Grounding a pragmatic theory of vagueness on experimental data: Semi-orders and Weber's Law

van Rooij, R.; Kochari, A.

DOI
10.31234/osf.io/v9xgq
10.1007/978-3-030-15931-3_9

Publication date
2019

Document Version
Author accepted manuscript

Published in
Vagueness and Rationality in Language Use and Cognition

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Grounding a Pragmatic Theory of Vagueness on Experimental Data: Semi-Orders and Weber’s Law

Robert van Rooi & Arnold Kochari*

Abstract

One of the traditional pragmatic approaches to vagueness suggests that there needs to be a significant gap between individuals or objects that can be described using a vague adjective like tall and those that cannot. In contrast, intuitively, an explicit comparative like taller does not require fulfillment of the gap requirement. Our starting point for this paper is the consideration that people cannot make precise measures under time pressure and their ability to discriminate approximate heights (or other values) obeys Weber’s law. We formulate and experimentally test three hypotheses relating to the difference between positive and comparative forms of the vague adjectives, gap requirement, and Weber’s law. In two experiments, participants judged appropriateness of usage of positive and comparative forms of vague adjectives in a sentence-picture verification task. Consequently, we review formal analysis of vagueness using weak orders and semi-orders and suggest adjustments based on the experimental results and properties of Weber’s law.

1 Introduction

Indirect versus direct comparisons. Consider the following figure, depicting the heights of John and Mary.

---

*Thanks to the Dutch NWO gravitation project Language in Interaction, the ESSENCE Marie Curie Initial Training Network, and the project ‘Logicas no-transitivas. Una nueva aproximacion a las paradojas’ for financially supporting this project. The authors also thank Nicole Standen-Mills for professional proof-reading.
According to Kennedy’s (2011) intuition, this picture allows us to say (1-a). At the same time, his intuition is that (1-b) is false.

(1) a. John is taller than Mary.
   b. Compared to Mary, John is tall, but compared to John, Mary is not tall.

Kennedy proposes that whereas for the truth of the explicit comparative (1-a) any (directly or indirectly observable) bit of John’s height that exceeds Mary’s suffices, there should be a significant gap between the heights of John and Mary for the implicit comparative (1-b) to be true. According to Kennedy (2011), such a significant gap is also required to make the positive sentence (2) true with respect to an implicit comparison class:

(2) John is tall.

As it turns out, Kennedy’s intuition concerning (2) fits well with a pragmatic approach towards clustering proposed by various philosophers to (dis)solve the Sorites paradox.

**The pragmatic gap requirement**  We can divide people into those that are tall and those that are not. How do we do so? What is the criterion? According to the most obvious solution, everybody is tall whose height is above a certain cutoff-point, where every bit of height suffices. But that gives rise to several questions: (i) How do we determine the cutoff-point? Is it the mean height of the relevant persons, the median height, or something else?¹ (ii) How does this proposal account for the intuition that some people are borderline cases of tall people? (iii) Doesn’t this presuppose that we need to be able to measure people’s height before we can classify people as tall, and isn’t that problematic? Suppose we have settled question (i). How, then to account for borderline cases? A natural proposal would have it that those individuals are borderline tall for which we cannot determine whether they are above or below the cutoff-point. This could be because our measurement is in many cases imprecise. But once we allow for this, we immediately get into

¹For experimental research on this, see Schmidt et al (2009) and Solt & Gotzner (2012). Solt & Gotzner (2012), for instance, show that classification is based on more than just ranking.
the well-known *Sorites problem*. In a *Sorites problem*, we have a long series of people ordered in terms of their height, where we assume that the difference in height between two subsequent persons is always very small. Now, if you decide that the first individual presented to you, the tallest, is tall, it seems only reasonable to judge the second individual to be tall as well, since there is only a miniscule difference in heights. But, then, by the same token, the third person must be tall as well, and so on indefinitely. In particular, this makes the last person also tall, which is a counterintuitive conclusion given that it is in contradiction with our intuition that this last, and shortest individual, is short, and thus not tall.

According to two pragmatic approaches towards vagueness, the problem of borderline cases is more of a theoretical than a practical problem. In practice—or so these pragmatic approaches claim—we are hardly ever confronted with the problem of borderline cases, because we have found pragmatic strategies to avoid them.

According to one such a pragmatic proposal, as suggested by Fara (2000) and foreshadowed by Tafel & Wilkes (1963), our clustering of classes into tall and non-tall objects can make use of any salient feature that distinguishes one group from the other, as long as this clustering is roughly consistent with a grouping based on height. One such salient feature in terms of which one group is distinguished from another might be the *spatial arrangement* of objects. Although this somewhat opportunistic pragmatic hypothesis may seem plausible, so far we have failed to find experimental evidence in support of this proposal.

To examine this proposal, we are testing in an ongoing series of experiments whether a specific arrangement of objects matters for the applicability of a vague adjective as a description of one of these objects. If it is indeed the case that any salient feature can be used for clustering for the purpose of vague adjective applicability, then spatial arrangement could be one good candidate for such a feature. It would predict that when the same object visually appears to be grouped with bigger objects, it should be classified as *big*, and when it visually appears to be grouped with smaller objects, it should be classified as *non-big*. In our experiments, we presented participants with 6 (or 11 in one of the experiments) objects standing next to each other. These objects were equally spaced from each other in terms of the relevant property (specifically, we used size, length, height, and width) and the target object was in the middle of this distribution. We created different conditions by manipulating the order in which the objects were arranged. The participants’ task was to decide whether the target object can be described as *tall/big/long/high*. In several experiments that we have conducted so far, the order did not matter for applicability of the vague adjective. Our intermediate results presented here
informally thus suggest that there is some kind of criterion of applicability of vague adjective which can not be overridden/substituted by the visual grouping of the objects.

A second pragmatic approach to clustering has explicitly been proposed as a solution to the Sorites paradox (Gaifman, 2010; Pagin 2010, 2011, Rayo, 2010, and van Rooij 2011b, arguably all based on Wittgenstein 1953). According to this hypothesis, we can appropriately classify some member of a group as being tall only if the clustering of the group based solely on height does not give rise to borderline cases. This is the case precisely if there is a significant gap in height between the shortest person classified as tall, on the one hand, and the tallest person of the other group, on the other. According to this gap hypothesis, $x$ cannot be classified as tall if there is another person $y$ not classified as tall, but still similar in height, $x \sim_T y$. This solution seems natural: we make a useful division of the set of all relevant objects into those that are tall and those that are not, and this division is easy if there exists a significant gap. This approach to clustering solves—or perhaps dissolves\(^2\)—the Sorites paradox, because the tolerance principle, according to which if one person is tall, and another has a very similar height, i.e., \(\forall x, y ((T(x) \land x \sim_T y) \rightarrow T(y))\), will not give rise to inconsistency.

**Analogue Magnitude Representation and Weber’s law**  In the discussion so far, it seems that we have implicitly assumed that we need to be able to measure people’s height in order to classify people as tall, or taller than another. But there is serious doubt that these and other classifications really require an ability to measure persons, objects, and events, in terms of meters, liters, hours and other units of precise measurements with which we are so familiar, making use of a mathematically defined system of numbers. Psychological evidence (starting from Moyer 1973) suggests that people represent these magnitudes in a more automatic, but imprecise (i.e., approximate), analogue way, and that we use them in this way, for instance, under time-pressure and other cognitive loaded circumstances.\(^3\) These so-called Analogue Magnitude Representations (AMRs) are taken to be primitive because in contrast to the standard way we think about measures, they are already present in infants, as well as in some other animals, such as primates, dogs, and birds.

One important feature of AMRs is that they obey *Weber’s law*, which holds

\(^2\)According to this approach, the Sorites paradox dissolves, because it does not even arise.

\(^3\)The different circumstances in which the different representations are used brings to mind the so-called Dual Process Theories of reasoning (cf. Evans, 2010, Kahneman, 2011), according to which there that are two distinct processing modes available for many cognitive tasks, one being automatic and fast, while the other is more controlled and slower. We will not dwell on this more general hypothesis in this paper, however.
that the ability to discriminate two magnitudes is determined by their ratio, i.e., that this ability grows with intensity (Weber 1834).\textsuperscript{4} According to this law, it is more difficult to discriminate two stimuli of high intensity than it is to discriminate two stimuli of low intensity, even though the absolute difference between the intensity of two high stimuli is exactly the same. As suggested above, AMRs exist for instance for numerical magnitudes, via our so-called ‘approximate number system’ (Dehaene, 2011). And indeed, experimental data (e.g. Dehaene, Dehaene 2011) confirms that under natural circumstances the approximate discrimination between the cardinality of two sets can only be done when they differ by a certain percentage. A natural hypothesis is that we use AMRs for all types of magnitudes, also including, for instance, length/height. But this also suggests that when we compare individuals in terms of their approximate height, our ability to distinguish these heights, at least when we can not measure them precisely, will obey Weber’s law. Applying what we know about AMRs and the principles of Weber’s law in terms of perception to vague adjectives, we would expect that there too, the higher the relative difference between objects, the more applicable a vague predicate would be as a description.\textsuperscript{5} This is in accordance with the gap hypothesis and can be seen as its extension.

