjective. This is a convention we will use throughout this paper. Membership of the negative adjective is then 1 minus the membership degree for the positive adjective (see Douven, 2016, and Douven et al., 2016, 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; Douven, 2016; Douven et al., 2016; Egré & Barberousse, 2014; Hampton, 1998, 2007). In this case the membership degree for the negative adjective is 1 minus the proportion of participants who judged the positive adjective to apply.

Figure 2 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 account 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 completions. The predicted membership curves are the averages across these 10,000 repetitions.

We discern two sampling schemes. The uniform sampling scheme draws cognitive reference points from the interval bounded by the smallest and the biggest typicality values that were generated by the participants. This corresponds to the procedure used by Douven and colleagues (Douven, 2016; Douven et al., 2016). 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.[7] In Figure 2 the predicted membership curves based on uniform sampling are depicted in red. The predicted membership curves based on normal sampling are depicted in blue.

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

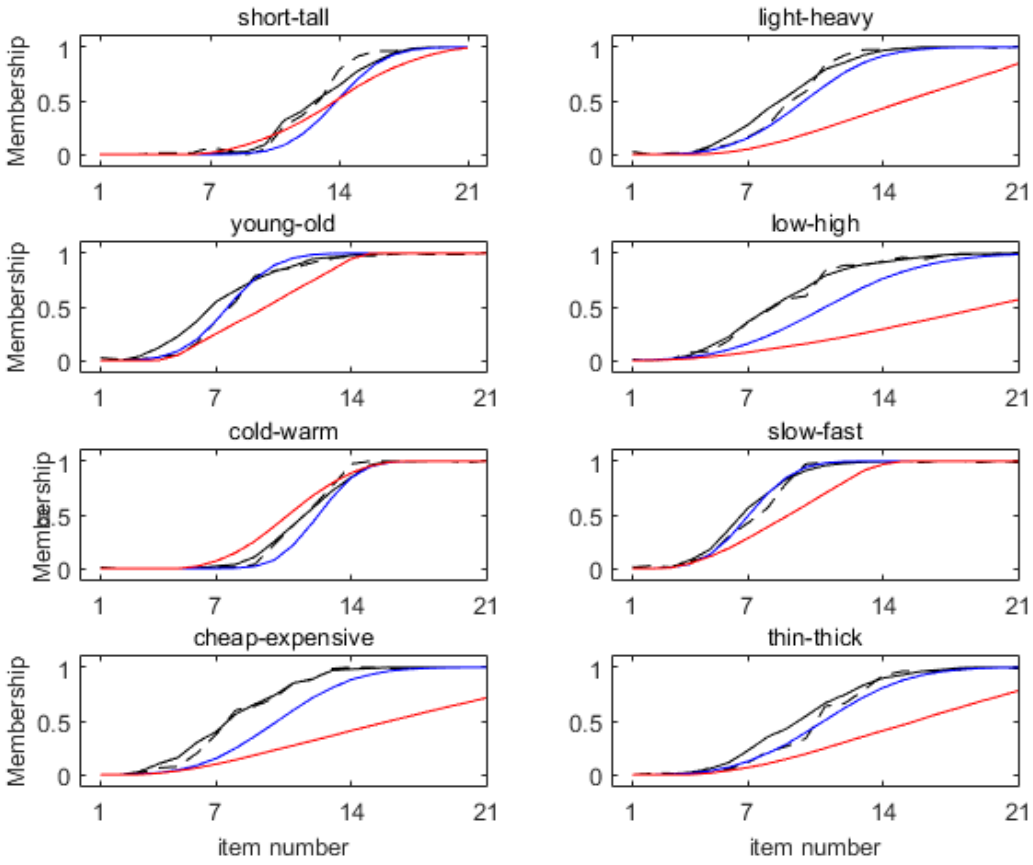

Figure 2. Observed degrees of membership (black) based on the dichotomous (dotted) and the continuous (solid) categorization task. Predicted degrees of membership based on normal (blue) and uniform (red) sampling from the generated typical values.

