Are *most* and *more than half* truth-conditionally equivalent?*

Milica Denić and Jakub Szymanik

Abstract

Quantifying determiners *most* and *more than half* are standardly assumed to have the same truth-conditional meaning. Much work builds on this assumption in studying how the two quantifiers are mentally encoded and processed (Hackl, 2009; Pietroski et al., 2009; Szymanik and Zajenkowski, 2010; Lidz et al., 2011; Steinert-Threlkeld et al., 2015; Talmina et al., 2017). There is however empirical evidence that *most* is sometimes interpreted as 'significantly more than half' (Ariel, 2003, 2004; Solt, 2011, 2016; Ramotowska et al., 2020). Is this difference between *most* and *more than half* a consequence of pragmatic strengthening of *most*, or is the standard assumption that the two quantifiers are truth-conditionally equivalent wrong? We report two experiments which demonstrate that *most* preserves the 'significantly more than half' interpretation in downward-entailing environments, which speaks against the pragmatic strengthening option and in favor of there being a difference between the two quantifiers at the level of truth conditions.

1 Introduction

Natural language quantification has been at the heart of study of formal semantics and logic (Mostowski, 1957; Barwise and Cooper, 1981; Westerståhl, 1985; Keenan and Stavi, 1986; Van Benthem, 1986; Higginbotham, 1994, a.o.), as well as of cognitive science and the psychology of reasoning (Braine and O'Brien, 1998; Chater and Oaksford, 1999; Johnson-Laird, 1983; Johnson-Laird and Bara, 1984; Geurts, 2003; Geurts and van Der Slik, 2005; Szymanik, 2016, a.o.). At the intersection of these fields, proportional quantifiers such as *most* and *more than half* have occupied a prominent place. The reason for this is two-fold.

Firstly, such quantifiers cannot be treated within standard first-order predicate logic, which has led to the use of the more powerful framework of generalized quantifiers in linguistics to accommodate them (Mostowski, 1957; Lindström, 1966; Barwise and Cooper, 1981).

Secondly, in relation to specifically the proportional quantifiers *most* and *more than half*, there is an important line of research which uses these two quantifiers as a case study to investigate their verification strategies and further proposals about how natural language quantifiers are mentally encoded and processed (Hackl, 2009; Pietroski et al., 2009; Szymanik and Zajenkowski, 2010;

*Many thanks to Fausto Carcassi, Alexandre Cremers, Manuel Križ, Leendert van Maanen, Mora Maldonado and Sonia Ramotowska for helpful comments and suggestions, as well as the audiences of SuB 2020 and the MLC seminar at the University of Amsterdam. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. STG 716230 CoSaQ.

Lidz et al., 2011; Steinert-Threlkeld et al., 2015; Talmina et al., 2017). The core assumption of this work is that the two quantifiers are truth-conditionally equivalent. These authors attribute differences in verification strategies between the two quantifiers to factors other than truth conditions, such as for instance the format of mental encoding of these truth conditions.

In this work, we challenge the assumption that *most* and *more than half* are truth-conditionally equivalent. We do so by studying how the two quantifiers are interpreted in linguistic environments with different logical properties in an experimental setting. This challenge calls for a greater care in how the results pertaining to cognitive processing of *most* and *more than half* are interpreted.

Challenging the truth-conditional equivalence assumption of *most* and *more than half* is also relevant more broadly. Namely, these two quantifiers exhibit a palette of curious properties which do not immediately follow from their generalized quantifier treatment (Solt, 2011, 2016). Understanding what underlies these properties may thus be fruitful more generally for our understanding of natural language quantification.

2 *Most* and *more than half*

Quantifiers *most* and *more than half* are commonly assumed to have the same truth-conditional meaning, as in (1). According to (1), if more than 50% of As are Bs, both (2a) and (2b) are true. 50% will henceforth be referred to as the threshold for truth of *most* and *more than half* according to the semantics in (1).

(1) ⟦ most ⟧ = ⟦ more than half ⟧ $= \lambda A.\lambda B.|A \cap B| > \frac{1}{2} \cdot |A|$

(2) a. Most of As are Bs.
 b. More than half of As are Bs.

However, *most* and *more than half* differ in multiple respects. We overview here major differences between the two quantifiers discussed in the semantic literature.

2.1 Differences between *most* and *more than half*

Thresholds

There is empirical evidence that while the threshold for *more than half* is indeed 50%, the threshold for *most* is often higher than 50%.