**Three hypotheses** The above discussion of the difference between explicit versus implicit comparatives, of the requirement of a gap and the magnitude thereof for the use of positive sentences, based on the idea that we represent magnitudes in an analogue way, gives rise to the following three hypotheses for one-dimensional adjectives:

1. **Comparative hypothesis**: For an explicit comparative of the form ‘John is taller than Mary’ to be true, there does not have to be a significant gap in height between John and Mary, so any difference would suffice.

2. **Gap hypothesis**: The truthful use of a sentence with a positive adjective like ‘John is tall’ requires there to be a significant gap in height between John and other people within the set of relevant persons.

3. **Weber’s law hypothesis**: Acceptability of a sentence with a positive adjective like ‘John is tall’ will increase with increasing relative difference

\textsuperscript{4}Strictly speaking, Weber’s law is about relative changes in stimulus magnitude, rather than relative differences between two stimuli. Thus Weber’s law on its own only says that the perceived change in a magnitude is proportional to relative changes in that magnitude. In this paper we assume that the likelihood of detecting a change in magnitude is proportional to the amount in perceived change, so that Weber’s law as interpreted strictly comes down to our interpretation.

\textsuperscript{5}See (Égré, 2017) for similar reasoning.
between the heights of John and other people within the set of relevant persons.

In the following section we will discuss to what extent these hypotheses are confirmed by experimental evidence. After that, we will look into the formal analysis of vagueness in terms of weak orders and semi-orders and address challenges to it based on the experimental evidence and AMRs.

2 Experimental Evidence

In this section, we present the results of two experiments on how vague adjectives are used by people when they compare a number of objects. In order to test the hypotheses presented above, we collected data in a sentence-picture verification task. In this task, participants read sentences describing an object using a vague adjective. They then saw a picture depicting this object and several others. The task was to decide whether the sentence was suitable as a valid description of the target object. We manipulated the relative differences in size between the target and other objects within the same image and expected that this would affect the acceptance rates.

Each presented picture contained 4 objects. The target object was of a different color than the other 3 comparison objects. This color was used to refer to this object in the sentences. The target object was always the biggest/longest/tallest/highest one in the picture. The comparison objects were smaller than the target object. They were all approximately of the same size, with a small, but noticeable difference. We manipulated the size difference between the target object and the biggest of the comparison objects. The ratio between the target and the biggest comparison object was one of the following: 12:12 (same size), 11:12, 9:12 or 7:12.

Participants had to read and make judgments about the same pictures in two conditions: once for a sentence describing the target using a positive form of a vague adjective and once for a similar sentence describing the target using a comparative form of this adjective. By comparing acceptance rates for the comparative and positive forms, we were able to see if people have different criteria for them. In sum, we had the following design: adjective form (positive vs. comparative) X difference ratio (12:12, 11:12, 9:12 and 7:12).

We expected that sentences in combination with ratio 12:12 would never be accepted since it should be a factually false statement for positive and comparative forms. Based on our comparative hypothesis, we expected the sentence with a comparative form to be accepted whenever participants saw the difference at all, i.e. in all ratios except for 12:12. Based on our gap hypothesis and Weber’s law hypothesis, we expected the acceptance rates of
sentences with positive forms of adjectives to be different for different ratios since a small gap may not be sufficient. Thus, the positive form of a vague adjective should be more acceptable when the difference ratio is 7:12 or 9:12 than when it is 11:12. Ideally, we should observe the following pattern of acceptance rates: 11:12<9:12<7:12.

Besides the acceptance rates, we also collected reaction times, i.e. the amount of time participants took to make a decision. Reaction times (RTs) are conventionally considered to reflect the difficulty/complexity of the decision being made. Consequently, if there are more processes taking place or there is more hesitation in a trial, the reaction time for this trial should be longer. Thus, based on our hypotheses, for positive-form judgment, in case of lower ratios the decision should take longer, because a positive vague adjective is less applicable there. The reaction times should thus form the following pattern 11:12>9:12>7:12. In addition, based on our comparative hypothesis we expected that there would be a difference in reaction times of acceptance between comparative and positive adjectives. Positive adjectives should take longer because participants need not only to visually register the difference, but also make sure that whatever criterion they use for applicability of positive adjectives is met.

From our pilot experiments we noticed that intermixing sentences with positive and comparative forms of the adjectives can confuse participants, so we decided to first present all the sentences with positive adjectives (where participants can stick to their idea of what tall, high, etc., means) and then present all sentences with comparative adjectives. During the first half of the experiment, the participants did not know that they would be judging comparative forms in the second half, and they were never told that they would see the same pictures twice.

2.1 Experiment 1

Materials We used 4 different adjectives and there were 12 different objects in total. The following adjective and object pairs were used: big: box, apple, square, circle; long: pencil, line, drinking straw; tall: bar, can, building, street lamp; high: stool. The colors were chosen arbitrarily, though in such a way that they would not be confused with each other based on the subjective judgment of the experimenter.

The sentences with positive forms of adjectives were always constructed in the following way: ‘The [target object color] [target object] is [adjective in positive form].’, e.g. ‘The grey box is big.’ Sentences with the comparative forms were constructed in the following way: ‘The [target object color] [target object] is [adjective in comparative form] than the [other color] ones’, e.g.
"The grey box is bigger than the orange ones." The pictures used with these sentences are given as examples in Figure 1.

Each of the twelve objects was presented in four ratios and in two adjective forms, so we had $12 \times 2 \times 4 = 96$ experimental trials in total. In addition to the experimental trials, we added 8 filler trials, each of which was also presented once in the positive and once in the comparative form, so there were 16 additional trials. These filler trials used different objects, but were constructed in such a way that they looked exactly like the experimental ones except that the answer was always ‘no’, because the target object was one of the smaller ones and thus the statements were factually wrong. The fillers were intermixed with the other trials and should not have been noticed as different.

A detailed description of the experimental procedure is provided in Appendix A.

![Figure 1: Examples of stimuli used in Experiment 1. From top to bottom, ratios: 7:12, 9:12, 11:12, 12:12. Sentences: 'The grey/orange box is big' or 'The grey/orange box is bigger than the orange/grey ones'. Orange color here printed as dark grey.](image)

**Participants** Participants were recruited via Prolific.ac, an online platform aimed at connecting researchers and participants willing to fill in surveys and questionnaires in exchange for compensation for their time. We recruited native American English speakers. Twenty-four participants completed the task in order and were included in the analyses reported below. The detailed eligibility and exclusion criteria are provided in Appendix A.
Results  Each of the 24 participants completed 112 trials, for a total of 2688 trials. We excluded all trials to which participants did not give a response (missing responses, N=39) and all trials with RTs lower than 250 ms (N=1) since it would be impossible to make a decision and to send a signal to execute the motor command (press the button) in such a short time. So in total we excluded 40 trials, or 1.5% of the total number of trials.

The acceptance rates for each ratio per sentence type are presented in Table 1 below. There seems to be no difference between the sentences with positive and comparative adjectives except for the ratio 12:12. In this latter ratio the target object was not the biggest anymore and, while 37% of participants still agreed that it could be referred to as *big/tall* etc, no participants agreed with the false statement that it was *bigger/taller* than other objects.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Sentence type</th>
<th>Total N trials</th>
<th>% of YES responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:12</td>
<td>positive</td>
<td>285</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>285</td>
<td>100</td>
</tr>
<tr>
<td>9:12</td>
<td>positive</td>
<td>287</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>286</td>
<td>100</td>
</tr>
<tr>
<td>11:12</td>
<td>positive</td>
<td>283</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>287</td>
<td>96</td>
</tr>
<tr>
<td>12:12</td>
<td>positive</td>
<td>272</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>282</td>
<td>0</td>
</tr>
<tr>
<td>Fillers</td>
<td>positive</td>
<td>190</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>191</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Mean acceptance rates per ratio and sentence type, Experiment 1.

Besides the acceptance rates, we also expected to see differences in reaction times between sentences with different forms of vague adjectives. The reaction times per ratio and sentence type are shown in Figure 2. In this and further reaction time analyses, we only look at reaction times to ‘yes’ responses, because ‘no’ responses are the result of a different decision and are not conventionally analyzed together with ‘yes’ responses. We can see that there is no difference between positive and comparative sentences except for the ratio 11:12, for which comparative sentence verification took slightly longer.

Discussion  In this experiment, there was no difference between how participants judged the applicability of positive and comparative forms of vague adjectives for any of the ratios in which we were interested. For any ratio except 12:12, the participants agreed that both the positive form and the comparative form were applicable. There was no difference between individual participants and adjectives either. This was an unexpected result for us.
Figure 2: Mean RTs per ratio and sentence type for ‘yes’ responses, Experiment 1. The value for the comparative sentences in ratio 12:12 is not shown since there was only one data-point. The error bars represent the standard error value.

