Vagueness, Logic and Use: Four Experimental Studies on Vagueness[*]

Phil Serchuk, Ian Hargreaves, and Richard Zach

Abstract: Although arguments for and against competing theories of vagueness often appeal to claims about the use of vague predicates by ordinary speakers, such claims are rarely tested. An exception is Bonini et al. (1999), who report empirical results on the use of vague predicates by Italian speakers, and take the results to count in favor of epistemicism. Yet several methodological difficulties mar their experiments; we outline these problems and devise revised experiments that do not show the same results. We then describe three additional empirical studies that investigate further claims in the literature on vagueness: the hypothesis that speakers confuse '*P′* with 'definitely *P*', the relative persuasiveness of different formulations of the inductive premise of the Sorites, and the interaction of vague predicates with three different forms of negation.

0. Introduction

The phenomena of vagueness raise important and complex issues in the philosophy of language and of logic. One such phenomenon is the Sorites Paradox:

(1) Someone who is 2 m in height is tall.

(2) If someone *x* mm in height is tall then someone *x - 1* mm in height is also tall.[1]

[*]This work was partially funded by an NSERC USRA. Thanks to Diana Raffman, Phil Kremer, Jeremy Fantl, Penny Pexman, the referees for *Mind and Language*, and audiences at the Unviersity of Toronto Grad Forum and the Paris Vagueness and Language Use conference for helpful comments.
Address for Correspondence: Phil Serchuk. Department of Philosophy, University of Toronto, 170 St. George St., Toronto, Ontario, M5R 2M8, Canada.
E-mail: phil.serchuk@utoronto.ca
[1]In classical logic (2) is logically equivalent to both

(3) Someone who is 1 m in height is tall.

(1) and (2) are, it seems, true, and (3) is false. Yet one can derive (3) from (1) and (2) using quantifier rules and the logical rule of inference known as Modus Ponens. Theories of vagueness aim at, among other things, an analysis and solution to this paradox. Different theories of vagueness will generally do this in different ways. For instance, some deny that (2) is true, while others deny that the rule of Modus Ponens is an acceptable rule in the presence of vague predicates like 'is tall'. Although many theories of vagueness are committed to claims about the psychology of vague concepts or the actual use of vague language by ordinary speakers, philosophers of vagueness rarely test these claims against empirical data. For instance, it is a standard part of defenses of theories which deny the truth of (2) to give an explanation for why (2) seems intuitively plausible. Yet what is plausible to a philosopher of logic need not at all be plausible to ordinary speakers.

The recent literature in so-called experimental philosophy is in large part dedicated to test claims of intuitive plausibility. Typically, this is done by comparing the intuitions of philosophers to those of the 'folk', where the 'folk' intuitions are obtained from some kind of empirical study. In this paper, we present the results of a series of such studies of the intuitions of ordinary English speakers about the use of vague language. We use our data to evaluate two such appeals to intuitive plausibility in the vagueness literature. One is the hypothesis that ordinary speakers systematically confuse vague predications, e.g., “John is tall”, with determinate counterparts of the form “John is definitely tall”. Generalized and restricted versions of the hypothesis have been given by

(2') It is not the case both that someone *x* mm in height is tall and someone *x - 1* mm in height is not tall, and

(2") Either someone *x* mm is not tall or someone *x-1* mm is.

a variety of theorists. For example, Fine (1975) and Keefe (2000) defend a restricted version of the hypothesis in support of a supervaluationist theory of vagueness. The second example comes from a paper by Weatherson (2005). In support of his own theory and against the degree theory of vagueness, Weatherson claims that

(2') It is not the case both that someone *x* mm in height is tall and someone *x - 1* mm in height is not tall.

is more persuasive to ordinary speakers than (2). (2') is equivalent to (2) in classical logic, and always equally or less true than (2) in fuzzy logic. We use empirical data to evaluate these claims in Sections 2 and 3.

Some proponents of experimental methods in philosophy go further than using empirical data to support or undermine claims regarding the 'intuitive plausibility' of specific philosophical claims, and take the results of empirical research to constitute evidence for or against philosophical theses. An early example of this approach is a paper by Bonini, Osherson, Viale, and Williamson (1999), which not only predates the recent resurgence of interest in experimental philosophy by several years, but is one of just a handful of papers that applies empirical methodology to philosophical work on vagueness.[2] Bonini et al. elicited the judgments of Italian speakers to ascriptions of vague predicates to borderline cases. From the results obtained they ruled out the gap, glut and degree theories of vagueness as not supported by speakers' actual use of vague predicates, and found support for Williamson's epistemic theory of vagueness.

[2]Early and *very* informal efforts are reported by Parikh (1991, 1994), who tests speaker disagreement about colour category boundaries and the assignment of fuzzy values to borderline cases. Hans Kamp ran an unpublished colour patch experiment in which he tried to find evidence of hysteresis (personal communication). Raffman (forthcoming) defends her theory of vagueness with empirical data showing a hysteresis effect in colour judgments. Alxatib and Pelletier (forthcoming) also use empirical methods to study vagueness; their work is described in later sections.

A large part of the current debate about philosophical methodology revolves around the question of whether empirical results of the sort presented by experimental philosophers have any role at all to play in philosophical investigations. Without lending support to either side of this debate, we would argue that if experimental research has a place in philosophical theorizing, it does so only insofar as the relevant experiments hold up to scrutiny. In particular, such experiments must be reproducible, the experimental designs must be sound, and the statistical methods used in their analysis must be appropriate. In light of this, in Section 1 we criticize the methodology used by Bonini et al. and present experimental data that undermine their conclusions. We identified a crucial ambiguity in their questionnaire. We then replicated their original study (using English speakers) and compared it to the results of an improved version of the questionnaire. The results from our revised experiment differ greatly from those found in Bonini et al. (1999), and do not support their conclusions.

We conclude by presenting data on speakers' use of negation in borderline cases. For reasons of space, we will not defend the use of empirical research in philosophy of language. If empirical studies have philosophical import (as we think they do) then empirical studies of negation will be an important component of such work. This is because some theories of vagueness, e.g., many-valued and truth-gap theories, leave room for different interpretations of negation. It is also an interesting question whether internal negation (e.g., “John is not tall”) and external negation (e.g., “It is not the case that John is tall”) should be treated differently in a logic of vagueness, and whether ordinary speakers treat such constructions differently may be taken by some to have a bearing on this question. Likewise, the treatment of negation by ordinary speakers might

be taken to favor one or another interpretation of negation. We describe an experiment we used to test these issues Section 4. Our goal in this paper is not to present a single unified theory of vagueness. Rather, we believe our work can help to refine existing theories of vagueness by providing theorists with new data on existing problems. Many of these problems are typically solved with armchair methods that, as we now show, do not correspond with the empirical data.

1. A Review of Bonini et al.

We begin by giving a critical evaluation of an experiment conducted by Bonini et al. (1999), who used empirical methods to test Williamson's epistemicist predictions about language use. After outlining their experiment we identify several problems that undermine its conclusion. The first concerns the implementation of their experiment: we show that their survey question is both ambiguous and question-begging, and that the ambiguity had a significant effect on their findings. We then argue that their experiment suffers from methodological problems which undermine its philosophical import.

Bonini et al. (1999) ran seven studies, the first six of which were nearly identical. In these first experiments they divided participants into two groups, truth-judgers and falsity-judgers. In the first three studies, truth-judgers were asked

> When is it true to say that a man is 'tall'? Of course, the adjective 'tall' is false of very small men and true of very big men. We're interested in your view of the matter. Please indicate the smallest height that in your opinion makes it true to say that a man is 'tall'.

> It is true to say that a man is ‘tall’ if his height is greater than or equal to ____ centimeters.

Falsity-judgers were asked

> When is it false to say that a man is ‘tall’? Of course, the adjective ‘tall’ is false of very small men and true of very big men. We're interested in your view of the matter. Please indicate the greatest height that in your opinion makes it false to say that a man is ‘tall’.
>
> It is false to say that a man is ‘tall’ if his height is less than or equal to ____ centimeters.

In each study participants were asked a similar question about up to six different vague predicates. Bonini et al. claim that the epistemicist, gap and degree theorist each predict that the average estimates of truth-judgers will be significantly larger than those of falsity-judgers (ibid, pp. 379-380). They call these 'gaps' (not to be confused with gap theory, or truth gaps) because there is a gap between where truth-judgers stop and where falsity-judgers start applying the predicate. They claim the glut theorist predicts gluts, i.e., an overlap between where truth-judgers stop and where falsity-judgers start applying the predicate. We describe and evaluate these claims in section 1.2.

Because all six studies yield similar results we focus on data from three predicates in their second study, which we call *Bonini-vague*. Their data (ibid, p. 383) are reproduced in table 1. The table shows the mean estimates and standard deviation (*SD*) of each group. To test if the differences between truth-judgers and falsity-judgers were greater than those that could be expected by chance alone, Bonini et al. used a Mann-Whitney *U*-Test (MW). MW ranks all of the scores from lowest to highest (assigning tied

ranks to tied scores), and then sums the ranks for each group. To see if the two groups differ, the sum of the ranks for each group is compared with one another, yielding a differences score. The standardized difference between these groups, or z, is then assessed for significance to determine whether it is likely to have occurred by chance alone. p is an indicator of significance. When $p > 0.05$, the observed difference is not large enough to allow us to infer that the groups are different, as these differences may be due solely to chance; however, when $p < 0.05$ we can be reasonably certain that the observed differences are significant, and we can use this information to inform our inferences.