The first type of evidence comes from corpus data: *most* is used in naturally occurring written text as if its threshold is higher than 50% (Solt, 2011, 2016). Solt finds that *most* in a sentence such as (2a) is used very rarely to describe situations in which 50-55% of As are Bs. Beyond 55%, the usage of *most* steadily increases to reach its peak at 80-85%.

The second type of evidence comes from reported judgments of oddness of sentences such as (2a) for %s greater — but not significantly greater — than 50%. Consider the following example from Solt (2011). Knowing that in 2008 50.7% of American population were female, (3a) is reported to be an infelicitous way to report that fact. On the other hand, (3b) is felicitous.

(3) a. ?Most Americans are female.
 b. More than half of Americans are female.

The third type of evidence comes from truth value judgments in an experimental setting: a sentence such as (2a) is sometimes judged false in situations in which more than 50%, but not significantly more than 50%, of As are Bs (Ariel, 2003, 2004; Ramotowska et al., 2020). To our knowledge, Ramotowska et al. (2020) study is the first to quantitatively evaluate the difference in thresholds between *most* and *more than half* from experimental data. They find that, while the average threshold of participants in their experiment for the quantifier *more than half* is indeed 50%, the average threshold for the quantifier *most* is 53%.

Vagueness

Solt argues that, even though the threshold for *most* is higher than 50%, the precise value for the threshold cannot be specified: "there is no value n such that speakers judge any proportion over $n\%$ to be most, while proportions less than or equal to $n\%$ are not most" (Solt, 2011, p.166). If this is correct, *most* is a vague quantifier, much like *tall* and *large* are vague predicates (Kennedy, 2007).

There are also experimental arguments in favor of the vagueness of *most*. The first comes from differences in verification times for quantifiers *most* and *more than half* reported in Ramotowska et al. (2020). Participants in their study were asked to evaluate the truth of sentences such as (2a) and (2b) against background information sentences of the form '$n\%$ of As are Bs', with $n \in [0, 100]$. They find that participants take longer to respond when n in the background sentence is closer to their individual threshold in the case of *most* than in the case of *more than half*. This effect may suggest that participants are more uncertain about what the threshold is in the case of *most* than in the case of *more than half*. We will see at a later point in the paper additional converging evidence for the vagueness of *most* coming from the differences in error pattern of *most* and *more than half*.

Verification

Verification strategies for quantifiers *most* and *more than half* have been studied in multiple experimental settings (Hackl, 2009; Szymanik and Zajenkowski, 2010; Lidz et al., 2011; Pietroski et al., 2009; Steinert-Threlkeld et al., 2015; Talmina et al., 2017). When it comes to contrasting the two quantifiers, Hackl argues that verifying a sentence such as (2a) involves tracking the difference between the As that are Bs and the As that are not Bs, and doesn't involve computing the total number of dots and dividing that number by two. Verifying a sentence such as (2b) on the other hand involves comparing the number of As that are Bs to half the number of As, and thus requires computing the total number of dots and dividing that number by two.

Domains

Finally, Solt (2011, 2016) observes that while *most* is felicitous with vague and uncountable domains, *more than half* is not — cf. the contrast in (4).

(4) Most/?? more than half of sadness diminishes over time.

2.2 Accounting for the differences

Interestingly, the differences between *most* and *more than half* discussed in Section 2.1 have not motivated the departure from the truth-conditional equivalence of (2a) and (2b) in the semantic literature. They are instead attributed to aspects of lexical semantics of the two quantifiers that do not affect their truth-conditional meaning (Hackl, 2009; Solt, 2011, 2016).

This is particularly striking for the case of differences in thresholds between *most* and *more than half*, which is the focus of our paper. In order to preserve the truth-conditional equivalence of *most* and *more than half* in light of the differences, their differences in thresholds are attributed to some form of pragmatic strengthening (Solt (2011, 2016); see also the discussion in Kotek et al. (2015)). We will refer to this general approach as Pragmatic Strengthening Hypothesis (PSH). This approach may come in (at least) two forms, which will be discussed in Section 3.

In this paper, we ask whether it is indeed plausible to maintain PSH, and by extension, the truth-conditional equivalence of the two quantifiers. We do this by deriving a general prediction of PSH, and demonstrating that it is not borne out in an experimental setting. Finally, we discuss the implications for the truth-conditional semantics of the two quantifiers.

3 Pragmatic strengthening hypothesis

We discuss two versions of PSH: the scalar implicature, and the R-implicature version.

3.1 Scalar implicature approach

One option is that the pragmatic strengthening that results in thresholds of *most* being higher than 50% is a scalar implicature (SI) (Grice, 1975, a.o.). To our knowledge, such an account has not been put forward, but could plausibly be formulated. For concreteness, we outline how the fact that the threshold of *most* is higher than the threshold of *more than half* could be accounted for as a higher-order SI (cf. Spector (2007) for higher-order SIs).