One case in which we would not see any difference between adjective forms would be if participants took the positive forms tall, long, big, high to be acceptable when the target object was simply the biggest/longest etc. If they interpreted a sentence like ‘The grey box is big’ as meaning that the grey box is the biggest one, then indeed, we would expect a ‘yes’ response in all trials. In order to see if this would be the case, we conducted another experiment where everything was kept the same except that the target object was no longer the biggest/tallest/highest/longest, so the superlative interpretation would no longer be possible.

2.2 Experiment 2

In this experiment, we tested the same hypotheses as in Experiment 1, using the same set-up, but we added one more object to each of the pictures. If the positive forms were indeed interpreted as referring to the biggest, etc., object, then by making the target object not the biggest in the group of objects to be judged, we expected to obtain different results, potentially agreeing with our hypothesis.

Materials We added one more object to each picture presented in this experiment. This new object was bigger than the target object, but of the same color as the comparison objects. Specifically, it was always 110% of the size of
the target object. Examples of altered stimuli can be seen in Figure 3.

In addition, we also had to change the formulation of the sentences in the second block where participants saw comparative forms of the adjectives. The new formulation was ‘The [target object color] [target object] is [adjective in comparative form] than the [the other color] ones on its two sides’, e.g. ‘The grey box is bigger than the orange ones on its two sides.’ The formulation had to be such in order for the sentences to still test whether the participants can visually register the difference. In order for this sentence to be true, the target object was never standing next to the newly added biggest object.

![Figure 3: Examples of stimuli used in Experiment 2. From top to bottom, ratios: 7:12, 9:12, 11:12, 12:12. Sentences: ‘The grey/orange box is big’ or ‘The grey/orange box is bigger than the orange/grey ones on its two sides’. Orange color here printed as dark grey.](image)

**Participants** The same eligibility requirements as in Experiment 1 were applied to participants in this experiment. Thirty-one participant completed the task in order and was included in the analyses. See Appendix A for details.

**Results** Each of the 31 participants included in the analysis did 112 trials, for a total of 3472 trials. We excluded all trials to which participants did not give a response (N=58) and trials with RTs lower than 250 ms (N=2). In total we excluded 60 trials or 1.73% of the total number of trials.

Data analyses reported here were conducted in the R environment (R Development Core Team, 2016). In order to test effects for significance, we ran mixed-effects models as implemented in the lme4 package (Bates, Machler, Bolker, & Walker 2015).
The acceptance rates per ratio and sentence type are presented in Table 2. There seems to be no difference between positive and comparative sentences in ratio 12:12. The difference is present for other ratios and forms the following pattern: 11:12<9:12=7:12. We modeled the acceptance likelihood using a logit mixed-effects model, which included the main effects of sentence type and of ratio as predictors (fixed effects) and random intercepts for subject and for depicted object (random effects). There was a significant main effect of sentence type ($\beta=-3.06$, $SE=0.15$, $z=-20.82$, $p<0.001$). We then modeled the ratio effect on positive and comparative sentences separately (still including random intercepts for subject and depicted object). In sentences with a positive form of the adjective, ratio 9:12 did not differ from ratio 7:12 ($\beta=0.002$, $SE=0.21$, $z=0.01$, $p=0.99$); 11:12 did differ from 7:12 ($\beta=-1.20$, $SE=0.21$, $z=-5.66$, $p<0.001$); and 12:12 also differed ($\beta=-3.60$, $SE=0.27$, $z=-13.36$, $p<0.001$). So the pattern 11:12<9:12=7:12 for positive sentences is supported by the significance levels. In the model that included comparative sentences only, only ratio 12:12 significantly differed from 7:12 ($\beta=-7.18$, $SE=0.47$, $z=-15.28$, $p<0.001$).

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Sentence type</th>
<th>Total N trials</th>
<th>% of YES responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:12</td>
<td>positive</td>
<td>365</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>370</td>
<td>96</td>
</tr>
<tr>
<td>9:12</td>
<td>positive</td>
<td>366</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>369</td>
<td>96</td>
</tr>
<tr>
<td>11:12</td>
<td>positive</td>
<td>365</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>369</td>
<td>93</td>
</tr>
<tr>
<td>12:12</td>
<td>positive</td>
<td>362</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>356</td>
<td>8</td>
</tr>
<tr>
<td>Fillers</td>
<td>positive</td>
<td>245</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>245</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Mean acceptance rates per ratio and sentence type, Experiment 2.

Next, we looked at the differences between ratios and sentence types in terms of RTs. The mean RTs in ‘yes’ responses for each ratio are presented in Figure 4. There seems to be a difference in RTs between sentence types in all ratios except for 12:12. In all cases, evaluating sentences with positive adjectives took longer than sentences with comparative adjectives. Moreover, the RTs increase with decrease in ratio sizes. In order to test for the statistical significance of these differences, we modeled RTs using a linear mixed-effects model. Here too, we included the main effects of ratio and sentence type as well as random intercepts for subject and depicted object. The p-values were calculated based on the Satterthwaite approximation for denominator de-
degrees of freedom as implemented in the package `lmerTest` (Kuznetsova, Brockhoff, & Christensen 2016). There was a main effect of sentence type ($\beta=155$, $SE=15.85$, $t=9.83$, $p<0.001$). The RTs in ratio 9:12 did not significantly differ from ratio 7:12 ($\beta=21.02$, $SE=17.14$, $t=1.23$, $p=0.22$), whereas 11:12 did differ from 7:12 ($\beta=120.5$, $SE=17.8$, $t=6.77$, $p<0.001$) and 12:12 also did ($\beta=297.52$, $SE=37.7$, $t=7.89$, $p<0.001$). Thus, the pattern 11:12<$9:12$=7:12 that we saw with acceptance rates is also present in RTs, but is valid for both positive and comparative sentences.

Figure 4: Mean RTs per ratio and sentence type for ‘yes’ responses, Experiment 2. The error bars represent the standard error value.

In order to examine whether all participants showed the same pattern of responses, we looked at inter-individual differences in acceptance rates. We calculated the acceptance rates for each sentence type in ratios 7:12, 9:12 and 11:12 combined. Ratio 12:12 was excluded since it is not of interest for our hypotheses. The resulting acceptance rates for each participant are presented in Figure 5. We see that while some of the participants exhibit the effect that we expected and which we can see in the overall acceptance rates, other participants always responded ‘yes’ in both types of sentences, and others still never accepted a positive form and almost always accepted a comparative. In order to make sure that the effects that we obtained were not simply a result of averaging, we did the analyses again excluding all subjects for whom the difference between acceptance of positive and comparative forms was more than 90% and less than 10%. This resulted in an exclusion of 17 participants, or more than half. We call the remaining 14 subjects *sensitive*.

The acceptance rates for these 14 *sensitive* subjects are presented in Table 3. The pattern is the same as the overall average. The mean RTs are presented
in Figure 6. Here too, the pattern is the same as for the average of all subjects. We re-ran both models, looking at the acceptance rates and RTs, with these 14 subjects only. The pattern of the significant effects was the same.

Finally, we looked at the differences in acceptance rates between adjectives. We included all subjects in this analysis. We calculated the mean acceptance rate in all ratios except for 12:12. The resulting means are presented in Table 4. All adjectives seem to show the same pattern of acceptance rates.

Discussion As opposed to Experiment 1, here participants judged the applicability of positive forms of the vague adjectives and comparative forms differently. It appears that our suspicion that subjects interpreted positive forms as superlatives in Experiment 1 was correct.

We predicted that we would see the following acceptance rates for different ratios: 11:12<9:12<7:12; we also predicted the following RTs: 11:12>9:12>7:12. This prediction was not fully confirmed, but we did find partial support for our hypotheses: while 7:12 and 9:12 did not significantly differ from each other, they did differ from 11:12. A small gap in the case of ratio 11:12 was indeed less acceptable than the large gaps present in 9:12 and 7:12. One might have expected that in the ratio 11:12, the acceptance rate of the positive forms
Table 3: Mean acceptance rates per ratio and sentence type, sensitive group of subjects of Experiment 2.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Sentence type</th>
<th>Total N trials</th>
<th>% of YES responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:12</td>
<td>positive</td>
<td>165</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>168</td>
<td>98</td>
</tr>
<tr>
<td>9:12</td>
<td>positive</td>
<td>166</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>166</td>
<td>99</td>
</tr>
<tr>
<td>11:12</td>
<td>positive</td>
<td>162</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>166</td>
<td>96</td>
</tr>
<tr>
<td>12:12</td>
<td>positive</td>
<td>163</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>163</td>
<td>8</td>
</tr>
<tr>
<td>Fillers</td>
<td>positive</td>
<td>112</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>111</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 6: Mean RTs per ratio and sentence type for ‘yes’ responses, sensitive group of subjects of Experiment 2. The error bars represent the standard error value.

would be close to 0 for the gap hypothesis to be supported. We did not observe this. However, given that a considerable variability between subjects, and potentially between different depicted objects, is to be expected, we see a lower acceptance rate in ratio 11:12 than in larger ratios as strong enough evidence.