Predicate	Truth-judgers (N=52)		Falsity-judgers (N=56)		z-score	
'tall'	179.55 cm	SD=6.7	164.13 cm	SD=14.7	z=7.48	p<.001
'old'	76.59 yr	SD=8.7	62.27 yr	SD=10.5	z=6.35	p<.001
'long'	165.16 min	SD=42.1	121.96 min	SD=43.8	z=4.95	p<.001

Table 1: Results from *Bonini-vague* (Mann Whitney *U*-Test)

For each question, the average response of truth-judgers is larger than that of falsity-judgers. The difference was statistically significant in every case. For example, the data show a 14-year gap between where falsity-judgers stopped taking the application of 'old' to be false and where truth-judgers started taking it to be true. Bonini et al. take these data to be consistent with three different theories (truth gaps, degree theory, and epistemicism). However, they argue for an epistemicist interpretation on two grounds. The first is familiar theoretical arguments against gap and degree theory. We will not discuss these arguments, though they are evaluated by Alxatib and Pelletier (forthcoming). The second is by means of a seventh study aimed at testing epistemicism

directly. This study, which we label *Bonini-crisp*, tested the hypothesis '*S* mentally represents vague predicates in the same way as other predicates with sharp true/false boundaries of whose location *S* is uncertain' (ibid, p. 387).

The questions in *Bonini-crisp* were similar to those in *Bonini-vague*, but they contained predicates that Bonini *et al*. claimed were both crisp and had uncertain boundaries. This time participants were divided into upper-judgers (mirroring the truth-judgers of *Bonini-vague*) and lower-judgers (mirroring the falsity-judgers). Upper-judgers were asked[3]

> When is a man at least of average height among 30-year-old Italians? Of course, a man of at least average height among 30-year-old Italians is not too tall and not too short. We are interested in your view of the matter. Please indicate the smallest height that in your opinion makes a man of at least average height among 30-year-old Italians.
>
> > A man is of at least average height among 30-year-old Italians if his height is greater than or equal to ____ centimeters.

Similarly, lower-judgers were asked

> When is a man not as tall as average among 30-year-old Italians? Of course, a man not as tall as average among 30-year-old Italians is not too tall and not too short. We are interested in your view of the matter. Please indicate the largest height that in your opinion makes a man not as tall as average among 30-year-old Italians.

[3]The questionnaires used by Bonini et al. were given to Italian students (in Italian). Bonini et al. published translations of the questions in *Bonini-vague*, but not of those in *Bonini-crisp*. Part of the question below was translated into English from a copy of the original Italian text sent to us by Daniel Osherson.

A man is not as tall as average among 30-year-old Italians if his height is less than or equal to ____ centimeters.

Participants were asked about six different predicates, including "of average height among 30-year-old Italians", "of average age for an adult Italian", and "the average [film length] of those shown at the Biannual Venice Film Festival". The data from *Bonini-crisp* (ibid, p. 391) for three predicates is reproduced in table 2.

Question	Upper-judgers (N=42)		Lower-judgers (N=43)		z-score	
average height	171.48 cm	SD=4.86	150.49 cm	SD=32.95	z=5.05	p<.001
average age	55.67 yr	SD=15.55	35.79 yr	SD=15.19	z=5.07	p<.001
average length	131.07 min	SD=63.78	83.67 min	SD=35.83	z=4.05	p<.001

Table 2: Results from *Bonini-crisp* (Mann Whitney *U*-Test)

For each question, the average response of upper-judgers is larger than that of lower-judgers. The difference is statistically significant in every case. For example, the data show a 20-year gap between where lower-judgers stopped applying "of average age for an adult Italian" and upper-judgers started.

To summarize: *Bonini-vague* asked participants about vague predicates and found a gap between the average responses of truth and falsity-judgers. *Bonini-crisp* asked participants about crisp predicates and found a gap between the average responses of upper and lower-judgers. Although this result is incompatible with gluts, Bonini et al. acknowledge that the data can be explained by at least three different theories of

vagueness: degree theory, truth gaps and epistemicism. Gap and degree theory are then rejected on familiar (e.g., see Williamson 1994) independent grounds. Bonini et al. conclude that participants mentally represented the vague predicates in *Bonini-vague* as they did the crisp predicates in *Bonini-crisp*, a result they claim epistemicism can best explain. In the following sections we outline several serious problems with Bonini et al.'s experiment and the conclusions they draw from it.

1.1 Problem 1: A Question-begging Assumption

Perhaps the most worrisome problem with Bonini et al.'s experiment is a presupposition in the question given to respondents. Recall the instruction sentence from *Bonini-vague*: "Please indicate the smallest height that in your opinion makes it true to say that a man is 'tall'." This language was necessary because the hypotheses given by Bonini et al. make predictions about the area of application for each predicate and its negation. The problem is that the hypotheses, and the questions used to test them, presuppose an epistemicist account of vagueness. The question presupposes that there is a precise boundary between tall and non-tall men, viz. "the smallest height". Although this presupposition is unproblematic for epistemicists and some truth-gap theorists, it is incompatible with many other theories of vagueness. Moreover, some theorists who accept the presupposition claim that we ought not to answer such questions. In fact, Williamson's own theory of vagueness maintains that participants are unjustified in giving any answer to the question posed in his own study; a genuine epistemicist would remain silent if presented with the instruction sentence. It is reasonable to assume that of the participants

who did answer, some would have preferred to remain silent.[4] Although we do not know how this "forced filling-in" affected the data, it is a source of noise that is inherent to the design of the experiment. This makes the design question-begging. It assumes that epistemicism is true, because if epistemicism is not true then the question is nonsense; moreover, even if epistemicism is true, there is simply no way for a participant who does not believe that vague predicates have precise boundaries to respond.

1.2 Problem 2: Ambiguous Questions

The conclusions reached by Bonini et al. are also threatened by an ambiguity in their questions. Consider this part of the 'tall' question put to truth-judgers (emphasis ours):

> ... Please *indicate the smallest height* that in your opinion makes it true to say that a man is 'tall'.
>
> It is true to say that a man is 'tall' if his height *is greater than or equal to* ____ *centimeter*s.

The instruction sentence asks participants to complete a necessity claim by asking (with a superlative) for the smallest height that makes it true to say a man is 'tall'. This is the data required to test the hypotheses given by Bonini et al. But the answer portion asks participants to complete a sufficiency claim by asking (with a comparative) for a number *n* such that it is true to say a man is 'tall' if his height is greater than or equal to *n*. These are different questions. Where the necessity claim challenges participants to give values at the outermost boundaries (if they exist), the sufficiency claim allows participants to avoid making a contentious or uncomfortable judgment by giving a safer value. For

[4]In the next section we describe a revised version of this experiment. Some of our participants left the questions blank, while others wrote unsolicited comments on the instruments citing the apparent absurdity of the question as the reason for their inability or unwillingness to answer.

example, suppose Alex believes that the smallest height that makes it true to say a man is 'tall' is 170 cm. To answer the necessity claim truthfully, Alex must write '170' in the answer portion. But the sufficiency claim allows Alex to give any number greater than 170: after all, it is true to say a man is 'tall' if his height is greater than or equal to 180, 200 or even 500 centimeters. This ambiguity can be found in all of the questions used by Bonini et al. As a result, truth-judgers answering the sufficiency claim may have given responses greater than what they believed was the smallest possible value (assuming they believed there was such a value), and in particular values greater than those they would have provided if presented only with the necessity claim. Similarly, falsity-judgers answering the sufficiency claim may have given responses smaller than what they believed was the greatest possible value (assuming they believed there was such a value), and in particular values smaller than what they would have provided if presented only with the necessity claim. Instead of being a genuine feature of vagueness, the observed gaps may have been created by the ambiguity in the questions used by Bonini et al.

1.3 Revising Bonini et al.

Our hypothesis is that the ambiguity amplified the gaps found in *Bonini-vague* and *Bonini-crisp*. To test our hypothesis we created four different studies, each based on a version of their question. The first two are *Replication-vague* and *Replication-crisp*. Participants in these studies were given English copies of the questions in *Bonini-vague* and *Bonini-crisp*. Results from the *Replication* studies were compared with results from our other two studies, *Revised-vague* and *Revised-crisp*. Participants in the *Revised* studies were only presented with the necessity claim; i.e., they were given a

disambiguated version of the question. We predict that *Replication-vague* will have larger gaps than *Revised-vague,* and that *Replication-crisp* will have larger gaps than *Revised-crisp*.

Participants. 368 undergraduates at the University of Calgary and 345 undergraduates at the University of Toronto participated. The experiments were conducted in 2005 and 2007.

Procedure. *Replication-vague* contained two control groups, truth-judgers$_{Rep}$ (*N*=142) and falsity-judgers$_{Rep}$ (*N*=141). These groups were given questions identical to those in *Bonini-vague* (i.e., ambiguous questions). Their results were compared to those from *Revised-vague*, which was composed of truth-judgers$_{Rev}$ (*N*=80) and falsity-judgers$_{Rev}$ (*N*=78). These groups were given disambiguated questions. For example, truth-judgers$_{Rev}$ were asked

> What is the smallest height a man can be so that he is still tall enough for it to be true to say that he is 'tall'? ____ feet and ____ inches.

Similar questions were asked about 'old' and 'long'. Falsity-judgers$_{Rev}$ were also given disambiguated questions.

Replication-crisp also contained two control groups, upper-judgers$_{Rep}$ (*N*=42) and lower-judgers$_{Rep}$ (*N*=43). These groups were given questions identical to those in *Bonini-crisp* (i.e., ambiguous questions). Their results were compared to those from *Revised-crisp*, which was composed of upper-judgers$_{Rev}$ (*N*=90) and lower-judgers$_{Rev}$ (*N*=97). These groups were given disambiguated questions. For example, upper-judgers$_{Rev}$ were asked

What is the smallest height a man can be so that he is still tall enough to be at least as tall as the average 30-year-old Canadian? ____ feet and ____ inches.

Similar questions were asked about average age and film length. Lower-judgers$_{Rev}$ were also given disambiguated questions.

Results. Table 3 compares data from *Replication-vague* with that from *Revised-vague*. Each box compares average responses from a *Replication* question against those from its *Revised* counterpart. For example, the average 'tall' response of truth-judgers$_{Rep}$ is 182.49cm (*SD*=5.5), whereas the average response of truth-judgers$_{Rev}$ is only 166.88cm (*SD*=19.28). The z-score column shows if there is a significant difference between the truth and falsity-judgers in a given row.