Figure 2 indicates that the normal predictions better approximate 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 also shows 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 provides a natural benchmark for assessing the quality of the fit. Both in Figure 2 and in Table 4 the adjectives *low/high* and *cheap/expensive* stand out as two contrary pairs for which the CS approach fares less well.

Table 4

Sums of squared deviations between the observed membership degrees and between the observed and the predicted membership degrees at the aggregated level.

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. We determined the PSE (Point of Subjective Equality) and the Slope of the membership curves following the specifications by Hampton and Williams (2016; see also Douven et al., 2016). The PSE is the point for which the membership degree equals .50. From Equation (1) it follows that the PSE is x = -a/b. The slope of this logistic function is 1/b.

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

The results 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 estimated 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 better approximate the slopes of the observed membership curves than the slopes of the uniform predictions do. Both predictions tend to overestimate the PSE, but this tendency is more pronounced for uniform predictions. Also in terms of PSE and Slope, the contrary pairs *low/high* and *cheap/expensive* are the least well approximated by the conceptual space predictions.

The results in Table 5 also add a number of insights that are not readily apparent in Figure 2. Except for the pairs *low/high* and *slow/fast*, the membership degree curves resulting from the continuous task tend to be flatter than the degree curves resulting from the dichotomous task (sixth and seventh column). Except for *short/tall* and *cold/warm*, the PSE also occurs earlier in the continuous membership degree curves compared to the binary membership degree curves (second and third column).

Table 5

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	1,26	1,39	1,16	1,93
light/heavy	9,12	8,81	9,97	15,37	1,30	1,50	1,56	3,11
young/old	8,00	7,19	7,62	9,53	1,49	1,59	1,06	1,63
low/high	8,71	8,63	11,13	18,75	1,89	1,89	2,31	4,63
cold/warm	11,51	11,58	12,30	10,77	0,95	1,19	0,89	1,37
slow/fast	7,27	6,94	7,07	8,99	1,19	1,20	0,99	1,66
cheap/expensive	8,10	7,85	10,21	16,18	1,46	1,63	1,73	4,05
thin/thick	10,59	9,72	10,99	15,82	1,68	1,83	1,86	3,46

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. Nevertheless, we also observed differences in the slope and point of subjective inequality of membership curves obtained with the different procedures. As was the case for typicality, the literature on noun concepts suggests at least two sources of the observed differences: the repetition of a categorization task at a later time is known to produce intra-individual differences (Hampton, Aina, Andersson, Mirza, & Parmar, 2012; McCloskey & Glucksberg, 1978) and task differences have shown to produce different membership curves

(Douven et al., 2016). Since no one task is arguably the better task for eliciting 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 to assess the absolute fit of the CS predictions against. The SSDs between the observations and the predictions of the CS approach with normal sampling of reference points, indicate a decent fit provided we leave out the two pairs the CS approach poorly accounts for (we defer a discussion of why the CS approach 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.

Irrespective of the means used to elicit membership degrees, the predictions of the CS account with normal sampling outperform the predictions of the CS account 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 under uniform sampling. It is an intrinsic property of the CS approach 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. To see this, imagine the situation where there is only one prototypical value for the positive adjective and one prototypical value for the negative adjective. Together they will determine a single threshold, located halfway between the two reference points, resulting in a discrete threshold function (i.e., this situation would not admit borderline instances). Normal sampling is a natural solution for down-weighting 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.

Based on the insight that more extreme typicality values impact the CS model predictions, we implemented a number of alternative models. Removing the first and last quartile of the

generated typicality values resulted in very steep curves (which resemble threshold curves) because it severely restricts the range. This was the case for both uniform and normal predictions. In line with the reasoning above, removing outliers (the 2,5% smallest and 2,5% largest generated values) tended to improve the predictions based on uniform sampling. It did not affect the results for normal sampling systematically, however (it led to similar, better, or worse predictions depending on the adjectival pair). We decided to report the results based on the complete set of generated values because the CS approach is intended to be a parameter free account of graded membership (see Douven et al. 2016), and the decision not to include certain typicality values introduces researcher degrees of freedom.

