

A QUANTITATIVE STUDY OF VAGUE LANGUAGE

NOAH D. GOODMAN

1. INTRODUCTION

When does a movie stop being *good*? Is a five-foot man *short*? And, how many hairs can you have and still be *bald*? These are all questions that seem not to have an answer, because the language used is *vague*. Vagueness is extremely common in everyday language, being perhaps the rule rather than exceptional. This has important implications. For instance, the legal doctrine of [whatsit] allows local statutes to be invalidated if they are determined to be vague [true?]. [something about AI / internet / NLP?] Vagueness may also expose for study critical features of the psychological structure of language and cognition. The lack of clear truth-conditions has been the most persistent obstacle to the use of mathematical logic as a model of human thought (which is indeed what mathematical logic was initially developed for [?]). The silver lining for experimental psychology is that vagueness implies quantitative, graded data where otherwise there might only be binary truth values; there is more information in a number than a bit. Indeed, it was precisely the vague boundaries of everyday concepts that drove psychological research to develop the notions of prototype and typicality [?]. While this line of research yielded successful quantitative models of category generalization [?, ?, ?], less progress has been made at connecting theories of vague language to psychological data. In this paper we describe and evaluate a quantitative theory of vague adjectives.

Vague concepts are characterized by having borderline cases: instances that are neither clearly in nor clearly out, even if one knows everything about that instance. For instance, having measured a man's height down to a millimeter, it may still be impossible to determine if that man is *tall*. The puzzle of vagueness goes back thousands of years, at least to the ancient greek philosopher [whatshisface], who created a series of logic puzzles that hinged on vague language. The most famous is known as the *sorites*, or paradox of the heap:

- (1) A million grains of sand together are a heap.
- (2) If you take one grain of sand from a heap, it is still a heap.
- (3) One grain of sand is a heap.

People tend to happily endorse the first two sentences, but resoundingly reject the last sentence. Yet by mathematical induction the last statement should follow from the first two. While this puzzle has often been taken as a problem for logic or metaphysics, we instead view it as a psychological

issue. What could be the process of interpretation and endorsement that would lead to the observed pattern of agreement?

[should mention sorites series form?]

As formulated above the sorites involves vague nouns, such as *heap*. Many features of noun concepts make it hard to isolate the origin of their vagueness; they are high-dimensional and likely depend on function and relation [?]. Adjectives, on the other hand, demonstrate the key properties of vagueness—borderline cases and the sorites—but usually depend on a single critical scale. For instance, *tall* seems to depend on height alone, and *expensive* on price. This makes adjectival vagueness a more tractable ground for model building—at least that is our motivation for considering adjective in this paper. Vagueness is characterized by borderline cases and the sorites. Borderline cases, and gradedness, are clearly possible for adjectives; it is not hard to create sorites arguments for adjectives paralleling the one above:

- (1) A million-dollar watch is expensive.
- (2) A watch that costs one dollar less than an expensive watch is still expensive.
- (3) A one-dollar watch is expensive.

In the remainder we refer to sentences like (1) as our *concrete* sentences and sentences like (2) as *inductive*. Vague adjectives have a third key property, related to the others: they are relative, depending critically on the class. For instance an *expensive watch* is likely very different in price from an *expensive house*. Thus the challenge is to experimentally establish that these three properties and explain their quantitative pattern with a formal model of interpretation. [revise that last sentence.]

It is remarkably easy for people to evaluate sentences like 1–3, forming an opinion about their truthiness consistently and without much reflection. This and the prevalence of vague utterances both suggest that vagueness is central to the design of language—a feature not a bug. We will suggest that vagueness indeed follows from basic aspects of the architecture of language: underspecified literal meanings coupled with pragmatic inference that fills in meaning by reasoning about the speaker’s intent, in a conversational context. [more on uRSA]

[related research??] [?]

In the remainder of this paper we [first?] focus on the scale of price and the adjective *expensive*. This is a useful case study because people have robust prior knowledge of prices for different common object types, and prices can be referred to directly (people can report and comprehend prices as numbers, unlike other scales, such as *goodness*). In section 2 we collect data for endorsement of concrete and inductive sentences, and show that the three characteristics of vague adjectives are evident: gradedness (including borderline cases), dependence on category (relativity), and sometimes strong

endorsement of inductive sentences. We then introduce a model of adjectival vagueness in section 3, which realizes adjective meanings as underspecified degree comparisons and fixes meaning through a pragmatic inference. [uRSA] This model depends on non-linguistic background knowledge; in section 4 we elicit the prior distribution on prices for several different categories of everyday item. We find that the model predictions, given these elicited priors, provide a good match the human data from Experiment 1. In section 5 we replicate and extend this finding, using a wider range of prices in the target sentences. [In section ?? we consider other adjectival scales, showing something.] We conclude by considering the implications of our results for language and cognition, and the relationship of our model to other theories of vagueness.