Related to the previous point, in the experiment planning stages, we had tentatively expected to observe the same acceptance rate for both positive and comparative forms in ratio 7:12, since the differences between the target and
comparison objects was relatively large there. However, the mean acceptance rates for the positive form were never higher than 54% (or 67% in the sensitive group). One possible explanation for this is that perhaps the difference was still not large enough to satisfy our participants’ criteria. However, although the differences are not statistically significant, they do go in the expected direction in the case of the RTs: participants were on average faster at giving ‘yes’ responses in the ratio 7:12 than in the ratio 9:12. This difference was small overall, but 52 ms. for the sensitive group of subjects. A more careful inspection with more experimental trials and participants would be needed to make a definite conclusion about the RT differences there and why we did not see difference between these ratios in acceptance rates.

We obtained the predicted difference in RTs between the judgments of positive and comparative forms in our data. Giving a ‘yes’ response in the positive adjective judgment took longer than in the comparative form judgment. We take this as evidence that more intensive processing takes place when determining whether a positive form can indeed be applied. The additional process that needs to take place, we assume, is verifying that some criterion is met. Potentially, this criterion might be whether the difference is large enough.

Another aspect that came as a surprise and deserves closer inspection is inter-individual differences between our subjects. Only 14 out of 31 subjects that we tested exhibited the pattern of effects that we expected. For some of the subjects, there was no difference between comparative and positive forms, i.e. they always accepted both. We cannot say that these subjects simply pressed ‘yes’ in every trial, because they all gave correct ‘no’ responses to filler trials that we included. Thus, it seems like they interpreted the positive form as meaning something like ‘one of the bigger ones’ and therefore it was always true. Another group of participants never agreed that a vague adjective can be used as a valid description of our target object. They did always see

<table>
<thead>
<tr>
<th>Adjective</th>
<th>Sentence type</th>
<th>Total N trials</th>
<th>% of YES responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>big</td>
<td>positive</td>
<td>366</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>370</td>
<td>95</td>
</tr>
<tr>
<td>long</td>
<td>positive</td>
<td>276</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>277</td>
<td>95</td>
</tr>
<tr>
<td>tall</td>
<td>positive</td>
<td>364</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>369</td>
<td>95</td>
</tr>
<tr>
<td>high</td>
<td>positive</td>
<td>90</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>comparative</td>
<td>92</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 4: Acceptance rates per adjective and sentence type, excluding ratio 12:12. Experiment 2.
a difference, because they agreed with the applicability of the comparative forms, so it cannot be that they simply did not notice size differences. Rather, it seems like they had some criterion that was never met in these experimental conditions. Potentially, they may have thought that a positive form can only be applicable when the object is the biggest/tallest etc., which was never the case in these pictures. However, this is just our speculation at this point. It would need to be tested in future studies.

It is also worth noting that the differences between both acceptance rates and the RTs for both positive forms and comparative forms are more pronounced in the group of sensitive subjects. The difference between RTs to positive and comparative sentences was around 200 ms for this group, whereas it was around 150 ms overall. These subjects seemed to have put extra effort into making a decision about the applicability of the positive form of the adjectives and that is why the expected difference is visible here.

2.3 Interpretation of the Experimental Data

We conducted two experiments that were the same except for one aspect and obtained very different results. Since the only difference between Experiment 2 and Experiment 1 was that we added a new object to the comparison objects in the pictures, we are inclined to conclude that in Experiment 1, participants thought it was acceptable to use a positive form because it referred to the biggest/tallest/longest/highest object on the screen (of course, for a definite conclusion an experiment directly comparing two situations would need to be conducted). In addition, there might have been a group of participants in Experiment 2 that had a strict requirement that the positive form can only be applied to the biggest/tallest, etc., objects. Overall, our evidence so far thus suggests that, at least in the set-up we used, the positive form of a vague adjective can be used as a description of the biggest/tallest etc. In other words, a gap between the target object and other objects is not necessary for the validity of a vague adjective if the target object is already the biggest.

In Experiment 2, we obtained evidence that there is a difference in the meaning of comparative and positive forms of the vague adjective. The comparative form was accepted in all ratios except for 12:12, so whenever there was any difference at all between the target and the comparison objects. This supports our comparative hypothesis. Moreover, we saw that judging the applicability of the positive form took longer than judging the applicability of the comparative form, so there is a difference also in the processing that took place in order to judge the applicability of two forms of the adjective.

The mean acceptance rate for the positive form of the vague adjective for the ratio 11:12 was lower than for the ratios 9:12 and 7:12. A parallel pattern
of differences was obtained with reaction times. This supports the prediction of the gap hypothesis, i.e. the difference between the target object and other objects has to be significant in order for the positive form to be justified. At the same time, this pattern supports the Weber’s law hypothesis: that with increasing relative differences in size/length etc., the applicability of the positive form of the vague adjective will increase.

3 Modeling by Semi-Orders and Challenges

In Section 1, we formulated three hypotheses, which were put to the test using the experiments described in the previous section. The results that we obtained were consistent with the first and third hypotheses, but they were unexpected under the second hypothesis. In this section we will discuss whether we can account for the data making use of semi-orders.

3.1 Semi-Orders and Indirect Versus Direct Comparisons

In van Rooij (2011a), it is proposed that the difference between explicit and implicit comparatives should be captured in terms of the difference between weak orders and semi-orders.\footnote{See Solt (2016) for a recent use of semi-orders to account for the difference between the meanings of ‘most’ and ‘more than half’.

6} A structure \( \langle X, >_P \rangle \), with \( >_P \) a binary relation on \( X \), is a strict weak order just in case \( >_P \) is irreflexive (IR), transitive (TR), and almost connected (AC):

\begin{definition}
A strict weak order is a structure \( \langle X, >_P \rangle \), with \( >_P \) a binary relation on \( X \) that satisfies the following conditions:

\begin{enumerate}
  \item [(IR)] \( \forall x : \neg (x >_P x) \).
  \item [(TR)] \( \forall x, y, z : (x >_P y \land y >_P z) \rightarrow x >_P z \).
  \item [(AC)] \( \forall x, y, z : x >_P y \rightarrow (x >_P z \lor z >_P y) \).
\end{enumerate}
\end{definition}

It is well-known (e.g. Krantz et al, 1971) that strict weak orders can be represented numerically by a real valued measurement function \( f_P \) as suggested above: for all \( x, y \in X \): \( x >_P y \) iff \( f_P(x) > f_P(y) \).\footnote{This is true in general only for countable sets. Things are more complicated otherwise.} If we now define the indifference relation, ‘\( \approx_P \)’, as follows: \( x \approx_P y \) iff \( _{def} \) neither \( x >_P y \) nor \( y >_P x \), it is clear that \( x \approx_P y \) iff \( f_P(x) = f_P(y) \). It follows immediately that the similarity relation is not only reflexive and symmetric, but also transitive: if \( x \approx_P y \) and \( y \approx_P z \), then \( x \approx_P z \) holds as well. Thus, the indifference relation defined for (strict) weak orders is an equivalence relation.
In the linguistic literature (e.g. Kennedy, 1999), comparisons based on adjectives like ‘hot’, ‘tall’, and ‘loud’ are normally based on exact measurement functions like our \( f_P \) above, which give rise to weak orderings. However, there are several phenomena for which a weak ordering seems already too strong. First, this is due to the phenomenon of multi-dimensional adjectives, which, arguably, give rise only to a (strict) partial order, i.e. an order that is just irreflexive and transitive. Second, and more relevantly for this paper, a weak order is too strong for phenomena related to vagueness. Because a weak order gives rise to an indifference relation that is transitive, we get into a Sorites problem:

A person may be indifferent between 100 and 101 grains of sugar in his coffee, indifferent between 101 and 102, ..., and indifferent between 4999 and 5000. If indifference were transitive he would be indifferent between 100 and 5000 grains, and this is probably false. (Luce, 1956)

To represent orders involving vagueness, the comparative ‘\( x \succ_P y \)’ should intuitively mean that \( x \) is significantly or noticeably greater than \( y \). Luce (1956) introduced orderings that capture exactly this, semi-orders. A structure \( \langle X, \succ_P \rangle \), with \( \succ_P \) a binary relation on \( X \), is a semi-order just in case \( \succ_P \) is irreflexive (IR), satisfies the interval-order (IO) condition, and is semitransitive (STr).

**Definition 2** A semi-order is a structure \( \langle X, \succ_P \rangle \), with \( \succ_P \) a binary relation on \( X \) that satisfies the following conditions:

- (IR) \( \forall x : \neg(x \succ_P x) \).
- (IO) \( \forall x, y, w : (x \succ_P y \land y \succ_P w) \rightarrow (x \succ_P w \lor v \succ_P y) \).
- (STr) \( \forall x, y, z : (x \succ_P y \land y \succ_P z) \rightarrow (x \succ_P v \lor v \succ_P z) \).