In the previous section we established that there is more variability among the typicality values for the positive adjective in contrary pairs. This is taken into account in the CS approach 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 distribution 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 approach 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 slopes of the predicted degree functions corresponds perfectly to the order of the variances pooled across the positive and negative generated instances (after they have been brought onto a common scale). This provides us with a means of assessing the specificity of the predictions produced by the CS approach: 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 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 contrary gradable adjectives in our study.

4.3. Membership degree at the individual level

4.3.1. Results

So far the CS approach 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 approach at the individual level as well. The former task provides us with a continuous membership curve for every individual, which we aim to predict with the CS approach using the values from the typicality selection task, in which participants select for each of the gradable adjectives the typical values from a list.

Again 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.[8] 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 in the previous section.

[8] 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.

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 lists the average 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 $\alpha=.0063$ 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).[9]

Table 6

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

Pair	Categorization data	Individual predictions			
		Normal	Uniform	t	Cohen's d
short/tall	1,24 (0,80)	1,10 (1,14)	1,21 (1,16)	-14,67	-1,64
light/heavy	1,16 (0,67)	0,98 (0,98)	1,08 (1,00)	-7,41	-0,83
young/old	1,36 (1,14)	1,27 (1,18)	1,41 (1,22)	-10,92	-1,22
low/high	1,72 (1,45)	1,71 (1,84)	1,85 (1,87)	-10,94	-1,22
cold/warm	0,80 (0,52)	0,70 (0,59)	0,80 (0,62)	-12,45	-1,39
slow/fast	1,04 (0,74)	0,74 (1,28)	0,84 (1,30)	-12,25	-1,37
cheap/expensive	1,12 (0,81)	1,84 (1,75)	1,96 (1,78)	-8,48	-0,95
thin/thick	1,68 (1,06)	1,61 (2,00)	1,78 (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 continuous and the binary

[9] 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' binary and continuous empirical membership curves and discarded the data from the 10% worst participants. This did not affect the finding that the normal CS model outperformed the uniform CS model.

membership curve for every participant. Both curves are intended to measure the same construct but an individual's binary curve clearly only provides a coarse indication of membership degree. As such, it provides a baseline against which to assess the fit of the normal CS model's predictions. The second column of Table 6 includes the average SSD between individuals' binary and continuous membership curves across all participants. While the average SSD for the normal CS model is lower than the average SSD for the binary curve 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, and determined the Point of Subjective Equality (PSE) and the Slope in the same manner we used for the aggregate curves in section 4.6.1.[10] Table 7 lists the mean and standard deviation of PSE and Slope for each of the adjective pairs. These results too indicate that the absolute fit of the predicted membership curves can be improved. The normal and the uniform CS model overshoot the observed PSE to the same extent. Paired samples t-tests (not shown) indicate that there is no significant difference in the absolute deviation from the observed Points of Subjective Equality of the Points of Subjective Equality resulting from the normal CS model and the uniform CS model for any of the adjective pairs. The absolute deviation from the observed Slopes, 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.

[10] 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

Mean (SD) 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	12,89 (1,36)	14,34 (1,77)	14,33 (1,78)	1,08 (0,46)	0,60 (0,32)	0,37 (0,32)
light/heavy	8,81 (1,69)	10,12 (2,17)	10,11 (2,15)	1,09 (0,45)	0,65 (0,44)	0,42 (0,39)
young/old	7,2 (1,71)	8,65 (1,83)	8,65 (1,83)	1,14 (0,58)	0,52 (0,26)	0,29 (0,25)
low/high	8,66 (2,50)	10,76 (3,53)	10,78 (3,53)	1,14 (0,58)	0,63 (0,36)	0,38 (0,38)
cold/warm	11,58 (1,35)	12,5 (1,38)	12,51 (1,39)	0,81 (0,42)	0,51 (0,25)	0,3 (0,28)
slow/fast	6,81 (1,67)	7,36 (1,88)	7,37 (1,88)	0,77 (0,37)	0,43 (0,17)	0,22 (0,20)
cheap/expensive	7,73 (2,34)	10,43 (2,90)	10,41 (2,91)	0,95 (0,49)	0,69 (0,36)	0,43 (0,35)
thin/thick	9,74 (2,16)	12,02 (3,32)	12,03 (3,34)	1,27 (0,55)	0,86 (0,60)	0,56 (0,53)

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 observed and the predicted PSEs are all significant at α=.0063, except for *short/tall* (p =.03) and *young/old* (p=.01). None of the correlations between the 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 across individuals.