Question	Truth-judgers$_{Rep}$ (*N*=142)/ truth-judgers$_{Rev}$ (*N*=80)		Falsity-judgers$_{Rep}$ (*N*=141)/ falsity-judgers$_{Rev}$ (*N*=78)		z-score	
tall	182.49 cm/ 166.88 cm	*SD*=5.5/ *SD*=19.28	177.30 cm/ 173.53 cm	*SD*=9.44/ *SD*=68.47	*z*=5.44/ *z*=2.47	*p*<.001/ *p*<.05
old	58.60 yr/ 38.71 yr	*SD*=10.90/ *SD*=16.15	52.30 yr/ 43.65 yr	*SD*=12.41/ *SD*=20.97	*z*=5.10/ *z*=1.53	*p*<.001/ N.S.
long	144.47 min/ 99.38 min	*SD*=34.14/ *SD*=43.66	124.08 min/ 106.64 min	*SD*=39.30/ *SD*=32.45	*z*=5.16/ *z*=0.16	*p*<.001/ N.S.

Table 3: Testing the ambiguity in *Bonini-vague* (Mann Whitney *U*-Test)

The control study *Replication-vague* yielded results similar to *Bonini-vague*: we observed statistically significant gaps between truth-judgers$_{Rep}$ and falsity-judgers$_{Rep}$ for all three predicates.[5] These gaps disappeared in *Revised-vague*, where there was no significant gap

[5]Although our observation of gaps mirrored those in *Bonini-vague*, the specific values participants gave were quite different. For example, the average 'tall' response of falsity- judgers$_{Rep}$ was 177.30 cm; it was 164.13 cm in *Bonini-vague*. We ignore this difference because Bonini et al. do not take the specific values obtained to be relevant to their hypothesis. The difference, if statistically significant, would suggest only that Italian students view vague predicates as having different boundaries than their English counterparts, not necessarily that they are mentally represented in different ways.

between truth-judgers$_{Rev}$ and falsity-judgers$_{Rev}$ for 'old' and 'long', and there was a statistically significant *glut* for 'tall'.

Table 4 compares data from *Replication-crisp* with data from *Revised-crisp*.

Question	Upper-judgers$_{Rep}$ (*N*=42)/ upper-judgers$_{Rev}$ (*N*=90)		Lower-judgers$_{Rep}$ (*N*=43)/ lower-judgers$_{Rev}$ (*N*=97)		z-score	
average height	173.52 cm/ 164.9 cm	*SD*=4.51/ *SD*=15.47	170.31 cm/ 167.67 cm	*SD*=7.15/ *SD*=8.79	*z*=2.62/ *z*=2.51	*p*<.01/ *p*<.05
average age	31.76 yr/ 33.99 yr	*SD*=8.47/ *SD*=11.26	35.24 yr/ 41.16 yr	*SD*=8.36/ *SD*=15.91	*z*=1.53/ *z*=2.87	N.S./ *p*<.01
average length	71.26 min/ 77.68 min	*SD*=31.12/ *SD*=57.6	67.78 min/ 78.25 min	*SD*=30.89/ *SD*=36.72	*z*=0.55/ *z*=0.67	N.S./ N.S.

Table 4: Testing the ambiguity in *Bonini-crisp* (Mann Whitney *U*-Test)

Interestingly, the results of *Replication-crisp* differ greatly from those of *Bonini-crisp.* Both studies show a significant gap between upper-judgers$_{Rep}$ and lower-judgers$_{Rep}$ on "average height". But unlike *Bonini-crisp*, *Replication-crisp* shows no statistically significant difference between upper-judgers$_{Rep}$ and lower-judgers$_{Rep}$ for "average age" and "average length". *Revised-crisp* showed statistically significant *gluts* between upper-judgers$_{Rev}$ and lower-judgers$_{Rev}$ for "average height" and "average age", and no significant difference for "average length".

Discussion. The data from each study are summarized in Table 5. Differences from the results of Bonini et al. are in bold.

Study	'tall'/average height	'old'/average age	'long'/average length
Bonini-vague	Gap	Gap	Gap
Replication-vague	Gap	Gap	Gap
Revised-vague	**Glut**	***No sig. difference***	***No sig. difference***
Bonini-crisp	Gap	Gap	Gap
Replication-crisp	Gap	***No sig. difference***	***No sig. difference***
Revised-crisp	**Glut**	**Glut**	***No sig. difference***

Table 5: Summary of Results

The data strongly support the hypothesis that the ambiguity amplified or created the gaps. The gaps found in *Replication-vague* and *Bonini-vague* were completely eliminated in *Revised-vague*.[6] *Revised-crisp* also eliminated the gaps found in *Bonini-crisp*, though comparisons here are made slightly more complicated because the results of *Replication-crisp* did not match those of *Bonini-crisp*. This suggests either that something went wrong in at least one study, or that Canadian students mentally represent crisp predicates differently than their Italian counterparts. Nonetheless, data from these studies still support our hypothesis. We claimed that the ambiguity in the wording used by Bonini et al. made a difference, and it clearly did: the results from *Revised-crisp* differ from those of both *Replication-crisp* and *Bonini-crisp*.

Recall that Bonini et al. used the apparently similar gaps in *Bonini-vague* and *Bonini-crisp* to argue that participants mentally represent vague predicates as they do crisp ones. Although we found some similarities between *Revised-vague* and *Revised-crisp*, they were not of the kind epistemicism predicts. *Revised-vague* showed a

[6]Interestingly, the gaps were eliminated in an unanticipated way. We expected that removing the ambiguity would cause truth-judgers to give smaller values and falsity-judgers to give larger values. But *both* truth and falsity-judgers gave smaller values in *Revised-vague* than they did in *Replication-vague*. However, falsity-judgers$_{Rev}$ didn't drop nearly as much as truth-judgers$_{Rev}$, and the difference was enough to turn the gaps into gluts.

significant glut for 'tall' and no significant difference for 'long'. Similarly, *Revised-crisp* showed a significant glut for "average height" and no significant difference for "average length". This shows not only that the ambiguity affected the data, but also that a disambiguated question yields data inconsistent with epistemicism. The hypothesis that speakers treat vague and crisp predicates identically is supported by 2 of the 3 predicates, but neither shows data that Bonini et al. think the epistemicist would predict. In section 1.5 we argue that these similarities offer no support for epistemicism.

According to the hypotheses advanced by Bonini et al., our data would provide some evidence for a glut theory and tell against epistemicism, gap and degree theory. However, we are not prepared to make this claim ourselves. Although the revised questions eliminated the ambiguity, the question-begging assumption (see section 1.1) is intrinsic to the design of the experiment, and in later sections we outline additional problems that tell against any meaningful interpretation of data. In our view, our data only license the conclusion that the ambiguity created the gaps observed by Bonini et al.; we do not believe that data from this kind of experiment, revised or otherwise, can reliably be used to argue for any theory of vagueness.[7]

1.4 Problem 3: Inappropriate Statistical Methods

In section 1, we described the hypotheses Bonini et al. give on behalf of the epistemicist, gap, glut, and degree theorist. For example, they claim the epistemicist predicts that the average estimates of truth-judgers will be significantly larger than those of falsity-judgers

[7]Alxatib and Pelletier (forthcoming) also criticize Bonini et al.'s experiment. They reinterpret Bonini et al.'s data and argue that it provides partial support for gap theories of vagueness. Although we agree with some of their criticisms, we do not believe that Bonini et al.'s data support gap theories. In our view, the problems with their methodology and experimental design (described throughout section 1) are simply too great for its results to be given any philosophical or psychological import.

(ibid, pp. 379-380). Although each of their hypotheses is framed in terms of *average* estimates, the statistical analysis Bonini et al. use, the Mann Whitney *U*-test (MW), does not test for significantly different averages (i.e., means): it tests for significantly different medians. Bonini et al. are aware of this: '[MW] was employed to evaluate the hypothesis that the medians of the two groups are identical' (ibid, p. 382). In general, MW is a perfectly good test for significance and, in the experiments reported by Bonini et al., it does show that the responses of truth-judgers were significantly different from those of falsity-judgers. But because MW tests for significance by comparing median values, it cannot support a hypothesis about two groups having significantly different means: two sets may have similar means but very different medians, e.g., {100, 101, 102} and {1, 2, 300}. This would be unproblematic if the hypotheses given by Bonini et al. could be reformulated in terms of median values. But their philosophical arguments hinge on truth and falsity-judgers having significantly different *means*.

> The average responses … reveal substantial gaps between the range of values in which the target sentence is deemed true, and those for which it is deemed false. For example, the range of indeterminacy for 'old man' is ten years, which is more than 15% of the size of the region in which the predicate is judged to apply falsely. Gaps of similar size show up throughout our studies. (ibid, p. 382)

The statistical significance of the gaps cannot be ascertained from MW because gap size is defined by Bonini et al. in terms of mean values: it is calculated by subtracting the mean falsity-judger value from the mean truth-judger value. Because Bonini et al. did not test for significantly different means, we do not know if the relevant values are significantly different, and thus we cannot know if the observed gap – whose size is a key

part of the epistemicist hypothesis – is a reliable observation or merely the product of chance.

One of the most common tests that is used to compare mean differences is Student's *t* test, one of a class of parametric tests that use the sample variance within a group of participants to estimate the variance of the population as a whole. In order for this estimate to be accurate, the variance of the participant group must conform to the standard distribution (i.e., a Gaussian distribution). When this assumption is met, parametric tests can provide sufficient power to accurately detect differences between means. Nonparametric tests like MW make no assumptions about the shape of the distribution. This liberates these tests from assumptions of normality, but does so at the cost of specificity. MW tests to see whether two group distributions are different in any way, but unless it can be shown that the two group distributions have an equal shape (e.g., variance, skew and kurtosis) these differences cannot be attributed to differences between the central tendencies of the groups. We ran MW on our data to more closely replicate and revise Bonini et al.'s experiment. However, we were unable to compare these results to those generated by parametric tests (i.e., tests that would check for statistically significant means) because our data sets had large standard deviations. We suspect that Bonini et al. used MW for precisely this reason; their data also has large standard deviations. What MW tells us (and Bonini et al.) is that truth-judgers gave significantly larger answers than falsity-judgers. But the hypotheses given by Bonini et al. require more than this; they require a significant difference in *average* estimates.