Although semi-orders cannot be adequately represented by measure functions that assign to each member of \( X \) just a single number, they can be represented by intervals. Scott & Suppes (1958) prove the following representation theorem for semi-orders:

**Fact 1** Let \( X \) be a finite set and \( \epsilon > 0 \). Then \( \langle X, \succ_P \rangle \) is a semi-order iff there exists a real-valued function \( g_P \) on \( X \) such that for all \( x, y \in X \): \( x \succ_P y \) iff \( g_P(x) > g_P(y) + \epsilon \).

To illustrate, suppose we have a semi-order \( \langle X, \succ_P \rangle \) with \( x \sim_P y \sim_P z \sim_P v \) but \( x \succ_P z, y \succ_P v \) and \( x \succ_P v \). If we assume that \( \epsilon = 1 \), this can be modelled by the following function \( g \): \( g_P(x) = 4, g_P(y) = 3.2, g_P(z) = 2.4, g_P(v) = 1.6 \).
As for weak orders, for semi-orders too we can define a similarity relation \( \sim_P \) in terms of ‘\( \succ_P \)’ as follows: \( x \sim_P y \) iff neither \( x \succ_P y \) nor \( y \succ_P x \). The relation ‘\( \sim_P \)’ is reflexive and symmetric, but need not be transitive. Thus, ‘\( \sim_P \)’ does not give rise to an equivalence relation. And this is required to solve the Sorites problem: small differences can be neglected, but many small differences together add up to significant differences, and these cannot be neglected. Measure theoretically ‘\( x \sim_P y \)’ is true iff the difference in height between \( x \) and \( y \) is less than \( \epsilon \). In case \( \epsilon = 0 \), the semi-order is a weak order.

As already suggested above, weak orders are very natural representations of standard explicit comparatives like (1-a). In van Rooij (2011a) it was proposed that the semi-order relation ‘significantly taller than’, i.e. ‘\( \succ_T \)’, is what is relevant to evaluate the truth of implicit comparatives like (1-b). Thus, (1-b) is true just in case John is significantly taller than Mary. This immediately explains why (1-a) can be inferred from (1-b), but not the other way around.

### 3.2 Challenges for the Semi-Order Approach

#### The pragmatic gap requirement
The gap hypothesis seems confirmed by the results of Experiment 2 as described above. Given that with a semi-ordering we can think of \( x \succ_{\text{tall}} y \) as true just in case \( x \) is significantly taller than \( y \), it seems we can naturally account for the pragmatic gap requirement involving clustering a group into the tall ones versus the rest.

However, the results of Experiment 1 suggest that the hypothesis that a significant gap is a necessary condition for dividing groups is too strong.\(^8\) even if there is no clear demarcation between the bigger and the smaller persons of the domain, the tallest person can still be called ‘tall’.\(^9\) Thus, the gap hypothesis seems not to allow for such exceptional situations, and the existence of a significant gap does not seem necessary. But once we are in this situation one wonders how to address the Sorites paradox.

#### Semi-orders and Weber’s law
Another challenge is how the semi-order approach can account for Weber’s law. In the introduction we argued that under time-pressure people represent magnitudes at least also in terms of so-called Analogue Magnitude Representations (AMRs). Fults (2011) has argued

\(^8\)There is, of course, another problem as well: it seems that the existence of a significant gap is not sufficient either. Think of the case where an ordered set of objects contains more than one gap with respect to height. Which gap should be used to distinguish the tall ones from the rest?

\(^9\)Since it was not an a-priori hypothesis that tested here, but a conclusion we came to based on looking at the data, this claim will need to be explicitly tested in a separate experiment in the future.
that this is relevant for how to interpret natural language comparatives like (1-a) and (1-b) and positives like (2). He claims that as a result of our use of AMRs for the analysis of such sentences, we can explain why there are no sharp cut-off points for the application of a predicate like ‘tall’ and that its application gives rise to variable uncertainty. Moreover, it explains immediately why vague predicates obey Weber’s law, and why what is a sufficient gap between two magnitudes is determined by their ratio and grows with intensity.

In van Rooij (2011a), semi-orders were argued to be important for the analysis of implicit comparatives and of positive sentences involving vague predicates. Based on the above hypothesis, however, Fults (2011) comes up with a number of challenges for the approach based on semi-orders whose numerical representation (via Fact 1) is based on a precise and fixed threshold $\epsilon$.

1. $\epsilon$ is too sharp a cutoff point: analogue magnitudes and vagueness do not have a sharp points of discrimination (the problem of higher-order vagueness).

2. semi-order are incapable of providing an account of variable uncertainty, since again, two values $f_P(x), f_P(y)$ are either within $\epsilon$ to each other or not; there are no intermediate levels.

3. $\epsilon$ is fixed, while analogue magnitudes and vagueness show scalar variabilities. If semi-orders are to be used for either, $\epsilon$ must change based on the value of the magnitudes being compared.

We would like to suggest that these three critiques involve two different intuitions: the first two critiques involve the sharpness of the threshold $\epsilon$, while the third critique involves, rather, the fixed character of $\epsilon$. We will respond to these critiques by claiming that both intuitions are correct, but that they can still be captured in terms of semi-orders. In section 4.2 we will show that for orderings to be semi-orders, cutoff points do not have to be sharp, but can also be probabilistic in nature. This will deal with the first two critiques. In the same section we will show that orderings where the threshold increases with increasing magnitude can be modelled in terms of semi-orders as well.

A final problem concerns the interpretation of the threshold. The representation theorem for semi-orders makes use of a threshold, $\epsilon$, which we interpreted as a significant gap. The most natural interpretation of this threshold,

\footnote{According to proponents of AMRs, these representations are associated with arithmetic computations, including comparison, addition, subtraction, multiplication, and division. If so, one can define another challenge for our representation based on semi-orders: how to account for this? We will ignore this challenge in this paper, but see Solt (2016) for a suggestion of how to meet this challenge.}
however, is as a \textit{just noticeable difference}, abbreviated by JND, or \textit{discrimination threshold}. It is also this interpretation of the threshold which explains why magnitudes observe Weber’s law. But one wonders whether the notions ‘significant gap’ and JND can be identified. According to Kennedy’s (2011) intuition, for instance, an explicit comparative can be true, though the corresponding implicit comparison is false. But if semi-orders with a threshold $\epsilon$ are used to account for the falsity of the implicit comparison, it is hard to explain why the explicit comparison can still be true, when $\epsilon$ is interpreted as the JND. Is there any relation between our interpretation of $\epsilon$ and the standard one as a JND?

4 Answering the Challenges

4.1 Revisiting the Gap Hypothesis

The results of Experiment 2 show that the \textit{gap hypothesis} is quite natural. The results of Experiment 1 suggest, however, that the existence of a gap is not a necessary condition for the appropriate use of sentences like ‘John is tall’: also in case there is no gap, the largest individual is still counted as tall. Of course, one could conclude from Experiment 1 that the \textit{gap hypothesis} should thus be rejected completely. However, we still feel that the existence of gap for the appropriate use of predicates is still a natural \textit{default}. How can we account for such a default?

In Cobreros et al (in press) a non-monotonic consequence relation $\models_{PrPr}$ is introduced to account for the intuition that although normally we conclude $\text{Tall}(y)$ from premisses $\text{Tall}(x)$ and $x \sim_{\text{Tall}} y$, this conclusion can be cancelled in case additional information is provided (in this case that $x$ is a borderline case of a tall person). According to this logic, it follows by default that in case $x$ is considered tall and $y$ is not, $y$ is not considered to be similarly tall as $x$, i.e., $x \not\sim_{\text{Tall}} y$. Interpreting the similarity relation as the indifference relation of a semi-order, this means that there is a significant gap in terms of height between $x$ and $y$. More generally, if we have a sequence of individuals $x_1, \cdots, x_n$, the first of which is obviously tall and the last obviously not, this analysis predicts that, by default, somewhere in the sequence there is a gap, i.e., a pair $x_i, x_j$ such that $x_i \not\sim_{Pr} x_j$. We take this to be in accordance with the \textit{gap hypothesis}. However, just as before, this default inference can be cancelled. Also when it is explicitly given as premisses that the first individual in the sequence is tall and the last one is not, it is still allowed by the logic that $x_1 \sim_{\text{Tall}} x_2 \land \cdots \land x_{n-1} \sim_{\text{Tall}} x_n$. As such, the Sorites problem is (dis)solved. Moreover, it is possible in this logic that there exists an $i$, $1 \leq i \leq n$, such that
∀j : i < j ≤ n : x_j is not counted as tall. It could be, for instance, that this ith individual is x_2, as observed in Experiment 1. Thus, this logic not only captures the idea that the use of a predicate like ‘tall’ normally comes with a gap, it is also consistent with the results of Experiment 1 that only the tallest individual in the sequence is classified as tall. This logic does not predict the results of Experiment 1, however, for the logic also allows any other individual in the sequence to be the last tall individual. At this point we don’t know yet whether this means that the logic |=PrPrPr is thus too liberal, or whether people find it indeed quite arbitrary which individual is the last tall person. For now, we leave this as it is, but only conclude that the results of Experiment 1 are not in contradiction with the gap hypothesis, at least once one weakens this principle—expressible as (Tall(x) ∧ ¬Tall(y)) → x ∼_{Tall} y—to a default.