Pair	PSE		Slope	
	Normal	Uniform	Normal	Uniform
short/tall	0,25	0,25	-0,05	-0,05
light/heavy	0,49	0,49	0,00	-0,01
young/old	0,29	0,28	-0,05	-0,04
low/high	0,58	0,58	0,09	0,10
cold/warm	0,48	0,48	0,22	0,26
slow/fast	0,41	0,40	-0,10	-0,12
cheap/expensive	0,49	0,49	0,05	0,10
thin/thick	0,41	0,41	0,01	-0,02

As we already explained in section 4.2.2 for the aggregate data, in the CS model there is a straightforward relation between the variances of the employed typicality regions, and the slope of the resulting membership curve. The slopes of the predicted individual degree functions correlate .99 or higher with the variances pooled across the positive and negative selected instances (after they have been brought onto a common scale) 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 slopes of the observed continuous membership curves varied between -.10 for *slow/fast* and .06 for *cold/warm*.

4.7.2 Discussion

The main result we found at the aggregated level is replicated at the individual level: the CS model that assumes a normally distributed typicality region outperforms the CS model that assumes a uniformly distributed typicality region. The normal CS model fits the individual continuous membership curves better than the uniform CS model does (significantly lower SSD). 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 approximate the observed membership curves than the uniform CS model does.

While the normal CS model fits the empirical data relatively better than the uniform CS model does, it provides a rather coarse approximation of the observed membership curves in an absolute sense. The resulting SSD is comparable to that between the continuous and the binary membership curves, where one can consider the latter a crude rendition of the former. The normal CS model also overestimates the PSE and curvature of the membership degree function. The model accounts for 18% of the variance in the empirical PSEs, but does not capture any of the variance in the empirical slopes.

What can we conclude from the fact that the CS model fares less well at the individual level than it does at the aggregated level? The most severe interpretation is that the CS model 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. This may be the case, but alternative interpretations should be entertained as well.

In particular, we may wonder whether the CS model can legitimately be used to predict individual membership curves. Do people entertain genuinely different ideas about what one considers *short* and *tall* men? Or do they share a common understanding of these notions? If it is the latter, both the typicality values and the membership judgments we observed might just be random instantiations of a common distribution and one shouldn't expect a particularly close correspondence between them (see, for instance, Connell & Lynott, 2014). The fact that we are not constantly confronted with insurmountable communication problems suggests a shared understanding of the meaning of terms as *short* and *tall*. Nevertheless, meaningful differences in application of terms like *short* and *tall* or *light* and *heavy* have been shown to exist, even when they are sufficiently contextualized (Verheyen, Dewil, & Egré, 2017). In Section 4.1.1 we provided evidence that the inter-individual typicality differences we observed are 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). Genuine inter-individual differences in membership (Hampton, Dubois, & Yeh, 2006; McCloskey & Glucksberg, 1978; Verheyen, Hampton, & Storms, 2010) and typicality (Barsalou, 1987; Hampton & Passanisi, 2016; Verheyen, Van Deun, & Storms, 2017) judgments have also been repeatedly shown for nominal concepts.

Inter-individual differences might be more pronounced for some adjective-comparison class combinations than others. Individual CS models might, for instance, do a better job when the comparison for *short* and *tall* is buildings, than when the comparison class is adults, because we have a better mutual understanding of the latter due to increased experience or explicit instruction. This prediction is not supported by our data, however, in that we observe a positive relationship between the pooled variance across participants and the SSD. The adjective-comparison class combination for which we observed the least amount of variability in typicality, is also the one for which the CS model yielded the lowest SSD (*warm*/*cold* summer's day). The adjective-comparison class combinations with the highest SSD, also rank among the highest in terms of pooled typicality variance (see the General Discussion for an explanation of why *low*/*high* ceiling and *cheap*/*expensive* smartphone are not well accounted for by the CS approach).