1.5 Types of Errors

Bonini et al. claim that the epistemicist, gap and degree theorist all predict that the average estimates of truth-judgers will be significantly larger than those of falsity-judgers.

> According to [the epistemicist], subjects react to questions about vague predicates as they do to questions about sharp predicates whose boundaries they think they do not know. To be more specific, we combine [epistemicism] with the auxiliary assumption that in responding to requests for the smallest number of which the predicate is true, *S* gives the least number for which he is reasonably confident that the predicate applies. In responding to requests for the greatest number of which the predicate is false, *S* gives the largest number for which he is reasonably confident that the predicate fails to apply. Using a statistical analogy, *S* is assumed to prefer type I error over type II. For example, truth-judgers lower the chance of accepting the truth of 'tall' in a region where it fails to apply by raising the chance of failing to accept 'tall' where it does apply. To explain the preference, we may conceive false application of a predicate as an error of commission and incorrect withholding of the predicate as an error of omission. There is evidence that people perceive errors of commission as graver than those of omission (Ritov and Baron, 1990; Spranca et al., 1991) and this would induce reluctance by truth-judgers to descend far down the height-continuum, and reluctance by falsity-judgers to ascend too high. Gaps result (ibid, p. 387).

So, according to Bonini et al., the desire to avoid errors of commission leads truth-judgers to give values above the unknowable boundary (since going too low might lead to

error) and falsity-judgers to give values below the unknowable boundary (since going too high might lead to error). The result is a gap between the responses of truth and falsity-judgers. Clearly the explanation is consistent with the data. But what is needed is an explanation of why the epistemicist specifically predicts gaps.

It is certainly not clear why the epistemicist could not predict gluts or a lack of significant difference between truth and falsity-judgers. For instance, the "average length" question in *Replication-crisp* showed a 4-minute gap between upper-judgers$_{Rep}$ and lower-judgers$_{Rep}$ that was not statistically significant. The epistemicist could plausibly maintain that the real (unknowable) boundary was somewhere in this area. It is very hard to see what the epistemicist gains by requiring that the gap be statistically significant. If anything, statistical *insignificance* provides better evidence for epistemicism than significant gaps: uniform responses between groups shows that truth and falsity-judgers, and upper and lower-judgers, stopped somewhere near the same (unknowable) boundary. Of course, gap (and glut) theorists could give a similar explanation for statistical insignificance: nothing in gap or glut theory requires that the gap (or glut) be statistically significant.

The epistemicist can also account for gluts. Even if people do perceive type I errors (error of commission, false positives) as being graver than type II errors (errors of omission, false negatives), what the epistemicist needs to show is that speakers actively avoid type I errors when using vague language. People make type I errors all the time: e.g., students answer exam questions incorrectly instead of leaving them blank and drivers get in accidents by going too fast or making a wrong turn. Facing gluts, the epistemicist could plausibly claim that the application of vague predicates was just

another circumstance where type II error was prevalent. Gap (glut) theorists could use a similar line of reasoning to explain gluts (gaps), since nothing in these theories requires that speakers know where the gluts (gaps) are, or that they err in a particular way.

So it is difficult to see what is gained by this kind of experiment. Some explanations may be better than others but, nonetheless, each possible outcome can be plausibly explained by nearly every theory of vagueness. This is precisely why Bonini et al. end up giving ordinary theoretical arguments against gap and degree theory: remember that on their own hypotheses, the data is equally compatible with those theories. So it is difficult to see what new information the experiment gave Bonini et al. since, apparently, they believe that gap and degree theory can be ruled out on unrelated grounds.

1.6 Mental Representation of Vague Predicates

Bonini et al. take the gaps in *Bonini-vague* and *Bonini-crisp* as evidence for epistemicism.

> The results support the hypothesis that estimates of an acknowledged, but unknown, boundary are generated in a manner similar to estimates of true and false regions of continua associated with vague predicates. In both cases, people seem to focus on regions which have little chance of straddling the dividing line at issue. This supports the conjecture that the psychological interpretation of vagueness rests on the assumption of a sharp but unknown boundary (ibid, pp. 391-392).

Yet this conclusion would follow only if there was a relationship between gap similarity and psychological interpretation. This need not be the case. Consider the predicate *P* ('equal to two plus two') and a blank predicate *Q* ('average mass of extraterrestrial life').[8] *P* has acknowledged and known boundaries; *Q* may or may not have crisp boundaries, but ordinary speakers know nothing about extraterrestrial life. Suppose we ran an experiment like *Bonini-crisp* with *P* and *Q*. Because both the greatest value for which it is false to say a number is *P* and the smallest value for which it is true to say a number is *P* is 4, there will be no difference between truth and falsity-judgers. Because neither truth nor falsity-judgers know the value of *Q*, their guesses should be statistically identical: a truth-judger is no more likely to give any particular number than a falsity-judger. So *P* and *Q* will show similarly-sized gaps, i.e., none. Irrespective of whether *P* and *Q* are vague, it would not follow from this result that *P* and *Q* had similar psychological interpretations. Presumably, we computed the extension of *P* and guessed the extension of *Q*. Our worry is that the gaps in *Bonini-vague* and *Bonini-crisp* are an equally uninformative measure.

For the sake of argument, let us now suppose that the gaps are philosophically meaningful. It would still not follow that our mental representation of vague predicates assumes a sharp but unknown boundary. The data equally license the conclusion that our mental representation of crisp predicates assumes *vague* boundaries. At best, the data show that we have similar representations of vague and crisp predicates. But it is question-begging to assume that this similarity is a result of their shared crispness; it might just as easily be a result of their shared vagueness. It is no less plausible to hold

[8]A blank predicate is one that speakers know nothing about. This ensures that speaker judgments are completely rooted in ignorance. In contrast, it is likely that participants knew some things about the (purportedly) blank, crisp predicates Bonini et al. asked about: e.g., they knew that the average Italian was not just three feet tall and that the average movie was more than thirty seconds.

that the predicates in *Bonini-crisp* are vague than it is to hold that the predicates in *Bonini-vague* are crisp. Consider the question in *Bonini-crisp* that asked when a man is of at least average height among 30-year-old Italians. Bonini et al. claim this is a crisp predicate. And it may be true that if we fixed the meaning of 'average', 'Italian', and '30-year-old', 'average height among 30-year-old Italians' would be crisp. But this need not be the case.

Speakers can use predicates like 'average' vaguely; i.e., without intending to talk about a mathematically determined mean. Suppose Robert and Susan are at a dinner party in Italy. Robert has just finished reading the latest census and knows that the average Italian is 177.75 cm tall. He asks Susan if she has met Marco. She is not sure and asks what he looks like. Robert replies "He is average height, for an Italian." In this case Robert is using "average height" vaguely. Robert may have no idea what a man who is specifically 177.75 cm tall looks like, and he is certainly not committing himself to any claim about Marco being precisely 177.75 cm tall. He might even know that Marco is only 176.15 cm tall. The description given by Robert is perfectly ordinary. For present purposes it does not matter how we explain this phenomenon – it may be a kind of ambiguity, approximation, context-sensitivity, expressive economy, or perhaps even linguistic incompetence. Speakers use 'average height' in these ways, and it is quite possible that some participants in *Bonini-crisp* did too. These participants would not have been mentally representing crisp predicates at all, and so it would not be surprising to find similarities in responses to *Bonini-crisp* and *Bonini-vague*.

Although we do not know how the participants in *Bonini-crisp* interpreted 'average', the fact that similar gaps were observed in *Bonini-vague* can be seen as

evidence that they read it *vaguely*. Further evidence can be found in the wording of the questions in *Bonini-crisp* (emphasis ours).

> When is a man not as tall as average among 30-year-old Italians? Of course, a man not as tall as average among 30-year-old Italians is *not too tall* and *not too short*.

This is a parallel version of the questions in *Bonini-vague*. While this was probably done to minimize the differences between *Bonini-vague* and *Bonini-crisp*, it invites participants to read 'average height among 30-year-old Italians' vaguely by characterizing 'average height' in terms of the vague predicates 'not too tall' and 'not too short'. So it is not at all surprising that Bonini et al. get similar results in *Bonini-vague* and *Bonini-crisp*, since in both cases participants are asked to give a precise boundary to a vaguely-interpreted predicate.

Each of the arguments we gave in this section is, in our view, sufficient to show that the experiment done by Bonini et al. is methodologically unsound and that its results should not be given any philosophical weight by philosophers of vagueness. However, we wholeheartedly agree with the idea that empirical work can meaningfully inform a philosophical study of vagueness, provided that the experiments are carefully designed from both a philosophical and experimental point of view. In the following sections we present data from three of our own experiments. We have tried to avoid the kinds of problems found in Bonini et al., and we make note of areas where, in hindsight, we believe improvements could be made.

2. The Confusion Hypothesis

Many theories of vagueness use a 'definitely' operator. Sometimes the operator is taken in a technical sense. For example, the supervaluationist notion of supertruth is typically explained in terms of definiteness: it is supertrue that *x* is *P* iff *x* is *P* on every precisification of *P*, which is to say that *x* is definitely *P*. If 'definitely' is just a special operator (e.g., one that distinguishes supertruth from truth), then correspondence with natural language may not be important. But philosophers of vagueness typically want more than this: a good theory of vagueness doesn't only solve problems of vagueness, it explains why we found the problem so compelling in the first place. For example, because the supervaluationist claims that the law of excluded middle holds in cases of vagueness, a good supervaluationist theory will explain away the (typically untested) intuition that ordinary speakers reject instances of it in cases of vagueness. Keefe (2000, p. 164) writes:

> A statement which is similar to the trivial 'either *a* is red or not' but which would be informative (because sometimes false) is 'either *a* is definitely red or definitely not-red', and it could be that the two are sometimes confused: if someone were to assert the former, then, on the assumption that they are obeying the Gricean rule of being informative, it would be reasonable to take them to mean the latter. And the fact that the former is *never* informative could explain why it is so common for our judgments of both sentences to be dictated by our judgments of 'either *a* is definitely red or definitely not-red' and why we thus consider 'either *a* is red or not' to be not true.

Here, Keefe explains away the (purportedly) ordinary intuition that 'either *a* is red or not' is not true by claiming that ordinary speakers confuse it with 'either *a* is definitely red or definitely not-red', which is false on most theories of vagueness.