4.2 Meeting Fults’ Objections

Cutoff-points and uncertainty On the intuitive understanding of the numerical representation of semi-orders, the threshold ε represents uncertainty as to whether x is A-er than y or not. What the standard numerical representation does not capture, however, is that this uncertainty is, in fact, not equally divided over ε, when the latter is thought of as an interval. What one intuitively wants to capture to account for the first two critiques of Fults (2011) is that the uncertainty peaks at the centre of the interval, and fades away at the edges. Uncertainty is naturally represented in terms of probability, and we propose to answer Fults’ first two critiques in terms of probabilistic choice models.

There are typically two interpretations for axioms that define orders, such as asymmetry and almost connectedness. On a normative approach, these axioms define what it means to be rational: an individual who violates the maxims is acting irrationally. This is typically assumed in decision theory, where an ordering is used to model preference. On a descriptive approach, on the other hand, the axioms describe testable conditions. Taking a descriptive approach, we see that individuals, or groups, often show inconsistent behavior. For instance, a subject might at one time find that sound x is louder than sound y, while at other times (s)he might make the opposite judgement. Or it might be that some agents of a group judge sound x as louder than sound y, while other agents make the reverse judgement. Still, even in these cases the agent, or group, might behave probabilistically consistent. Let P_{Axy} be the

\[11\] In the logic |=PrPrPr, be borderline tall is modelled by saying that you are both tall and not tall. Although Cobreros et al (in press) claim that there is good empirical motivation for this modeling of borderline cases, this particular modeling of borderline cases is not required for a non-monotone logic like |=PrPrPr to work.
frequency with which an agent considers \( x \) \( A \)-er than \( y \) for some adjective \( A \), or the proportion of individuals that considers \( x \) \( A \)-er than \( y \). If it is assumed that \( P_A^{xy} + P_A^{yx} = 1 \) for all \( x, y \in X \), \( \langle X, P_A \rangle \) is called a forced choice pair comparison system in mathematical psychology. To make sense of the notion of probabilistic consistency as mentioned above, a measurement representation can be given to such a system. On one natural proposal, what \( P_A^{xy} \) measures is a comparative strength of \( A \)-ness, and this can be captured by a real-valued function \( f_A \) as follows: \( P_A^{xy} > P_A^{yx} \) iff \( f_A(x) > f_A(y) \) (cf. Luce & Suppes, 1965). Under some natural conditions, this kind of ordering relation will give rise to a (strict) weak order. But, of course, many times we would say that an agent or group only considers \( x \) \( A \)-er than \( y \) if \( x \) is judged to be \( A \)-er than \( y \) a sufficiently large percentage of the time, or by a sufficiently large proportion. In such a case we can define a relation \( \succ^\epsilon_A \) on \( X \) as follows: \( x \succ^\epsilon_A y \) iff \( P_A^{xy} > \epsilon \), with \( \epsilon \in [\frac{1}{2}, 1] \). Luce (1958) has already shown that this ordering relation is a semi-order, in case the forced choice comparison system \( \langle X, P_A \rangle \) satisfies certain ‘natural’ conditions.\(^{12}\) Of course, if \( \epsilon = \frac{1}{2} \), the semi-order is a strict weak order.

How is this all related to Fults’ first two critiques? That is, how to make use of forced-choice comparison systems to capture the intuition that the uncertainty peaks at the centre of the interval \( \epsilon \) and fades away at the edges? To answer that question, notice that in case \( P_A^{xy} = \frac{1}{2} \) there is maximal uncertainty about whether \( x \) is \( A \)-er than \( y \) or not. In case \( P_A^{xy} = 1 \), \( x \) is considered to be certainly \( A \)-er than \( y \). Fults’ (2011) first two critiques are thus captured once we make the natural assumption that \( P_A^{xy} = \frac{1}{2} \) exactly at the centre of the interval \( \epsilon \), marking the difference between \( x \) and \( y \) and that \( P_A^{xy} \) approaches 1 or 0 at the edges.

**Weber’s law and semi-orders** Revealing as the representation theorem of Scott & Suppes for semi-orders (Fact 1) may be, it seems to contradict Weber’s law, according to which the threshold grows with intensity. But if this threshold grows with intensity, and thus is not constant, to capture Weber’s law algebraically the representation theorem involving semi-orders (Fact 1) does not seem to be appropriate. What we need is a representation theorem where the threshold is variable.

As it happens, there exists a representation theorem with variable thresholds for so-called interval orders, that was introduced by Wiener (1914). The structure \( \langle X, \succ \rangle \) is an interval order in case the relation \( \succ \) is irreflexive, and satisfies the interval-order condition:

\[^{12}\text{In particular, when } \langle X, P_A \rangle \text{ satisfies strong stochastic transitivity: } \forall x, y, z \in X : \text{ if } P_A^{xy} \geq \frac{1}{2} \text{ and } P_A^{yz} \geq \frac{1}{2}, \text{ then } P_A^{xz} \geq \max\{P_A^{xy}, P_A^{yz}\}.\]
Definition 3 An interval order is a structure \( \langle X, \succ \rangle \), with \( \succ \) a binary relation on \( X \) that satisfies the following conditions:

\[(IR) \quad \forall x : \neg(x \succ x).\]
\[(IO) \quad \forall x, y, v, w : (x \succ y \land y \succ v \land v \succ w) \rightarrow (x \succ w \lor v \succ y).\]

Notice that any interval order is also a strict partial order, because (IR) and (IO) together entail that the resulting ordering is transitive. Just like for strict partial orders and semi-orders, the indifference relation \((xIy \iff \neg(x \succ y) \land \neg(y \succ x))\) for interval orders need not be transitive, and thus does not give rise to an equivalence relation. As a result there is no possibility of getting a full numerical representation, where each member of \( X \) is mapped to a single number. But just like semi-orders, interval orders can also be represented in terms of threshold (Fishburn, 1970), but this time this threshold need not be fixed.

Fact 2 Let \( X \) be a finite set and \( \epsilon > 0 \). Then \( \langle X, \succ \rangle \) is an interval-order iff there exists two positive real-valued functions \( g \) and \( \epsilon \) such that for all \( x, y \in X \):

\[x \succ y \iff g(x) > g(y) + \epsilon(y),\]

with \( \epsilon(y) \geq 0 \) for all \( y \).

We know that every semi-order is an interval order. Indeed, this just follows if we take \( \epsilon \) to be a constant function. However, it also follows that the semi-order discussed in section 3.1 just below the statement of Fact 1, \( \langle X, \succ_p \rangle \) with \( x \sim_p y \sim_p z \sim_p v \) but \( x \succ_p z, y \succ_p v \) and \( x \succ_p v \), can be represented as an interval order by the following functions \( g \) and \( \epsilon \): 
\[g(x) = 2, g(y) = 1.5, g(z) = 1.1, g(v) = 0.8, \quad \text{and} \quad \epsilon(x) = 0.7, \epsilon(y) = 0.6, \epsilon(z) = 0.5 \text{ and } \epsilon(v) = 0.4,\]

if we define \( x \succ y \iff g(x) > g(y) + \epsilon(y) \). Still, not all interval orders are semi-orders: it is possible that we have an interval order \( \langle X, \succ \rangle \) such that \( x \succ y \succ z \), but where it holds that \( vI_x, vI_y \) and \( vI_z \). This is possible in case \( \epsilon(v) \) is very large compared to \( \epsilon(x), \epsilon(y) \) and \( \epsilon(z) \).

Fact 2 shows that the phenomena that are captured by Weber’s law can be represented by an interval-order. However, we will show that although Weber’s law has variable thresholds, it can still be represented in a qualitative way as a semi-order.

It is interesting to observe what happens if we demand that for \( x \succ y \) to hold, we demand not only that \( g(x) > g(y) + \epsilon(y) \), with \( \epsilon(y) > 0 \) for all \( y \), but also make the following constraint: 
\[g(x) \geq g(y) \iff \epsilon(x) \geq \epsilon(y).\]

Notice that it now follows that \( xIy \iff g(x) \not> g(y) + \epsilon(y) \) and \( (g(y) \not> g(x) + \epsilon(x)) \). But this means that \( xIy \) holds iff \( |g(x) - g(y)| \leq \min\{\epsilon(x), \epsilon(y)\} \). With this in hand, we can show that our example of an interval-order that is not a semi-order cannot be represented by the functions \( g \) and \( \epsilon \) observing the new constraint: it is impossible that \( xPyPz \), but where it holds that \( vIx, vIy \) and \( vIz \). For if \( vIy \), then either (i) \( \min\{\epsilon(v), \epsilon(y)\} = \epsilon(v) \) or (ii) \( \min\{\epsilon(v), \epsilon(y)\} = \epsilon(y) \).
But if (i), then \( g(v) \leq g(y) \), and thus \( x \succ v \), which contradicts our assumption that \( x \I v \). And if (ii) then \( g(v) \geq g(y) \), from which it follows that \( v \succ z \), which contradicts our assumption that \( v \I z \).