Alternatively, our procedure might have failed to produce accurate data at the individual level. According to this account, the difficulty of the CS approach to account for the observed membership degree at the individual level may be due to an insufficiently precise assessment of the variability of the typicality region through the typicality selection task. Since in the CS approach the thresholds and thus the boundary regions are derived from cognitive reference points, the precisification of these typical instances is a requirement for the CS approach to deliver meaningful predictions. We also touched upon this matter in the discussion of the aggregate data (Section 4.2.2), where we mentioned how employing different cut-offs for eliminating outliers from the typical instances affected the CS models' predictions, and speculated that one reason why the normal CS model might

outperform the uniform CS model is that it provides for a natural way of down-weighting outlying typicality values. Whether or not one considers the superior account of the normal CS model theoretically justified because of the graded nature of typicality or merely a technical matter involving the down-weighting of extreme typicality values, these considerations point to a potential shortcoming of the CS approach: the challenge of determining the boundaries (degrees) of contrary 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 approach requires one to establish the boundaries of the typicality regions of each of the adjectives. Or put differently, to establish degree membership, the CS approach requires one to establish typicality degree first. While this observation does not imply any kind of vicious regress for the model (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).

5. General Discussion

5.1. Main findings

We set ourselves three goals in this paper. Our first and main goal was to see if the CS account of degrees of membership for borderline cases of a vague predicate, tested so far only on color adjectives and on some nominal categories, remains successful when tested on relative gradable adjectives. Our answer to this first question is generally positive. For 6 of the 8 pairs of relative adjectives that we tested in combination with nouns, we saw that observed degrees of membership are well predicted by the typicality values generated across participants (see Section 5.2 for a discussion of the two other pairs).

This finding lends further support to the CS model, since the adjectives we selected come in antonym pairs, in the same way in which, in the context of Douven's previous studies, the lexical items tested form contrastive pairs in context (vase vs. bowl, green vs. blue). Moreover, we get a confirmation that even though *grammatically* the positive form of relative gradable 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. Consequently, what we see is that those values do indeed determine the extent to which an item is likely to be placed under the category.[11] 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 combination for which participants did not see actual exemplars, but were asked to imagine values relative to abstract physical scales (more on this below). This finding is worth highlighting, for Douven and associates left as an open question whether the CS account of vague categories can be extended to abstract semantic categories.

Beside this main goal, we had two subordinate goals, of a more methodological nature, regarding the notion of typicality. The first of those was to examine whether the model makes better predictions when typical values themselves are equipped with a gradient. We undertook 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 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

[11] Kamp and Partee mention that even though compounds like "tall man" can have a prototype, the resemblance to the prototype does not determine the extension of an item under the category. One argument may be that someone who is very much taller than a typically tall man would still count as *tall*, although that person is not typical for a tall man. However, in the CS approach a very tall man would under ALL completions resulting from sampling typically *tall* and typically *short* instances, be closer to the prototype for *tall* than to the prototype for *short*, and therefore receive a membership degree of 1. For items that are atypical in the direction opposite to the antonym (such as a very tall man), intuitively the degree of membership should remain 1, and thus one could argue that resemblance or distance to the prototype does determine the membership for these extreme cases.

relative gradable adjectives some instances may be considered more representative or are more available than others. More can be said about the choice of normal vs. uniform sampling, however (see below Section 5.2).

The second subordinate goal we had was to investigate inter-individual differences concerning the connection between typicality and degrees of membership. Because different individuals generate and select different values as typical, we assessed the CS approach at the individual level. Our findings here were less successful. 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 very good. We cannot take this finding to be conclusive against the CS account, since our method for determining the variability among typical instances might not have been as precise in the individual cases as in the group case. More work is therefore required to examine the performance of the CS approach regarding inter-individual differences. In the following section we take up this and other challenges our study raises for the CS approach.