2. EXPERIMENT 1

2.1. Participants. $N = 70$ participants for Experiment 1a and $N = 50$ participants for Experiment 1b were recruited through Amazon’s Mechanical Turk.¹ Only participants with U.S. IP addresses and approval rate above XX were recruited [check]. Participants were paid for their time.

2.2. Materials. We generated a set of concrete sentences of the form “A {X} that costs \${P}\$ is expensive.” There were five target categories “X”: headphones, laptop, watch, coffee maker, and sweater. The target price “P” had XX different values for each category, chosen by the experimenters’ intuition for a reasonable range of plausibility.

We also generated a set of inductive sentences. In Experiment 1a, they had the relative clause form “A {X} that costs \${P}\$ less than an expensive {X} is also expensive.” The inductive sentences for Experiment 1b instead had the conditional form “If a {X} is expensive, then another {X} that costs \${P}\$ less is also expensive.” In both cases, we used the same five target categories “X”. There were XX target price gaps “P” per category, again chosen by intuition with the goal of yielding both good inductive premises and more questionable ones.

2.3. Methods. Participants completed 55 trials in about ten minutes. Participants were instructed:

In this experiment, we would like you to give us your opinion about some statements. Each statement is about the prices of different household items.

For each question, please indicate how much you agree with the statement. Choose a circle that is closer to the words “Completely agree” if you agree and choose a circle that is closer to the words “Completely disagree” if you disagree.

¹The number of subjects was chosen based on... [what]?

Each participant was then asked to rate a series of sentences on a nine-point scale. [WHAT WAS THE DESIGN?] Each participant then saw all five categories with all four price values for each kind of sentence [yielding $5 \times 5 \times 2 = 50$???]; trials were presented in random order.

2.4. Results. Judgements for inductive sentences of Experiment 1a and 1b had an extremely strong correlation of $r = 0.98$. [double shcek that this is correlation of just inductive sentences.] We therefore combine data from the two experiments for the remainder of the analyses.

it is reliable (people agree: split half? beta?). we qualitatively capture the three characteristic effects of vagueness: we see that judgements depend on the class (and on the price and frame). we see gradedness in endorsement of concrete premises, including borderline cases. we see strong endorsement of some inductive premises. importantly this is also graded, with some inductive sentences judged as not good at all.

this suggests rich quantitative patterns to explain.

3. MODEL

we make an uRSA model building on (or just using?) [?].

should we be using an S2 model, since this is agreement data?

but to apply model need priors....

alternative models? simpler Bayesian models – L0 and un-lifted L1. implementable fuzzy logic model??

4. EXPERIMENT 2

priors. do BDA with log-normal family? linking fn?

4.1. Comparison to endorsement data. model fit.

5. EXPERIMENT 3

expt 3a and 3b: getting a wider range of endorsement.
model fits.

6. EXPERIMENT 4???

7. GENERAL DISCUSSION

-yay.

-extending to other adjectives and scales?

7.1. Relation to other theories of vagueness. is this a version of epistemicism? of supervaluation? relation to fuzzy logic?

"Stephen Schiffer (2003, 204) denies that classical probability calculations apply in vague contexts. Suppose Ned is borderline old and borderline bald. According to Schiffer we should be just as confident in the conjunction ?Ned is old and bald? as in either conjunct. [...] He crafts the rules for vague

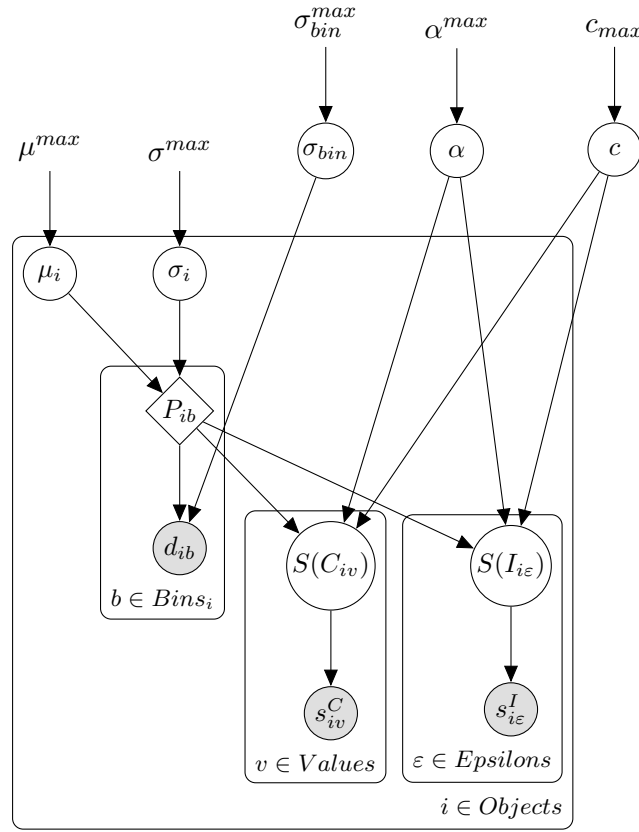


FIGURE 1. Graphical model

partial belief to provide a psychological solution to the sorites paradox.” (SEP)

maybe un-lifted threshold model is a version of epistemicism (there is a threshold, each listener is uncertain what it is).