Keefe is not the only theorist to defend such a hypothesis, though she is perhaps the most well known. Her claim is a version of what is known as a *confusion hypothesis*. Our current interest is in the most general form of the confusion hypothesis, similar to the formulation given by Williams (2006, p. 412).

> The confusion hypothesis maintains that we confuse an utterance of 'There is something that is *F*' with the claim that *there is something that is definitely F*: our intuitions about the former track the truth-values of the latter. The need to explain the seductiveness of the sorites is incumbent on *all* non-logically revisionary treatments of vague language; and the confusion hypothesis requires only that one's favoured treatment allow the construction of the notion 'Definitely'. So, for example, Greenough (2003) defends an epistemic version of the confusion hypothesis; and Edgington (1997) appeals to the same moves in the context of a degree theoretic account. Hence, when supervaluationists such as Fine (1975) and Keefe (2000) postulate confusion to explain the sorites, it is no idiosyncratic idea.

We take Keefe to be giving a restricted version of the hypothesis that applies only to disjunctions that are classical tautologies, and Williams to be giving a generalized non-tautological existential version. If we can show that confusion occurs in the general case, then there will be strong evidence that confusion hypotheses can offer a plausible model of the ordinary intuitions that many theorists want to reject. Of course, additional testing

will be required to test less general models. Data showing that confusion does not occur in the general case will not disprove any more specific model. For example, the restricted confusion hypothesis Keefe describes could hold even if confusion does not occur in the general case, since speakers might only make the confusion in certain disjunctions. But her explanatory burden would be higher, and it would no longer be so easy for a theorist to use an untested confusion hypothesis in defense of her particular theory.

2.1 Experiment #1

We test a generalized version of the confusion hypothesis: speakers use the same truth-value to describe 'x is φ' and 'x is definitely φ', where 'φ' is a vague predicate.[9]

Participants. 350 undergraduates at the University of Calgary participated. The experiment was conducted in 2005.

Procedure. Participants were divided into two groups: *heavy* (*N*=176) was asked about 'heavy' and *rich* (*N*=174) was asked about 'rich'. Groups were given paper surveys to fill out in a classroom setting. For example, the question for *rich* read

> Imagine that on the spectrum of rich women, Susan is somewhere between women who are clearly rich and women who are clearly non-rich.[10] We are

[9]It would be interesting to see if confusion hypotheses can be iterated. For example, one might see if speakers use the same truth-value to describe 'x is definitely φ' and 'x is definitely definitely φ', and so on. We doubt anyone would claim that confusion hypotheses are transitive (e.g., that there is a confusion between 'x is φ' and 'x is definitely definitely definitely definitely definitely φ').

[10]In retrospect, it would have been better to ask about "the spectrum of richness", since "the spectrum of rich women" could be understood as excluding any women who are clearly non-rich. Fortunately, the results do not indicate that participants read the question in this way. If they did, we would expect participants to deny that Susan wasn't rich (see questions (6) and (7) in section 4.1). But fewer than 20% of respondents answered in this way; more participants said it was true that she was "not rich".

interested in your opinion about the status of the following twelve sentences.[11] Please check *one* box only for each.[12]

[] true [] not true, but also not false [] partially true and partially false

[] false [] both true and false [] true or false, but I don't know which

Participants were asked to describe the following sentences

(4) Susan is rich.

(5) Susan is definitely[13] rich.

Results. We test our hypothesis by comparing the responses of each participant to (4) and (5). Table 6 contains data for both *heavy* and *rich*, separated by '/'. The numbers in the boxes show how many participants answered (4) with the title of the row and (5) with the title of the column. For example, 6 participants in *heavy* and 4 in *rich* described both (4) and (5) as neither true nor false.

[11]Participants were also asked about sentences that were not related to this experiment. Those questions are omitted from this list. Some of these additional data are reported in section 4. Originally, (4) was the first listed sentence and (5) the ninth.

[12]Alxatib and Pelletier (forthcoming) worry that giving participants just four choices in an experiment like this ('true', 'false', 'other-valued', and 'can't tell') presupposes a degree-theoretic framework because there would be no way of knowing how "other-valued" was interpreted; they suggest that participants might understand it as 'somewhere in between'. Our experiment has no such ambiguity since participants were explicitly allowed to choose between answers corresponding to gap, glut and degree theory.

[13]Although the terms 'definitely', 'determinately', and 'clearly' are used interchangeably in the literature, we chose to ask about 'definitely' because it is most commonly used. Since we asked about 'definitely', we used 'clearly' in the preamble to avoid extra noise. It would be interesting to see if using 'determinately' or 'clearly' in the preamble (or asking about 'clearly *P*' or 'determinately *P*′) would make a difference. Moreover, 'definitely' can be used in other ways – e.g., as an epistemic modal. Although we tried to make it clear that we were asking about truth, further studies are needed to sort out all of these issues.

heavy/rich	Answer to (5)							
Answer to (4)		True	Neither	Partially	False	Both	Don't know	Total
	True	5/5	0/1	2/1	10/2	0/1	1/0	18/10
	Neither	1/1	6/4	0/1	45/42	1/0	1/2	54/50
	Partially	1/1	3/6	4/1	33/20	0/2	4/1	45/31
	False	1/0	1/0	0/0	20/47	1/0	0/0	23/47
	Both	1/1	1/2	0/2	8/6	3/3	2/1	15/15
	Don't know	1/2	2/0	0/0	13/11	0/0	5/8	21/21
	Total	10/10	13/13	6/5	129/128	5/6	13/12	176/174

Table 6: Data for *heavy* and *rich*

The distribution of responses was statistically significant for both groups (Chi-square, p=.000). The data do not support a generalized confusion hypothesis. Only 24% of participants in *heavy* and 39% of those in *rich* answered (4) and (5) identically. The difference can mainly be attributed to the fact that a majority (73%) of participants in each group described (5) as 'false', while only 13% in *heavy* and 27% in *rich* described (4) as 'false'.

2.2 Experiment #2

In this experiment we compare the judgments of ordinary speakers about the boundaries of 'φ' and 'definitely φ'. Like the questions in Bonini et al. (1999), these questions presuppose that vague predicates have sharp boundaries. However, this presupposition is not relevant to the experiment. We are not interested in the specific values given by participants or how they might respond to the presupposition. The generalized confusion hypothesis predicts that speakers respond to inquiries about the boundary of a vague

predicate 'φ' as they do to inquiries about that of 'definitely φ'. To test this specific hypothesis, the question is as good as any.

Participants. 164 undergraduates at the University of Calgary participated. The experiment was conducted in 2005.

Procedure. We used a modified version of *Bonini-vague*. Participants were divided into four groups and given paper surveys to fill out in a classroom setting. The group *largest* (*N*=41) was asked about the largest value *n* such that 'φ(n)' would not apply; the group *smallest* (*N*=43) was asked about the smallest value *n* such that 'φ(n)' would apply.

largest: How tall could a man be without being considered 'tall'? The largest height a man could be without, in your opinion, being considered 'tall' is ____ feet and ____ inches.

smallest: How short could a man be while still being considered 'tall'? The smallest height a man could be while still, in your opinion, being considered 'tall' is ____ feet and ____ inches.

Responses from *largest* and *smallest* were compared with those from participants asked about 'definitely φ'. The group *def-largest* (*N*=39) was asked about the largest value *n* such that 'definitely φ(*n*)' would apply and the group *def-smallest* (*N*=41) was asked about the smallest value *n* such that 'definitely φ(*n*)' would not apply.

def-largest: How tall could a man be without being considered 'definitely tall'? The largest height a man could be without, in your opinion, being considered 'definitely tall' is ____ feet and ____ inches.

def-smallest: How short could a man be while still being considered 'definitely tall'? The smallest height a man could be while still, in your opinion, being considered ‘definitely tall’ is __ feet and __ inches.

Each group was asked about 'tall', 'old', and 'long'.

Results. By comparing data from *largest* against *def-largest* and *smallest* against *def-smallest,* we obtain six measures (each pair of groups has three predicates to compare). The confusion hypothesis holds that speakers confuse 'φ' with 'definitely φ', so it predicts no significant difference for any of the six measures: participants in *largest* and *def-largest*, and *smallest* and *def-smallest*, should have interpreted their question identically. In contrast, a significant difference would suggest that speakers did distinguish between 'φ' and 'definitely φ'. Table 7 shows the raw data for all groups. We use a one-way ANOVA to test for significance.

Predicate	*def-largest*/ *def-smallest*		*largest*/ *smallest*		*def-largest* v. *largest*/ *def-smallest v. smallest*	
'tall'	181.71 cm/ 184.55 cm	*SD*=4.84/ *SD*=11.39	176.19 cm/ 183.12 cm	*SD*=9.22/ *SD*=22.93	*z*=11.06/ *z*=0.13	<.005/ N.S.
'old'	52.32 yr/ 60.73 yr	*SD*=15.07/ *SD*=16.07	44.93 yr/ 53.84 yr	*SD*=10.97/ *SD*=18.86	*z*=6.27/ *z*=3.24	*p*<.05/ N.S.
'long'	128.58 min/ 162.83 min	*SD*=38.20/ *SD*=129.87	120.49 min/ 132.09 min	*SD*=32.09/ *SD*=67.97	*z*=1.06/ *z*=1.87	N.S./ N.S.

Table 7: Data for Experiment #2

We set aside data from the 'long' comparison between *def-smallest* and *smallest* since the standard deviation is too big for the ANOVA to be a useful measure of significance. Of

the remaining five measures, only two show a significant difference: 'old' and 'long' in *def-largest* and *largest*. The remaining three show no significant difference.