The idea that the threshold grows with intensity means that for all \( x,y \in X \): \( g(x) \gtrless g(y) \) iff \( \epsilon(x) \gtrless \epsilon(y) \). In Appendix B, we show that any interval-order that is represented by functions \( g \) and \( \epsilon \) satisfying this constraint is, in fact, a semi-order.

**Fact 3** If we can define an order \( \langle X, P \rangle \) in terms of two real-valued measure functions \( g \) and \( \epsilon \) on \( X \) such that for all \( x,y \): \( xPy \) iff \( g(x) > g(y) + \epsilon(y) \), with \( \epsilon(y) > 0 \) for all \( y \) and for all \( x \) and \( y \): \( g(x) \gtrless g(y) \) iff \( \epsilon(x) \gtrless \epsilon(y) \), then this order is a semi-order.

Observe that the ‘if’ cannot be turned into a ‘if and only if’. The above numerical representation is not a necessary condition for semi-orders. Let us provide two examples to show this. First, Fact 1 shows that in order for \( xPy \) to hold in a semi-order, it doesn’t have to hold that \( \epsilon(x) > \epsilon(y) \), if \( g(x) > g(y) \): a constant \( \epsilon \) will do. Second, recall that any strict weak order is also a semi-order. However, there are strict weak orders that cannot be numerically represented by the above, because almost connectedness (i.e., \( \forall x,y,z(xPy \rightarrow (xPz \lor zPy)) \)) does not hold, as can be shown by the following counterexample: \( g(x) = 1, g(z) = 0.9 \) and \( g(y) = 0.8 \) and \( \epsilon(x) = 0.25, \epsilon(z) = 0.2 \), and \( \epsilon(y) = 0.15 \). Then it is clear that although \( xPy \) because \( g(x) = 1 > 0.95 = g(y) + \epsilon(y) \), it is neither the case that \( xPz \) (because \( g(x) = 1 \not> 0.9 + 0.2 = g(z) + \epsilon(z) \)) nor that \( zPy \) (because \( g(z) = 0.9 \not> 0.8 + 0.15 = g(y) + \epsilon(y) \)). In other words, even though the above numerical representation of an order \( \langle X, P \rangle \) is a sufficient condition for it to be a semi-order, it is clear that this numerical representation is not a necessary one.

Notice also that the relation \( I \) is not only reflexive, but because of \( xIy \) iff \( |g(x) - g(y)| \leq \min\{\epsilon(x), \epsilon(y)\} \) also symmetric. The similarity relation does not have to be transitive, as our earlier example illustrates: \( xIz \) and \( zIy \), but \( xPy \).

From fact 3 it immediately follows that Weber’s law can be captured in terms of a semi-order:

**Fact 4** Weber’s law can be captured in terms of a semi-order.

Recall that Weber’s law states that the size of the difference threshold, \( \Delta g(x) \), is lawfully related to \( g(x) \), the magnitude of stimulus \( x \) as follows:
In this ‘law’, \( \Delta g(x) \) represents the difference threshold, \( g(x) \) represents the stimulus intensity of \( x \), and \( k \) signifies that the proportion on the left side of the equation remains constant despite variation for the \( g(x) \) term. Stating it otherwise, \( \Delta g(x) = k \times g(x) \). If the difference threshold is a constant proportion equal to 0.1, then the size of the threshold for an experience caused by something with an intensity of 100 units would be 10 units (i.e., \( \Delta g(x) = 0.1 \times 100 = 10 \)), and an experience caused by something with an intensity of 1000 units would be 100 units (i.e., \( \Delta g(x) = 0.1 \times 1000 = 100 \)). Now we can say that \( xPy \) iff \( g(x) > g(y) + \Delta g(y) \) and that \( xIPy \) iff \( |g(x) - g(y)| \leq \Delta g(y) \). We have represented interval-orders above as follows: \( xPy \) iff \( g(x) > g(y) + \epsilon(y) \). These two are obviously equivalent, if we take \( \epsilon(y) \) to be \( \Delta g(y) \) for each \( y \). However, because it now immediately follows that \( \epsilon(x) > \epsilon(y) \) iff \( g(x) > g(y) \), our interval order is, in fact, a semi-order, because this was the condition we required for an interval order to be a semi-order.

### 4.3 The Interpretation of the Threshold

In the beginning of the paper we followed Kennedy’s (2011) proposal that whereas for the truth of the explicit comparative (1-a) (repeated as (3-a)) any (directly or indirectly observable) bit of John’s height that exceeds Mary’s suffices, there should be a significant gap between the heights of John and Mary for the implicit comparative (1-b) (repeated as (3-b)) to be true.

(3) a. John is taller than Mary.
   b. Compared to Mary, John is tall, but compared to John, Mary is not tall.

Later in the paper, however, we interpret a significant gap as the standard interpretation of a threshold (Luce, 1956): as a just noticeable difference. But this suggests that whereas (3-b) can only be true if there exists a JND between the heights of John and Mary, no such JND need to exist in order for (3-a) to be true. But if no JND need to exist between the heights of John and Mary, one wonders how (3-a) could be true at all? Perhaps one could say that for the truth of (3-b) it is only demanded that there be a difference, not that it is noticeable. Perhaps. We take it to be more natural, however, that two notions of just noticeable difference are involved, although both depend on length.

Perhaps economics provides the most obvious case to illustrate two different ways to think about just noticeable differences that depend on the same objective measure. Bernoulli (1738) has proposed a difference between perceived happiness and amount of money, although the former might depend on the latter. Although it is clearly the case that after counting, a person can notice the
difference between €995 and €1000 equally well as the difference between €15 and €20, it can still be the case that this person cannot perceive a difference in happiness between owing €995 and €1000, but can do so between owing €15 versus €20. Perceived amount of money clearly gives rise to a non-zero notion of just noticeable difference in both cases, while perceived pleasure need not be. This is why Bernoulli concluded that utility should not be measured in terms of euros (or florins, in his case). Although utility depends on euros, the dependence is logarithmic rather than linear in nature. One might think of Bernoulli’s strategy for making a distinction between amount of euros and utility in a slightly different way: as one between exact counting, on the one hand, and approximate counting, on the other.\textsuperscript{13} Thus, in both cases, it is only numbers of euros that counts, but in the one case this number is more exact than in the other case.

Can we explain the results of Experiment 2, and Kennedy’s (2011) intuition, according to which we would not count John as being tall (compared to Mary), although he is still observed to be taller, in terms of a similar strategy? In van Rooij (2011a) it was proposed that the difference between explicit and implicit comparatives like (3-a) and (3-b), and the difference between weak orders and semi-orders, can be thought of in terms of the distinction between a direct versus indirect observable difference: whereas direct observable difference gives rise to a semi-order, indirect observable difference gives rise to a more fine-grained weak order. But what does it mean to be a direct or indirect observable difference? One way to account for the distinction between the two is to assume that what counts for ‘direct distinguishability’ is speed, and the ability to make the distinction under high cognitive load. It might well be that although precise counting and exact measuring is possible when enough time is allowed, this is not possible under these time-pressured circumstances. What counts in these latter cases is approximate magnitudes, and these give rise to Weber’s law. Thus, we propose that we should interpret the thresholds involved in our use of semi-orders to account for the data as just noticeable differences, and to explain Kennedy’s (2011) intuition—with which we agree—that there is a difference (in English) between the use of explicit versus implicit comparatives in terms of the circumstances under which we would use such comparatives. In contrast to explicit comparatives, implicit comparatives are used only in case precise measures do not matter, or when there is not time enough to determine them. We propose a similar hypothesis to explain the data of Experiment 2: the use of ‘tall’, instead of the more precise ‘taller than’, is limited to

\textsuperscript{13}But as one reviewer rightly remarked, this cannot be everything to Bernoulli’s strategy: Bernoulli also introduced logarithmic functions for purely formal reasons, i.e., to prevent choices from having infinite expected utility. There are also general considerations about the diminishing marginal value of money.
those circumstances in which all that counts, or all that can be done, is (to make) a division of a group into a few relevant classes based on (approximate) height.

5 Conclusion

The problem of how we group individuals or objects when we use a vague adjective like tall is one that has puzzled linguists and philosophers for a long time. Unlike the case of a comparative form of such adjectives, where it seems like the definition is straightforward, we do not know what exact criteria are used to determine when a positive form of a vague adjective is appropriate as a description. In this paper, we focused on the gap hypothesis as a potential solution. According to this proposal, for example in the case of height, we would only use tall if there was a significant gap between the shortest person in the group of people classified as tall and the tallest person in the non-tall group. This hypothesis is related to the fact that we as humans do not have precise information about, for example, heights of people. In a situation where we cannot take a ruler and measure the person’s height, we make an estimate. For such a situation, Weber’s law states that discriminability is determined by the ratio of difference between the two values and the higher the difference ratio, the easier it is for us to discriminate them.