5.2 Challenges for the CS approach

In contrast to the majority of vagueness accounts, thresholds play a secondary, rather than a primary role in the CS approach in that they are derived from typical instances of two contrasting categories. The CS approach posits that the degree membership 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'. The CS approach thus tackles the question of how to achieve separateness and clarity of continuous categories (Rosch, 1978) not by looking at the categories' boundaries, but at their typical cases. As such the approach places tremendous emphasis on the specification of these typical instances and at least in its current form suggests typicality is all there is to deriving degree membership. One reading of our results is

that the former is not readily achieved. We hypothesized that the rather poor performance of the CS approach at the individual level could be due to a misspecification of the variability among typical instances and at the aggregated level documented how different ways of dealing with outliers affected the CS models' predictions. Our main finding, that a CS model assuming normal sampling of typical instances outperforms 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 reporting abstract values (e.g., due to lack of knowledge) is limited. Whether the CS approach should embrace idiosyncratic typicality distributions that allow a gradient deserves further consideration. Relative gradable adjectives are particularly suited to address this question since they are prone to context-sensitivity and subjectivity. 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, 2017).

Throughout this paper the results for *cheap/expensive* and *low/high* 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 approach. These explanations of course need to be tested in future research.

In the case of *cheap*/*expensive* smartphones, the explanation for the oddity might lie in the greater subjectivity of this particular assessment (Kennedy, 2013; Kamp & Sassoon, 2014). 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.[12]

For the subjectivity argument to hold, however, one would have to entertain the possibility that subjectivity 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* wasn't well accounted for at either level. The research on subjectivity would then suggest that in addition to typicality, other factors might also contribute to the application of terms (e.g., egocentric reference, see Verheyen, Dewil, & Egré, 2017, or practical interests, see Fara, 2000, 2008, or a representation of what counts as ideal, see Bear and Knobe, 2017).

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 representation as being centered around a single value (normal CS approach) or having no mode whatsoever (uniform CS approach).

The observation that a concept has multiple reference points is not an uncommon one. Rosch (1975a) already observed the special status of the multiples of 10 for numbers, of basic emotions for emotions, of focal colors for colors, and of vertical, horizontal, and diagonal lines for line orientations (see Medin & Shoben, 1988, and Voorspoels, Vanpaemel, & Storms, 2008, for other illustrations). Multimodality might apply to an individual's typicality distribution, so this explanation isn't invalidated by the fact that the CS approach 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

[12] That is not to say that *tall* has no subjective meaning whatsoever. See Dunning and Cohen (1992) and Verheyen, Dewil, and Egré (2017) for evidence that the standard for application of *tall* is influenced by one's own height.

another illustration of the fact that for the CS approach to be able to provide a good indication of degree membership, one has to first get a good measure of the typicality distribution.

5.2. Scaling

We would like to conclude this paper with a note on psychological scaling. Douven and colleagues evaluated the CS approach for shapes (pictures of containers in Douven, 2016) and colors (blue and green in Douven et al., 2016). 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 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 studies 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 and the observation that variability of typicality is higher at the positive end of the scale than at the negative end, might be seen as a manifestation of the accumulation of error with magnitude the first account of number representation proposes. According to the second account, as numbers grow they are increasingly positioned closer to each other with a constant variance.

We have investigated what the impact is of entertaining this type of representation for our findings regarding degree membership at the aggregate level (Section 4.2) by assuming a logarithmic transformation of numerosity, instead of a linear one. 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 one's gender, or one's 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 because of practical reasons, such as the additional requirement of obtaining 8 x 41 x 20 judgments of pairwise similarity from 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 & 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

SV and PE were 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. The funding agency had no role in in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; or in the decision to submit the article for publication.

References

Alxatib, S., & Pelletier, F. J. (2011) The psychology of vagueness : Borderline cases and contradictions. *Mind & Language, 26*, 287-326.

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.

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.

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.

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, in press*.