2.3 Discussion

Neither of our experiments support a generalized version of the confusion hypothesis, which predicts confusion in all six cases. Although Experiment #1 showed that speakers distinguished between (4) and (5), one might worry that any confusion between (4) and (5) was neutralized by their apparent juxtaposition on the survey instrument: each participant was asked about both 'φ' and 'definitely φ'. This worry can be set aside by considering Experiment #2, where participants were asked for the boundaries for either 'φ' or 'definitely φ'. Here, we found that in three out of five pairs speakers gave statistically identical boundaries to 'φ' and 'definitely φ'. That this did not occur in the remaining two cases suggests that the confusion hypothesis is not generalizable. Moreover, we cannot be certain that confusion is responsible for the lack of significance in the three comparisons where no difference was found, since it is possible that 'φ' and 'definitely φ' *sometimes* share a boundary. Nonetheless, the data do show that confusion occurs in at least some cases. This leaves open many possibilities. For example, these data do not tell against restricted versions of the hypothesis (e.g., Keefe's, see section 4). Further empirical research will be required to more precisely characterize the phenomenon.

3. Fuzzy Logic and the Inductive Premise of the Sorites

Degree theorists typically hold that notions of fuzziness and partial truth are rooted in common sense. This is especially true in computer science[14]; philosophical accounts are more tempered. For example, Machina claims that his 'inclinations [about degrees of truth] are at least verbally in agreement with the common sense view,' though he recognizes 'that agreement cannot be taken at face value as an indication that the common man thinks of degrees of truth in the same way' (Machina 1976, p. 54). Arguments like this give fuzzy logic much to prove in the empirical arena.

The inductive premise of the Sorites argument is typically expressed as a conditional, e.g.,

$S\rightarrow$ If a is tall, then a' is tall.

In classical logic, $S\rightarrow$ is equivalent to both

$S^{\wedge}$ It is not the case both that a is tall and a' is not, and

$S^{\vee}$ Either a is not tall or a' is.

This equivalence does not hold in fuzzy logic, where $S^{\wedge}$ is equivalent to $S^{\vee}$ but not $S\rightarrow$.[15] In every Sorites series $S\rightarrow$ will always have a truth-value near 1, while the truth values of $S^{\wedge}$ and $S^{\vee}$ will go as low as 0.5. The degree theorist holds that the Sorites fails because the inductive premise is not completely true; it looks persuasive because we 'confuse truth for

[14]Perhaps this is why degree theory is the only theory of vagueness that has been put to use. For over three decades fuzzy logic has been used in computer science applications to model vagueness (e.g., see Langari et al. (eds.), 1995), albeit at varying levels of complexity and to varying degrees of success. Elkan (1994) concluded that the fuzzy logic component of such systems had no role in their successes. Although his arguments were largely debunked (see the responses accompanying his paper, and Serchuk 2008), his general conclusions are still well-regarded in mainstream AI.

[15]The standard definitions for fuzzy operators are:

$t(P \vee Q) = \max(t(P), t(Q))$

$t(P \wedge Q) = \min(t(P), t(Q))$

$t(P\rightarrow Q) = \min(1, 1 - t(P) + t(Q))$

near-truth' (Weatherson 2005, p. 61). As Edgington (2001, p. 375) puts it, 'the difference between clear truth and almost clear truth – between 1 and 0.99 – is an insignificant difference upon which, normally, nothing hangs.'

Weatherson (2005) holds that the degree theorist cannot explain the plausibility of each form of the inductive premise of the Sorites. He argues that even if fuzzy logic plausibly models $S\rightarrow$, it does not give a plausible account of $S^{\wedge}$ and $S^{\vee}$. On his view, the degree theorist cannot claim that we confuse truth for near-truth when confronted with $S^{\wedge}$ and $S^{\vee}$, since their truth-values are much lower than those of $S\rightarrow$. Weatherson gives a thought experiment in support of this thesis. His goal is to show that $S^{\wedge}$ is the most plausible version of the inductive premise, followed by $S\rightarrow$ and $S^{\vee}$. This is a strong argument against the degree theorist, for her explanation of the Sorites depends on there being a (rough) correlation between degrees of truth and plausibility.

> [The fuzzy logician] has no explanation for why premises like $S^{\wedge}$ look persuasive. This is quite bad, because $S^{\wedge}$ is *more* plausible than $S\rightarrow$, as I'll now show. Consider the following thought experiment. You are trying to get a group of (typically non-responsive) undergraduates to appreciate the force of the Sorites paradox. If they don't feel the force of $S\rightarrow$, how do you persuade them? My first instinct is to appeal to something like $S^{\wedge}$. If that doesn't work, I appeal to theoretical considerations about how our use of *tall* couldn't possibly pick a boundary between *a* and *a'*. I think I find $S\rightarrow$ plausible *because* I find $S^{\wedge}$ plausible, and I would try to get the students to feel likewise. There's an asymmetry here. I wouldn't defend $S^{\wedge}$ by appealing to $S\rightarrow$, and I don't find $S^{\wedge}$ plausible because it follows from $S\rightarrow$ … I don't think anyone has put

> forward a Sorites argument, where the major premises are like $S^{\vee}$ … There's a good reason for this: $S^{\vee}$ is *not* intuitively true, unless perhaps one sees it as a roundabout way of saying $S^{\wedge}$. In this respect it conflicts quite sharply with $S^{\wedge}$, which *is* intuitively true (ibid, pp. 61-63).

The result of Weatherson's thought experiment is not obvious.[16]

3.1 The Experiment

We devised an empirical test of a version of Weatherson's thought experiment. We tested the claim that $S^{\wedge}$ is more persuasive than $S\rightarrow$, and that $S^{\wedge}$ is more persuasive than $S^{\vee}$.

Participants. 243 undergraduates at the University of Calgary participated. The experiments were conducted in 2005.

Procedure. Participants were divided into two groups. The group *heap* (N=119) was asked about 'heap' and *rich'* (N=124) about 'rich'. Groups were given paper surveys to fill out in a classroom setting. The question given to participants in *heap* read:

> 'Heap' is a vague concept: it seems that our use of the word 'heap' does not determine a number X so that any collection of X grains of sand or more are a heap, and anything with fewer than X grains of sand is not a heap (such a number X would be a 'borderline'). Consider the following sentences, where X stands for an arbitrary number. We'd like to know which, in your opinion, express the vagueness of 'heap' most persuasively and which ones the least. Please rank them in order of persuasiveness on the table below. Please break ties.

[16]For example, Machina (1976) claimed it was a point in *favour* of fuzzy logic that it gave $S\rightarrow$ a greater degree of truth than $S^{\wedge}$.

> [The conjunctive form of the inductive premise for 'bald'] essentially says that it never happens that of two persons differing by just a hair that it can be said that one is bald and the other isn't. Since [in fuzzy logic] it can be somewhat true as well as somewhat false that *one* individual is bald, when his hair is very sparse, it can naturally also be quite true that *two* individuals, roughly alike, are both bald to some extent and not bald to some extent. Hence, the low truth value of [the conjunctive form] which denies this can happen (ibid, pp. 200-201).

(A) If *X* grains of sand are a heap, then *X -1* grains of sand are also a heap.

(B) The following statement is false: *X* grains of sand are a heap, but *X-1* grains of sand are not a heap.

(C) *X* grains of sand does not mark the borderline between being a heap and not being a heap.

(D) Either *X* grains of sand are not a heap or *X-1* grains of sand are a heap.

Option (A) corresponds to $S\rightarrow$, (B) to $S^{\wedge}$, (C) to theoretical considerations (*TC*), and (D) to $S^{\vee}$. This does not precisely match the setting of Weatherson's thought experiment. We decided it would be too difficult to give paper surveys to non-philosophers explaining the Sorites and its prerequisite logical concepts. So we instead asked participants to decide how persuasively the different formulations of the inductive premise expressed the phenomenon of vagueness as described in ordinary English. We believe this sufficiently captures the spirit of Weatherson's thought experiment.

Results. We refer to collapsed data (*N*=243) since no significant difference was found between the responses of participants in *heap* and *rich'*. We test the hypotheses in two ways. First, we consider the mean ranking of the answers. This is shown in table 8.

Option	Mean	*SD*
TC	1.840	1.054
$S\rightarrow$	2.247	0.960
$S^{\wedge}$	2.679	0.960
$S^{\vee}$	3.235	0.995

Table 8: Mean Rankings (*N*=243)

The mean for each answer corresponds with its average placement in an ordering, with 1 denoting the first ranking slot (most persuasive) and 4 the fourth (least persuasive). For example, if answer A had a mean of 1.000 it would mean that every participant ranked it as most persuasive. If A had a mean of 2.500, it would mean that, on average, equally many participants ranked it second and third (or first and fourth). We found statistically significant differences (p = .000, paired sample T-test) in the mean rankings of every possible sentence pair; i.e., the mean ordering of the sentences is statistically significant for all possible pairs. The data support the hypothesis that $S^{\wedge}$ is more persuasive than $S^{\vee}$ (since 2.679 < 3.235) but not $S\rightarrow$ (since 2.679 > 2.247).

We can also test the hypotheses by measuring the relative plausibility of each answer pair. This is done in table 9.

Row was ranked as more persuasive than *column*	*TC*	$S\rightarrow$	$S^{\wedge}$	$S^{\vee}$
TC	-	158 (65.0%)	175 (72.0%)	192 (79.0%)
$S\rightarrow$	85 (35.0%)	-	153 (63.0%)	188 (77.4%)
$S^{\wedge}$	68 (28.0%)	90 (37.0%)	-	163 (67.1%)
$S^{\vee}$	51 (21.0%)	55 (22.6%)	80 (32.9%)	-

Table 9: Relative Plausibility (*N*=243)

Each box in the table shows the number and percent of participants who ranked the option corresponding to the row as being more persuasive than the option corresponding to the column. For example, 158 participants ranked *TC* as being more persuasive than $S\rightarrow$. As before, the data support the hypothesis that $S^{\wedge}$ is more persuasive than $S^{\vee}$ but not

$S\rightarrow$. 67.1% of participants thought $S^{\wedge}$ was more persuasive than $S^{\vee}$, but only 37% thought $S^{\wedge}$ was more persuasive than $S\rightarrow$.