More than half may activate alternatives obtained by replacing 'half' with greater proportions (Solt, 2016). Many of these alternatives, such as *more than 51%, more than 54%...* would be filtered out from implicature computation due to granularity considerations (Cummins et al., 2012). Let us assume however that *more than half* has at least one alternative where 'half' is replaced by n, $0.5 < n < 1$ which is not filtered out due to granularity. For concreteness, let's take this alternative to be *more than 75%*. A sentence such as (2b) would have as an alternative the sentence in (5), and as its SI the negation of (5).

(5) More than 75% of As are Bs.

In other words, a sentence such as (2b) enriched by its SI would be interpreted as (6):

(6) More than 50% and less than 75% of As are Bs.

If a sentence such as (2a) were to compete with the pragmatically enriched meaning of (2b) in (6), and thus as a consequence had the implicature that (6) is false, (2a) would effectively be interpreted as (5). In other words, in this example, the threshold of *most* would be 75% as a consequence of a higher-order scalar implicature due to the competition with the pragmatically enriched meaning of *more than half*.

The sketch of a SI account described here is not intended as fully developed proposal: our goal is to merely demonstrate that such an account may be pursued as an explanation for differences in thresholds between *most* and *more than half*.

3.2 R-implicature approach

A proposal put forward by Solt (2016) is that the difference in thresholds between *most* and *more than half* is due to an R-implicature (RI). RI involves pragmatic strengthening to a stereotypical case (Horn, 1984).

Solt (2016) proposes that, while *most* and *more than half* are truth-conditionally equivalent, stereotypical situations in which a sentence such as (2a) would be used are those in which significantly more than 50% of As are Bs. The reason for this stems from Solt's proposal about the way quantities relevant for truth of sentences (2a) and (2b), such as $|A \cap B|$ and $|A - B|$ are measured and represented. Simplifying a lot, these quantities are by default measured and represented in a precise way in the case of *more than half*, and in a noisy way in the case of *most*. According to Solt's proposal, a sentence (2a) is verified by checking that $|A \cap B|$ is greater than $|A - B|$. As the representations of these quantities are noisy, (2a) would typically be used only when these quantities are sufficiently distinct, i.e. when $|A \cap B|$ is sufficiently greater than $|A - B|$.

Crucially, because of this, a sentence (2a) R-implicates that significantly more than half of As are Bs. A sentence such as (2b), on the other hand, does not have a comparable RI.

4 Moving forward

We propose to evaluate the plausibility of PSH by comparing how *most* and *more than half* are interpreted in *downward-entailing (DE) environments* as opposed to *upward-entailing (UE) environments*.

DE environments support entailments from supersets to subsets, ie. from a more general to a more specific term (and not vice-versa). An example is in (7).

(7) John didn't eat a chocolate cookie.
$\not\Rightarrow$ John didn't eat a cookie.
$\Rightarrow$ John didn't eat a dark chocolate cookie.

UE environments support entailments from subsets to supersets, ie. from a more specific to a more general term (and not vice-versa). An example is in (8).

(8) John ate a chocolate cookie.
$\Rightarrow$ John ate a cookie.
$\not\Rightarrow$ John ate a dark chocolate cookie.

Why would contrasting *most* and *more than half* in DE and UE environments be a relevant test for PSH? The reason is that, because of the entailment properties of these environments, increasing the threshold of *most* in UE environments *strengthens* the sentence, but doing so in DE environments *weakens* the sentence. To see this, note that (9b) (asymmetrically) entails (9a), but (10a) (asymmetrically) entails (10b).

(9) a. More than 50% of As are Bs.
b. More than 75% of As are Bs.

(10) a. It is not the case that more than 50% of As are Bs.
b. It is not the case that more than 75% of As are Bs.

This is true more generally: inferences which strengthen a sentence in an UE environment weaken it in a DE environment. This is why they are typically not derived when a sentence is embedded inside a DE environment, such as for instance the scope of negation. Let us examine the absence of relevant inferences in DE environments in greater detail.

4.1 SIs and DE environments

A sentence such as (11a) is commonly assumed to activate (11b) as alternative; note that (11b) asymmetrically entails (11a). Because of this, (11a) triggers as its SI the negation of (11b). In other words, (11a) enriched by its SI is interpreted as (11c) (Grice, 1975, a.o.).

(11) a. John ate a cookie or an apple.
 b. John ate a cookie and an apple.
 c. John ate a cookie or an apple, but he didn't eat both.