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

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science, 4*, 427-455.

Bonini, N., Osherson, D., Viale, R., & Williamson, T. (1999). On the psychology of vague predicates. *Mind & Language, 14*, 377-393.

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).

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.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coltheart, V., & Evans, J. St. B. T. (1981). An investigation of semantic memory in individuals. *Memory & Cognition, 9*, 524-532.

Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science, 6*, 390-406.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Decock, L., & Douven, I. (2014). What is graded membership? *Noûs, 48*, 653-682.

Dehaene, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychological Science, 12*, 244-246.

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.

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition, 151,* 80-95.

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic, 42*, 137–160.

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2016). Measuring graded membership: The case of color. *Cognitive Science*. doi: 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.

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.

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

Egré, P. (2016). Vague judgment: A probabilistic account. *Synthese*, DOI 10.1007/s11229-016-1092-2.

Egré, P., & Barberousse, A. (2014). Borel on the heap. *Erkenntnis, 79*, 1043-1079.

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

Fara, D. G. (2008). Profiling interest relativity. *Analysis, 68*, 326-335.

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

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

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137-165.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*, 355-384.

Hampton, J. A., Aina, B., Andersson, J. M., Mirza, H., & Parmar, S. (2012). The Rumsfeld effect: The unknown unknown. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*, 340-355.

Hampton, J. A., Dubois, D., & Yeh, W. (2006). The effects of pragmatic context on classification in natural categories. *Memory & Cognition, 34*, 1431-1443.

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.

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*. doi:10.1007/s10988-016-9202-7

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

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.

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.

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry, 56*, 258-277.

Kennedy, C., & McNally, L. (2005). Scale structure and the semantic typology of gradable predicates. *Language, 81*, 345–381.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy, 4*, 1-45.

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.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 1-36.

Lee, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification, 19*, 69–85.

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

Luce, R. D. (1972). What sort of measurement is psychophysical measurement? *American Psychologist, 27*, 96-106.

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.

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.

McCloskey, M., & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition, 6*, 462-472.

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.

Medin, D. M., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*, 158-190.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation, 20*, 2597-2628.

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.

Rips, L. J., & Turnbull, W. (1980). How big is big? Relative and absolute properties in memory. *Cognition, 8*, 145-174.

Rosch, E. H. (1975a). Cognitive reference points. *Cognitive Psychology, 7*, 532-547.

Rosch, E. H. (1975b). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

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

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017). *Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives*. Manuscript submitted for publication.

Sæbø, K. J. (2009). Judgment ascriptions. *Linguistics and Philosophy, 32*, 327–352.

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. (2017). *Subjective meaning in gradable adjectives: The case of tall and heavy*. Manuscript submitted for publication.

[dataset] Verheyen, S., & Egré, P. (2017, March 28). Typicality and graded membership in dimensional adjectives: Data. Retrieved from osf.io/djkdg

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.

Verheyen, S., Van Deun, K., & Storms, G. (2017). *Typicality in conceptual space: Representations, variability, and limitations.* Unpublished manuscript.

von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics, 3*, 1–77.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: a typicality-based evaluation. *Psychonomic Bulletin & Review, 15* , 630-637.

White, A., Malt, B. C., Storms, G., & Verheyen, S. (2017). *Mind the generation gap: Differences between young and old in everyday lexical categories*. Unpublished manuscript.

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

Appendix

This section documents the information that can be found in Figure 2 and Tables 4-5 of Section 4.2 (Membership degree at the aggregated level) in the main article. Instead of a linear representation of numerosity, a logarithmic transformation is assumed.