We have presented experimental evidence that confirms this intuition about the difference in meaning between the positive and comparative forms of the vague adjectives: while any difference in sizes between the target object and the other objects within the context was sufficient for the comparative forms to be acceptable, acceptability of the positive form depended on the size difference. For using the positive form as a description, the larger size differences resulted in higher acceptance rates (and lower reaction times) than smaller ones. Our results also suggest that a positive form can in principle be used to describe simply the biggest, longest, etc., object within the context regardless of whether there is a gap or not.

One way to capture the difference between the comparative and positive forms existence of which our experiments confirmed is in terms of weak orders and semi-orders (van Rooij, 2011a). However, a number of challenges to semi-orders have been raised by Fults (2011) based on Weber’s law and Analoge.

14 Of course, one can think of circumstances where we talk about high magnitudes, but still a very fixed threshold is involved. For instance, it might be that one only has to pay a higher percentage of income taxes when one earns more than exactly €100,000.

15 This hypothesis would also explain, we believe, why, intuitively, the threshold has to be larger for the case of positives like (2) than for (implicit) comparatives.
Magnitude Representations. Namely, semi-orders make use of a sharp threshold of applicability that would not be possible if we indeed make approximations of values such as people’s heights using AMRs. In addition, so far it has been proposed that the threshold is fixed across different values, which again is problematic under AMRs that are rather ratio-based. Besides these challenges, our experimental results showed that the requirement of a significant gap might be too strong for vague adjectives. In Section 4, we showed how the analysis of vagueness presented in van Rooij (2011a) can be adjusted to meet these objections and refer to ongoing work that can solve the latter issue with the significant gap not being required. We thus showed that an analysis in terms of weak orders and semi-orders can still account for the use of vague adjectives and the difference between comparative and positive forms.

One question that we feel is still up for discussion is the interpretation of the threshold for comparative and positive forms that we discussed in Section 4.3. Intuitively, and based on our experimental results, an explicit comparative such as taller can be used for a situation where a positive form tall would not be acceptable. So there should be two different thresholds. We put forward our explanation of the difference as being in the contexts where the positive and comparative are used. We suggested that the positive form is used in situations when we cannot make exact measurements but just approximate ones, and that explains why Weber’s law is important there, whereas a comparative form would be used when precise values are known and are important.

There should be follow-up experimental work to investigate the exact processes that result in differences in reaction times between the comparative and positive adjective judgments that we obtained. In addition, we need to find out where the differences between participants stem from. Our participants either interpreted the meaning of vague adjectives differently or interpreted the set-up differently. It would be useful to know whether their choices could be altered if they were given a different task. For example, they could be given a more explicit communicative task where they have to describe an object to another person.
Appendix A. Experimental Procedure and Participants

Procedure The experiment was built using a JavaScript library for online chronometric experiments—JsPsych—and run in participants’ web browsers. This library allows for collection of participants’ responses as well as fairly accurate reaction time data (de Leeuw 2015, de Leeuw & Motz 2016). The first screen that participants saw displayed general information about the experiment. On the second screen, they had to give consent to participate, and agree with storage of the data obtained. They then filled in a short questionnaire regarding their background, which served as a check that indeed all eligibility requirements were met. This was followed by detailed instructions for the task. Participants then did three practice trials, after which the experiment itself started. There were two blocks—one with sentences with positive vague adjectives and one with sentences with comparative vague adjectives. Each of the blocks contained 56 trials. Between the blocks, participants could take a break.

Trials within each block were presented in random order. Each trial started with a display of the sentence in the middle of the screen. When the participants finished reading the sentence, they pressed the space bar to see the picture itself. They then had to press either P if they agreed that the sentence could be used to describe this picture, or Q if they did not agree. The participants had 2300 ms to give a response, and if no response was given, the trial ended automatically and it was recorded as a missing response. The next trial then started after a random inter-stimulus interval between 700 and 1200 ms.

The instructions that participants received were the following:

‘In this experiment, you will see pairs of sentences and pictures. Your task is to indicate whether you think the sentence is a good description of the given picture. In other words, do you think people would use this description for the given picture? You will need to think about the exact meanings of words in order to make correct judgments.

Every trial will start with a sentence [...] Please pay attention to the meanings of sentences and give correct answers, but also try to do it quickly. [...]’

Participants Eligible participants had to meet the following criteria: native speaker of English; age 18–35 years old; born and currently living in the USA; right-handed. Completing the experiment should have taken approximately 15 minutes and the participants were paid 1.80 £.


Experiment 1: Twenty-five participants completed the task. One participant was excluded due to reading the instructions for under 10 seconds. The remaining 24 participants were included in the analysis. Nine were male, 15 female; their mean age was 24.96 (range 18–33). They took 11:29 minutes on average to complete the task (min. 08:33 and max. 17:11).

Experiment 2: Forty participants completed the task. One participant reported being color-blind and was excluded. Six further participants were excluded for reading the instructions for under 10 seconds. Two participants were excluded for pressing a single button as a response throughout the experiment. Thus, 9 participants in total were excluded and the analyses were performed on the remaining 31 participants. The mean age of these participants was 26.26 (range 20–34). Thirteen were female, 17 male, and 1 of other gender. Participants took 11:19 on average to complete the task (min. 07:11, max. 17:32).

Appendix B. Proof of Fact 3

To prove Fact 3, we have to show that the resulting order obeys the conditions for being a semi-order. That is, it has to be (i) irreflexive (IR), (ii) satisfy the interval order condition (IO), and (iii) be semi-transitive (STr).

(IR) Irreflexivity follows immediately if $\epsilon(x) \geq 0$.

(IO). Suppose $xPy$ and $vPw$. To prove $xPw$ or $vPy$. Suppose $\neg xPw$. It follows that (i) $g(x) > g(y) + \epsilon(y)$, and (ii) $g(v) > g(w) + \epsilon(w)$, and (iii) $g(x) \neq g(w) + \epsilon(w)$.

Now either (a) $\epsilon(v) = \epsilon(x)$, (b) $\epsilon(v) < \epsilon(x)$, or (c) $\epsilon(v) > \epsilon(x)$.

(a) But then by the constraint it follows that $g(v) = g(x)$, and with (ii) and (iii) we have a contradiction.

(b) But then by the constraint it follows that $g(x) > g(v)$ and thus with (ii) $g(x) > g(w) + \epsilon(w)$, which is in contradiction with (iii).

(c) But then by the constraint it follows that $g(v) > g(x)$ and thus with (i) $g(v) > g(y) + \epsilon(y)$. But then $vPy$.

Similarly, we can prove that if $\neg vPy$ it follows that $xPw$, which is enough to prove what we wanted.

(STr). Suppose $x > y$ and $y > z$. To prove for any $v$: $xPv$ or $vPz$. So suppose $xPy$ and $yPz$ and $\neg xPv$. That is, suppose (i) $g(x) > g(y) + \epsilon(y)$ (ii)
\[ g(y) > g(z) + \epsilon(z), \text{ and (iii) } g(x) \not> g(v) + \epsilon(v). \] Now either (a) \( \epsilon(v) = \epsilon(x) \), (c) \( \epsilon(v) > \epsilon(x) \) or (e) \( \epsilon(v) < \epsilon(x) \).

(a) By the constraint it follows that \( g(x) = g(v) \). Notice that from (i) \( g(x) > g(y) + \epsilon(y) \) and (ii) \( g(y) > g(z) + \epsilon(z) \) it immediately follows that \( g(x) > g(z) + \epsilon(z) \) (transitivity). Thus \( vPz \).

(b) By the constraint it follows that \( g(v) > g(x) \). Notice that from (i) and (ii) it follows that \( g(x) > g(z) + \epsilon(z) \), as in (a). Thus now also \( g(v) > g(z) + \epsilon(z) \). But this means that \( vPz \).

(c) By the constraint it follows that \( g(x) > g(v) \), and thus \( g(v) + \epsilon(v) \geq g(x) > g(v) \). Now there are three possibilities: either (c1) \( \epsilon(v) = \epsilon(y) \), (c2) \( \epsilon(v) < \epsilon(y) \), or (c3) \( \epsilon(v) \geq \epsilon(y) \).

(c1) By the constraint it follows that \( g(y) = g(v) \). But this is impossible, because we have assumed that \( xPy \) and \( \neg xPv \).

(c2) By the constraint it follows that \( g(y) > g(v) \). Thus, \( g(y) + \epsilon(y) > g(v) + \epsilon(v) \). But this is impossible, because we have assumed that \( xPy \) and \( \neg xPv \).

(c3) So, (c3) has to be the case. But this means with the constraint that \( g(v) > g(y) \). Because \( g(y) > g(z) + \epsilon(z) \), it follows that \( g(v) > g(z) + \epsilon(z) \). Thus \( vPz \).

Thus, if \( \neg xPv \), then \( vPz \).

Similarly, we can also prove that if \( \neg vPz \), then \( xPv \). Now we have proved semi-transitivity.

References