Observe now what happens when (11a) is embedded in a DE environment, as in (12). If the embedded sentence 'John ate a cookie or an apple' had the same implicature in (12) as in (11a), (12) would be true in two situations: (i) the situation in which John ate neither a cookie nor an apple; (ii) the situation in which he ate both a cookie and an apple. This interpretation is not naturally available[1]; (12) is naturally interpreted to be true in (i) but false in (ii).

(12) John didn't eat a cookie or an apple.

4.2 RIs and DE environments

While RIs are less studied and less well understood than SIs, Horn (1984) suggests that RIs are not derived under negation unless they have already been lexicalized, i.e. unless they had become part of the lexical meaning of an expression that used to trigger them as pragmatic strengthening implicatures. An example of a non-lexicalized RI discussed by Horn (1984) is the strengthening of ability modals. (13a) has as its RI (13b). In other words, (13a) enriched by its RI is interpreted as (13c).

(13) a. Mary was able to solve the problem.
 b. Mary solved the problem.
 c. Mary was able to solve the problem and solved it.

Observe now what happens when (13a) is embedded in a DE environment, as in (14). If the embedded sentence 'Mary was able to solve the problem.' had the same implicature in (14) as in (13a), (14) would be true in two situations: (i) the situation in which Mary wasn't able to solve the problem; (ii) the situation in which Mary was perfectly able to solve the problem, but didn't solve it (because for instance she decided not to). This interpretation is not naturally available; (14) is naturally interpreted to be true in (i) but false in (ii).

(14) Mary wasn't able to solve the problem.

4.3 General prediction of PSH

We have seen that inferences that lead to pragmatic strengthening in UE environments would lead to weakening in DE environments, and are thus in general not derived. The prediction for the differences in thresholds between *most* and *more than half*, if such difference was caused by pragmatic strengthening of *most*, is straightforward. *Most* should have a significantly higher threshold than *more than half* in UE environments, but the two quantifiers should have comparable thresh-

[1]This interpretation can be forced however with a pitch accent on 'or' (Cohen, 1971; Horn, 1989; Fox and Spector, 2018).

Table 1: **Predicted thresholds by PSH**

	More than half	**Most**
Upward monotonicity	50%	$50 + n\%$
Downward monotonicity	50%	50%

olds in DE environments. These predictions are summarized in Table 1. In what follows, we report two experiments which investigate whether the thresholds of the two quantifiers are affected by the monotonicity properties of their environment.

5 Experiment 1

Experiment 1 investigates thresholds of quantifiers *most* and *more than half* in an UE environment, as in (15a) and (16a) (unembedded *most* and *more than half*), and in a DE environment, as in (15b) and (16b) (*most* and *more than half* in the complement clause of *It is not the case that*).

(15) a. **Most** of the birlers are enciad.
b. It is not the case that **most** of the birlers are enciad.

(16) a. **More than half** of the birlers are enciad.
b. It is not the case that **more than half** of the birlers are enciad.

5.1 Task

Participants were told that they would see a sentence paired with background information and that their task was to decide whether the sentence was true with respect to the background information. The background information was presented in a box with gray background, and the test sentence was presented below the box in large font. An example of a test item is in Figure 1. Participants were instructed to record their responses by clicking on buttons labeled ‘False’ and ‘True’. They first saw two practice trials, one involving a true sentence, and one involving a false sentence; these practice trials were accompanied by suggested responses. The purpose of these examples was to familiarize the participants with the task. They then began the test phase of the experiment, the first two items of which were identical to the two practice trials.

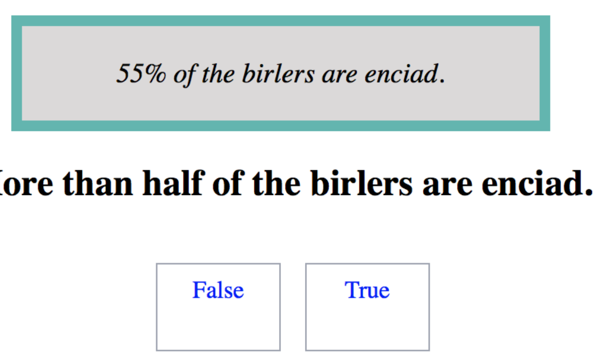


Figure 1: A test item from Experiment 1

5.2 Materials

Background information consisted of a single sentence which had the format in (17), with A a pseudo-noun with (regular) plural morphology, B a pseudo-adjective or a pseudo-noun with (regular) plural morphology , and $n \in [0, 100]$.[2]

(17) $n\%$ of the A are B.