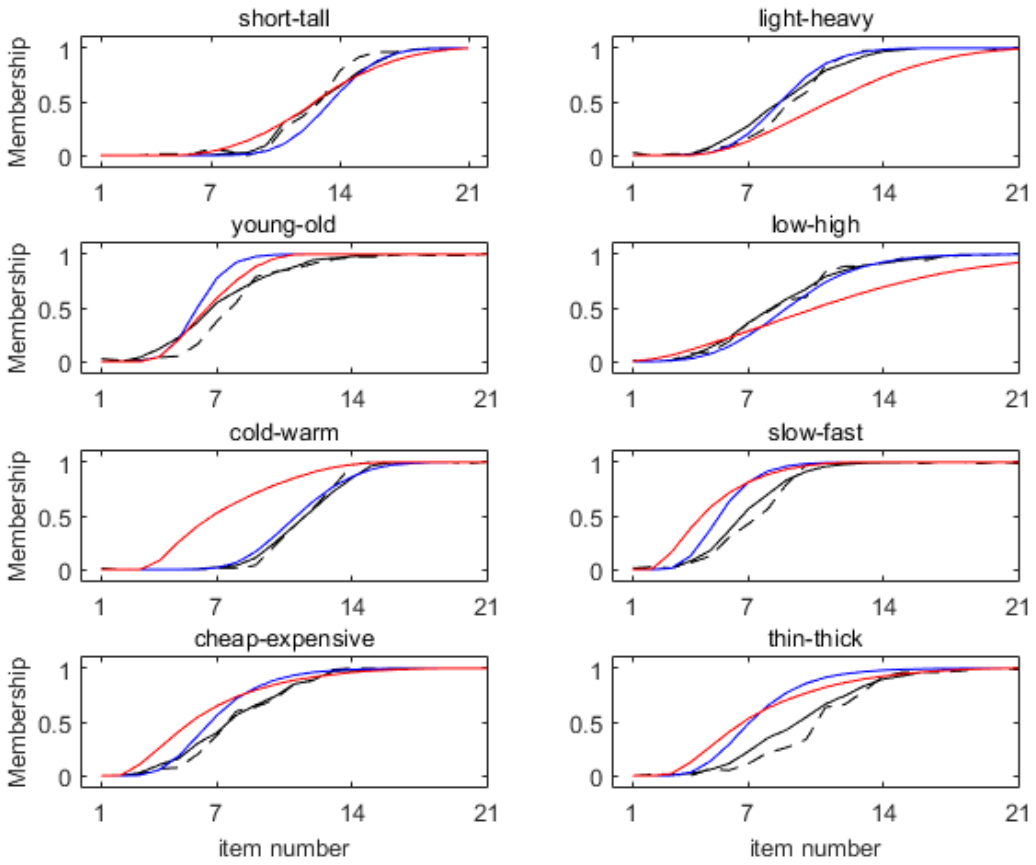

Figure A1. Observed degrees of membership (black) based on the binary (dotted) and the continuous (solid) categorization task. Predicted degrees of membership based on normal (blue) and uniform (red) sampling from the logtransformed generated typical values.

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 binary 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 between the observed membership degrees and between the observed and the predicted membership degrees.

Pair	Categorization data	Binary 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

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

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

Pair	PSE				Slope			
	Observed		Predicted		Observed		Predicted	
	Binary	Continuous	Normal	Uniform	Binary	Continuous	Normal	Uniform
short/tall	12,59	12,87	13,56	12,77	1,26	1,39	1,15	1,85
light/heavy	9,12	8,81	8,76	11,55	1,30	1,50	1,15	2,26
young/old	8,00	7,19	6,05	6,68	1,49	1,59	0,69	0,98
low/high	8,71	8,63	9,22	11,10	1,89	1,89	1,69	3,61
cold/warm	11,51	11,58	11,31	7,38	0,95	1,19	1,27	1,66
slow/fast	7,27	6,94	5,68	5,02	1,19	1,20	0,87	1,21
cheap/expensive	8,10	7,85	7,04	6,38	1,46	1,63	1,28	1,96
thin/thick	10,59	9,72	7,46	7,57	1,68	1,83	1,34	2,09

On average, the slopes of the observed membership degree curves are better predicted assuming a linear representation of numerosity than assuming a log-transformed representation (expressed as mean absolute deviation: MAD=0,24 vs. MAD=0,30 for binary categorization and MAD=0,24 vs. MAD=0,37 for continuous categorization). For binary 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).