Discussion. The data show that Weatherson's thought experiment (or, at least, our version of it) fails to capture the intuitions of ordinary speakers about the Sorites. Although fuzzy logic plausibly models $S\rightarrow$ as having a greater degree of truth than $S^{\wedge}$, Weatherson is right that the equivalence between $S^{\wedge}$ and $S^{\vee}$ is counterintuitive. Importantly, any logic with DeMorgan's law cannot 'tell a story about why [$S^{\wedge}$] is intuitively plausible that does not falsely predict [$S^{\vee}$] is [too]' (ibid, p. 63). While Weatherson might be right that this is a general problem for logics with DeMorgan's law, it is particularly damaging for fuzzy logic. Part of what is supposed to make fuzzy logic appealing is its (apparent) correspondence with pre-theoretic ways of thinking about vagueness. The data suggest that this is not always the case.

It could be suggested that participants were influenced by the awkwardness of some sentences, particularly the conjunction (B) and disjunction (D). The worry here is that speakers ranked the sentences not according to the relevant semantic criteria, but according to what sounded more natural or was easier to understand. We grant that some sentences are less ordinary and natural than others, and we acknowledge that speakers may have ranked sentences in this way – we did not ask participants to explain their ordering. But this is no threat to our conclusion. Weatherson's thought experiment involves explaining the Sorites to untutored undergraduates using the unexplained notion of 'intuitive truth'. It seems to us that the awkwardness of the disjunction is precisely why an untutored undergraduate would not find it persuasive. Of course the disjunction

becomes compelling once you understand its (classical) semantic features; Weatherson recognizes that it becomes 'intuitively true' once one comes to see 'it as a roundabout way of saying $S^{\neg}$' (p. 63). What is interesting is whether it is intuitively true *beforehand*, and in such a case criteria like naturalness are perfectly appropriate.

The degree theorist is not without options. She might grant that persuasiveness is a valid criterion, but maintain that our experiment did not test the most natural candidates. For example, perhaps speakers would find a negated conjunction of the form "it does not happen that" more compelling than the "the following statement is false" form we tested. Additional empirical work would be needed to evaluate the many different possibilities. The degree theorist could also deny that $S^{\vee}$ is a valid form of Sorites reasoning, though she would probably have to reject DeMorgan's law too. An option here is to use Łukasiewicz strong conjunction, where $t(P \,\&\, Q) = \max(0, t(P) + t(Q) - 1)$. This ensures $S^{\wedge}$ is given a higher truth-value than $S^{\vee}$, but at the expense of giving $S^{\wedge}$ an identical truth-value to $S{\rightarrow}$.[17] The best option available to the degree theorist is probably to just bite the bullet and admit that her treatment of $S^{\vee}$ is implausible, but maintain that this is an acceptable price for a plausible model of $S{\rightarrow}$ and $S^{\wedge}$. As Weatherson observes, nobody has 'put forward a Sorites argument where the major premises are like $S^{\neg}$' (ibid, p. 63). No doubt this is because sentences like $S^{\vee}$ are too awkward for ordinary use. This suggests that while fuzzy logic may not be the right logic of vagueness, it might be close enough for some purposes (e.g., as a heuristic).

[17]MacFarlane (2010) rejects this move on independent grounds. Mirroring Machina's argument (see n. 16), he gives the example of Borderline Jim. Jim is borderline in multiple categories, e.g., baldness, tallness, intelligence, etc. Let P denote 'Jim is bald' and Q denote 'Jim is tall', where $t(P)=0.5$ and $t(Q)=0.5$. Then the strong conjunction of P and Q will be completely false, since $t(P \,\&\, Q) = 0$. MacFarlane, citing Schiffer (2003), rightly observes that 'it seems perfectly appropriate to endorse the conjunctive proposition that Jim is tall and bald ... to about the same (middling) degree as we endorse the conjuncts separately' (p. 13). This is the result given by normal fuzzy conjunction, where $t(P \wedge Q) = 0.5$.

4. Types of Negation

We will assume that however negation works in borderline cases, the negation of a proposition *P* is clearly (i.e., not borderline) true when *P* is clearly false, and is clearly false when *P* is clearly true. This leaves open the status of negation in borderline cases. Horgan (1994, p. 165) proposes that we distinguish between strong and weak negation.

> We want statements like [the inductive premise of the Sorites] to turn out neither true nor false ... it is natural and useful to enrich the object language ... by adding another form of negation. Let ⌐Φ be true when it's not the case that Φ is true; Φ itself might be false, or might lack truth value altogether. Call this *weak* negation. *Strong* negation[18], by contrast, will work in the manner of negation in classical logic: ~Φ will be true when Φ is false ... Although these two forms of negation do not seem to have cleanly distinguishable modes of expression in ordinary language, I do think they both *occur* in ordinary language. So I now stipulate the following usage, to apply henceforth in this paper: 'it's not the case that' is to be understood as the ordinary-language counterpart of ⌐, whereas other negation constructions in English will be counterparts of ~.

Horgan uses this framework to model the inductive premise of the Sorites. On his view, neither the inductive premise nor its strong negation is either true or false; their weak negations are therefore true.[19] A third candidate (not considered by Horgan) is Gödel

[18]Strong negation is also known as Kleene negation. See Kleene 1952.

[19]For example, Horgan claims that both the inductive premise "For any *n*, if an *n*-haired person is bald then an (*n+1*)-haired person is bald" and its strong negation are neither true nor false. Thus their weak negations, "It's not the case that for any *n*, if an *n*-haired person is bald then an (*n+1*)-haired person is bald" and "It's not the case that not every *n* is such that if an *n*-haired person is bald then an (*n+1*)-haired person is bald" are true (Horgan, p. 165).

(intuitionistic) negation, $\sim_g P$: $\sim_g P$ is clearly true iff P is clearly false, and false otherwise. The three options are summarized in Table 10.

P	Strong ($\sim P$)	Weak ($\neg P$)	Gödel ($\sim_g P$)
True	False	False	False
Other	Other	True	False
False	True	True	True

Table 10: Truth Tables for Three Kinds of Negation

Suppose Bill is borderline tall such that (on a given theory of vagueness) T(b) (i.e., Bill is tall) is neither true nor false. Then the strong negation $\sim T$(b) (i.e., Bill is not tall) will also be something other than true or false, the weak negation $\neg T$(b) (i.e., It is not the case that Bill is tall) will be true, and the Gödel negation $\sim_g T$(b) will be false. There is no proposed linguistic form for Gödel negation. Obviously there are real technical and syntactic differences between these negations. Not only do they have different truth conditions, but their syntactic forms vary. Weak negation negates an entire sentence, whereas strong negation negates only the predicate. We devised an experiment to see if there is a corresponding semantic difference in ordinary language.

4.1 The Experiment

The first experiment tests the hypothesis that speakers treat negation in at least two different ways. The hypothesis would be supported if speakers assign different truth-values to 'not φ' and 'it is not the case that φ'. We also want to see if strong negation is typically expressed by 'not' and weak negation by 'it is not the case that'.

Participants. 350 undergraduates at the University of Calgary participated. The experiments were conducted in 2005.

Procedure. See section 2.1 for an explanation of the preamble and answer set. We re-used groups *heavy* (*N*=176) and *rich* (*N*=174). In this experiment, participants were asked about the truth of the following six sentences.[20]

(6) Susan is not rich.

(7) It is not the case that Susan is rich.

(8) Either Susan is rich or Susan is not rich.

(9) Susan is rich or it is not the case that Susan is rich.

(10) Susan is rich and Susan is not rich.

(11) Susan is rich and it is not the case that Susan is rich.[21]

Although we use Horgan's 'it is not the case that' to express weak negation in ordinary language, Horgan is only committed to the existence of weak negation, not to its having a particular linguistic form in ordinary language.

Results. We test the hypothesis by comparing the responses of each participant to pairs (6) and (7), (8) and (9), and (10) and (11). Table 11 shows responses from *heavy* and *rich*.

[20]Participants were also asked about sentences that were not related to this experiment. Those questions are omitted from this list. Some of these additional data are reported in section 2. Sentences (6)-(11) were presented in a different order.

[21]Although our focus is on negation, this is one of the first attempts to empirically test ordinary intuitions about the law of excluded middle and the law of non-contradiction.

heavy/rich	Answer to (7)							
Answer to (6)		True	Neither	Partially	False	Both	Don't know	Total
	True	11/29	0/1	0/1	6/4	0/1	1/1	18/37
	Neither	20/14	18/16	5/2	5/4	1/1	5/6	54/43
	Partially	14/10	6/8	12/10	11/2	1/0	5/3	49/33
	False	4/16	0/1	3/0	14/9	0/1	3/0	24/27
	Both	1/3	1/1	0/1	3/3	4/4	3/1	12/13
	Don't know	2/2	1/3	0/0	2/2	0/1	14/13	19/21
	Total	52/74	26/30	20/14	41/24	6/8	31/24	176/174

Table 11: Responses to (6) and (7), *N*=176/174

The data support our hypothesis: 59% of participants in *heavy* and 53% in *rich* gave different answers to (6) and (7).[22] Some theorists hold that (6), a case of strong negation, is something other than 'true' or 'false'. This leaves open the question of what the strong negation of such sentences is. We captured this by allowing participants to choose between 'partially', 'both', and 'neither'.[23] 65% of participants in *heavy* and 51% in *rich* answered (6) as predicted, i.e., with one of the four values other than 'true' or 'false'. Some theorists also hold that (7), a weak negation, is true. Yet only 29% of *heavy* and 43% of *rich* described (7) as 'true'. Neither form is a plausible candidate for Gödel negation, where 'false' is the predicted result.

[22]The number of participants who gave different answers to a given pair is the minimum number of participants who are guaranteed to have distinguished between members of that pair. Some respondents who described each member identically may have distinguished between the two forms of negation, but thought they had identical truth-values.

[23]Because of the brevity of an early abstract that they accessed, Alxatib and Pelletier (forthcoming) mistakenly claim that we only gave subjects three or four options (ibid, p. 9). This leads them to (wrongly) conclude that we cannot tell what kind of negation participants were using. In fact, we gave participants six options (see section 2.1) precisely so we could tell what kind of negation they were using. Participants using strong negation would have chosen 'false', participants using weak negation would have chosen 'both', 'neither', or 'partially', and participants who (as Alxatib and Pelletier worry) wanted to '[advert] not being able to tell' (p. 9) would have chosen 'don't know'.