supervaluation seems to be related to disambiguating by choosing a meaning at the literal level that is applied to all instances of a word? but however it works it doesn’t seem to account for graded endorsement – just rejection of truthity.

contextualism: ”Scott Soames (2002, 445) answers that all vague words literally are indexical.” (SEP) ”Hans Kamp, the founder of contextualism, maintained that the extension of vague words orbits the speaker’s store of conversational commitments. In a more psychological vein, Diana Raffman says changes in context trigger gestalt shifts between look-alike categories.” (SEP) ”Stewart Shapiro integrates Kamp’s ideas with Friedrich Waismann’s concept of open texture. Shapiro thinks speakers have discretion over borderline cases because they are judgment dependent.” (SEP)

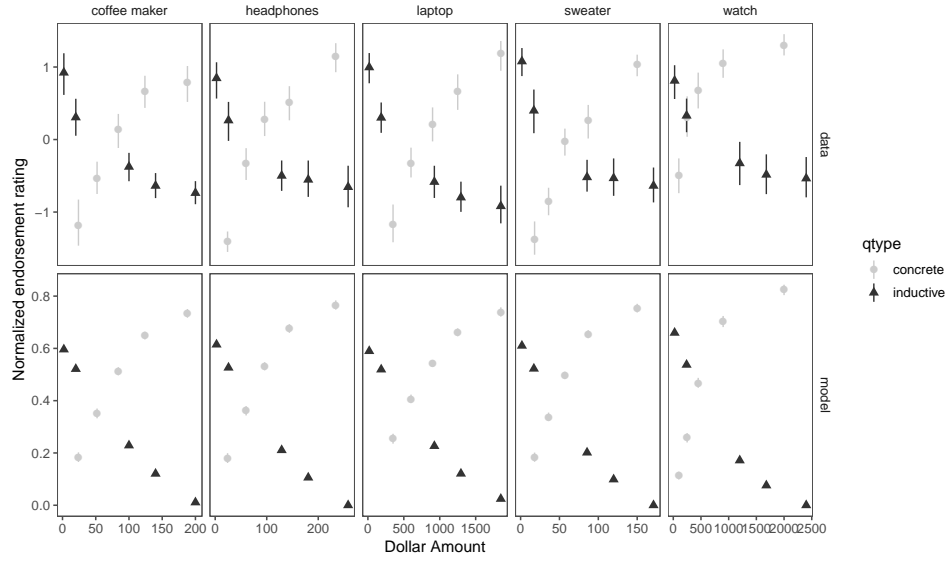


FIGURE 2. Model fit for final experiments

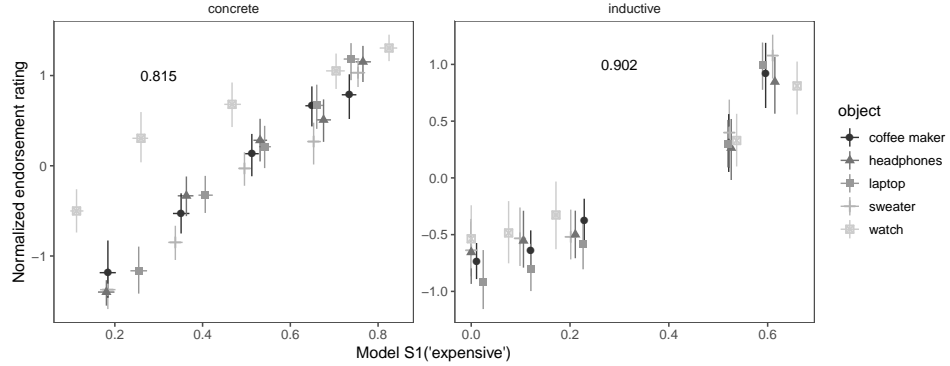


FIGURE 3. Final experiments correlations

It's worth reflecting on Kamp and Partee's "Prototype theory and compositionality" (Cognition, 1995). (<http://semantics.uchicago.edu/kennedy/classes/s06/readings/kamp-partee95.pdf>)

7.2. Other vague language. vague nouns.
higher-order vagueness.

8. CONCLUSION

APPENDIX A. DISCRETIZATION FOR RSA MODEL FIT