Table 12 shows responses to (8) and (9). We now report collapsed data (*N*=350) since there is no significant difference between the responses of *heavy* and *rich* to questions (8)-(11).

	Answer to (9)							
Answer to (8)		True	Neither	Partially	False	Both	Don't know	Total
	True	67	3	5	17	6	15	113
	Neither	7	12	2	7	2	4	34
	Partially	12	3	4	5	6	5	35
	False	41	14	12	56	7	7	137
	Both	7	4	2	3	2	1	19
	Don't know	7	0	0	0	1	4	12
	Total	141	36	25	88	24	36	350

Table 12: Data for (8) and (9), *N*=350

The data support our hypothesis: 59% of participants gave different answers to (8) and (9). Some theorists claim that (8) is something other than 'true' or 'false', since neither disjunct is true or false. Speakers did not interpret (8) in this way: only 29% of participants gave such a response. Although one might expect that speakers think (8) is true, most did not describe it as such. In fact, a plurality (39%) of respondents described (8) as 'false'. This is a strange result, particularly because only a handful of respondents described either of (8)'s disjuncts as 'false' (see Table 6 and Table 11). However, this result is consistent with Keefe's restricted confusion hypothesis (see section 2). She predicts that speakers will think (8) is 'not true' (Keefe 2000, p. 164) because they confuse '*Fa* or ~*Fa*' with 'definitely *Fa* or definitely ~*Fa*'. Her claim is partially supported by our data, which show that 68% of participants described (8) as something other than

'true'. Most theorists claim that (9) is true, since one of its disjuncts is. This only partially accords with use: a plurality (40%) of respondents described (9) as 'true'. This suggests that weak negation is less susceptible than strong negation to Keefe's restricted confusion hypothesis, though because we did not ask about 'definitely *Fa* or definitely ~*Fa*' there is insufficient evidence to know if it holds at all.

Table 13 shows responses to (10) and (11).

	Answer to (11)							
Answer to (10)		True	Neither	Partially	False	Both	Don't know	Total
	True	13	1	10	33	0	9	66
	Neither	2	9	0	10	5	2	28
	Partially	1	2	3	18	3	3	30
	False	4	6	1	173	2	9	195
	Both	2	1	3	10	3	2	21
	Don't know	3	0	0	3	0	4	10
	Total	25	19	17	247	13	29	350

Table 13: Data for (10) and (11), *N*=350

Because (10) and (11) have the same-truth value (something other than true or false), our hypothesis must change: we would not expect (10) and (11) to be described differently. And indeed they are not: 59% of participants answered (10) and (11) identically. An answer other than 'true' or 'false' was only given by 22% of respondents to (10) and 14% to (11). The data are consistent with Gödel negation, where 'false' is the predicted result: this response was given by 56% of respondents to (10) and 71% to (11). Surprisingly, speakers were more likely to describe the classical contradiction as 'false' when it was generated with a *weak* negation. This seems to contradict our earlier finding that strong

negation was more susceptible to confusion, since if a similar confusion occurred here we would expect more participants to describe (10) as 'false'.

4.2 Discussion. The experiment supports the notion that there are at least two forms of negation in natural language. But speakers did not always use negation as predicted. Speakers who distinguished between 'not' and 'it is not the case that' did not always treat 'not' as a strong negation and 'it is not the case that' as a weak negation. Neither form was consistently read as a Gödel negation. This suggests that the standard semantic treatments of negation do not consistently correspond with our ordinary use of negation in borderline cases. This offers tentative support to the view advanced by Alxatib and Pelletier (forthcoming). Instead of holding that different negations correspond to different linguistic forms, they argue that a sentence like

> '*a* is not tall' is actually three-ways ambiguous: on one reading, the negation is identified with [strong] negation ... on a second reading 'not' is identified with [weak] negation ... and on a third reading 'not' is identified with intuitionistic [Gödel] negation. (ibid, p. 25)

On this view, the inconsistency in use that we observed could be explained by appealing to an ambiguity in the linguistic forms of negation.

Data from (8) and (9) suggest that speakers are willing to violate the law of excluded middle in borderline cases. Responses to (10) and (11) show that they tend to preserve the law of non-contradiction. These results differ greatly from those found by Alxatib and Pelletier (forthcoming). When presented with an image of borderline tall man, over 44% of their respondents described him as 'tall and not tall' (ibid, p. 26). In

contrast, our 'both' option was selected by only 19% of respondents to (10) and by 7% of respondents to (11). Unfortunately, Alxatib and Pelletier report very little statistical data and, in particular, they do not show that the data cited above is statistically significant; nevertheless, the difference is quite striking.[24] Moreover, there are some important differences between our experimental design and theirs. For example, our borderline case is linguistically stipulated, whereas theirs is shown visually. So contextual details seem to play a major role in how plausible contradictions are to ordinary speakers. Further research should be conducted to better understand this phenomenon.[25]

There is an important connection between these experiments and those in section 2. The version of the confusion hypothesis that one subscribes to, and the interaction it predicts between negation and 'definitely', will determine how one reads our results. For example, consider the result that participants were more likely to think (7) was true than they were to think that (6) was. This could be explained by a confusion hypothesis if participants confused (7) with the much weaker claim "it is not the case that Susan is definitely rich". On the other hand, there doesn't seem to be any evidence of confusion in (6): if participants read (6) as "Susan is not definitely rich" we'd expect to see lots of 'true' responses (which we didn't), and if they read it as "Susan is definitely not rich" we'd expect to see lots of 'false' (which we didn't). The interaction between negation and confusion hypotheses is complex, and more focused testing is required before any conclusions about it can be generated from these data.

[24]Alxatib and Pelletier (ibid, p. 22, n. 14) report one statistical measure for one question and one pair of answers. No statistical data is provided for any of the other measures.
[25]See Ripley (2009) for unpublished experimental data on the responses of ordinary speakers to contradictions in borderline cases.

5. Conclusion

It was once the standard view that philosophical theories should be evaluated on the basis of conceptual analysis alone. Although this view has been tempered, there is still no consensus concerning the role of empirical data in the construction of philosophical theories. Although we did not set out to justify the use of empirical data by philosophers, we believe that our results show the value of considering these data when developing and defending theories of vagueness. We have seen how claims about what is intuitive can so easily be wrong, and there is no reason to think that such mistakes are rare; we could, no doubt, find suspect claims throughout the literature. Moreover, it is quite clear that philosophers of vagueness are interested in theories that have at least some bearing on ordinary language. Priest writes:

> The meanings of vague predicates are not determined by some omniscient being in some logically perfect way. Vague predicates are part of *our* language. As a result, their meanings must answer in the last instance to the use that *we* make of them (Priest 2004, p. 13).

Of course, it does not follow from this that empirical work, let alone experiments and paper surveys, are the *best* way of learning about our language. We must work harder to ensure that our experiments are rigorously designed and statistically sound. In section 1 we showed how an empirical project can fail to be philosophically relevant when these conditions are not met. The data presented by Alxatib and Pelletier are genuinely interesting, but because they do not test for statistical significance we simply cannot know if they are philosophically viable: scientists, psychologists and statisticians have

known for a very long time that it is not enough simply to discover that *x*% of participants do some particular thing. One of our referees commented that 'experimental philosophy is no longer in need of a basic defense'. This may be true, but philosophers who run experiments (including ourselves) clearly have a lot to learn. Nevertheless, in sections 2, 3 and 4, we showed that empirical linguistic data has an important role to play in the development of philosophical theories of vagueness. This is especially true when theorists make claims about the beliefs and practices of ordinary speakers. And given the current state of empirical work in the field – namely, the almost complete lack of it – we think that this methodology and our results offer an innovative and underused way of thinking about problems of vagueness.

Department of Philosophy

University of Toronto

Department of Psychology

University of Calgary

Department of Philosophy

University of Calgary

References

Alxatib, S. and Pelletier, F. J.: The psychology of vagueness: Borderline cases and contradictions. Mind & Language, forthcoming.

Bonini, N., Osherson, D., Viale, R., and Williamson, T. 1999: On the psychology of vague predicates. *Mind & Language*, 14:4, 377-393.

Edgington, D. 1997: Vagueness by degrees. In R. Keefe and P. Smith (eds), *Vagueness: A Reader*, Cambridge, MA: MIT Press.

Edgington, D. 2001: The philosophical problem of vagueness. *Legal Theory*, 7:4, 371-378.

Fine, K. 1975: Vagueness, truth and logic. *Synthese*, 30, 265-300.

Greenough, P. 2003: Vagueness: A minimal theory. *Mind*, 112, 235-281.

Keefe, R. 2000: *Theories of Vagueness.* Cambridge: Cambridge University Press.

Kleene, S. 1952: *Introduction to Metamathematics*. Amsterdam: North Holland.

MacFarlane, J. 2010: Fuzzy epistemicism. In Richard Dietz and Sebastiano Moruzzi (eds.), *Clocks and Clouds*. Oxford: Oxford University Press, 438-463.

Machina, K. 1976: Truth, belief and vagueness. *Journal of Philosophical Logic*, 5, 47-48.

Parikh, R. 1991: A test for fuzzy logic. *ACM SIGACT News*, 22:3, 49-50.

Parikh, R. 1994: Vagueness and utility: the semantics of common nouns. *Linguistics and Philosophy*, 17, 521-535.

Raffman, D. Forthcoming. *The Many Ways of Vagueness*.

Ripley, D. 2009: Contradictions at the borders. Unpublished. Available at http://sites.google.com/site/davewripley/papers.

Ritov, I. and Baron, J. 1990: Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, 3, 263-77.

Schiffer, S. 2003: *The Things We Mean*. Oxford: Oxford University Press.

Serchuk, P. 2008: Elkan's 'Paradox' and the Correctness of Fuzzy Logic. *Journal of Multiple-Valued Logic and Soft Computing*, 14:1-2, 33-50.

Spranca, M., Minsk, E. and Baron, J. 1991: Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27, 76-105.

Weatherson, B. 2005: True, truer, truest. *Philosophical Studies*, 123, 47-70.

Williams, J. R. G. 2006: An argument for the many. *Proceedings of the Aristotelian Society*, 103:3, 409-417.