Depending on the factor MONOTONICITY, test sentences had the format in (18a) (*upward monotonicity*), or (18b) (*downward monotonicity*), with A and B the same pseudowords as in the background information sentence. In the case of target items, Q reflects the level of the factor TARGET QUANTIFIER: Q was *most* or *more than half*. With respect to crossing MONOTONICITY and TARGET QUANTIFIER, the experiment employed a between-subject design: each participant was exposed to only one of the four combinations of MONOTONICITY$\times$TARGET QUANTIFIER. There were thus four experimental versions.

(18) a. Q of the A are B.
b. It is not the case that Q of the A are B.

One example target item for each combination of factors is in examples (19)-(22).

(19) MONOTONICITY: *upward*, TARGET QUANTIFIER: *most*
Most of the birlers are enciad.

(20) MONOTONICITY: *downward*, TARGET QUANTIFIER: *most*
It is not the case that most of the birlers are enciad.

(21) MONOTONICITY: *upward*, TARGET QUANTIFIER: *more than half*
More than half of the birlers are enciad.

(22) MONOTONICITY: *downward*, TARGET QUANTIFIER: *more than half*
It is not the case that more than half of the birlers are enciad.

Control items also had the format as in (18a) and (18b). They differed from target items in that Q reflects the level of CONTROL QUANTIFIER: Q was existential *some* or universal *all*. Control items were the same across four experimental versions.

One example control item for each combination of factors is in examples (23)-(26).

(23) MONOTONICITY: *upward*, CONTROL QUANTIFIER: *some*
Some of the birlers are enciad.

(24) MONOTONICITY: *downward*, CONTROL QUANTIFIER: *some*
It is not the case that some of the birlers are enciad.

(25) MONOTONICITY: *upward*, CONTROL QUANTIFIER: *all*
All of the birlers are enciad.

(26) MONOTONICITY: *downward*, CONTROL QUANTIFIER: *all*
It is not the case that all of the birlers are enciad.

[2]Experiment 1 design file with As, Bs, and ns for all experimental items, as well as scripts used for the analyses, can be consulted at https://semanticsarchive.net/Archive/TIxOTBiY/. Pseudowords used in Experiment 1 largely overlap with pseudowords used in the experiment reported in Ramotowska et al. (2020).

Each experimental version had 48 target items, and 96 control items (48 control items contained the CONTROL QUANTIFIER *some*, and the other 48 the CONTROL QUANTIFIER *all*; half of the occurrences of each control quantifier were in *upward*, and the other half in *downward* monotonic environment). Crucially, the truth evaluation of both target and control items depended on the % exhibited in the background sentence, which we will refer to as the test % in the context of Experiment 1. For the 48 target items, test %s were sampled without replacement and kept constant across four experimental versions. The same procedure was applied to select the test %s of 48 control items containing the CONTROL QUANTIFIER *some*, and the 48 control items containing the CONTROL QUANTIFIER *all*.

Note that when the quantifiers are in the complement clause of *It is not the case that*, they are in a finite clause which rules out the option of quantifier raising (Beghelli and Stowell, 1997; Cecchetto, 2004; Farkas, 1981). This ensured that the quantifier phrase headed by Q is interpreted in the scope of negation, which eliminates the possibility that these sentences have multiple readings due to scope ambiguities.

5.3 Participants and exclusions

200 participants were recruited at Prolific, and directed randomly to one of the four experimental versions. The experiment was a web-based binary truth value judgment task, hosted on Alex Drummond's Ibex platform for psycholinguistic experiments. One participant was excluded for failing to complete the task. Two participants were excluded for not being native speakers of English. Furthermore, we excluded 29 participants who made more than 20% of errors on what we will refer to as uncontroversial control items. These were all control items apart from (i) those in which the control quantifier *some* was in *upward monotone* environment and test % $>$ 50 and (ii) those in which the control quantifier *all* was in *downward monotone* environment and test % $<$ 50. The reason why (i) and (ii) are 'controversial' control items is that they may be perceived in our experimental context as having a stronger alternative, obtained by replacement of the control quantifier by *most/more than half*. This may lead to 'controversial' control items giving rise to scalar implicatures altering the cases in which these items are judged true or false. Furthermore, for each experimental version we excluded responses whose response time was more than 2 SD away from the average (response times were log-transformed). Finally, three participants were excluded because their estimated threshold wasn't in the range (0,100). 165 participants remained for the analysis, with comparable number of participants across four experimental versions (41, 37, 41, 46).

5.4 Participants' threshold estimation

To estimate participants' thresholds on target items, we use the same method as Ramotowska et al. (2020). For each participant, we fit the logistic function in (27):

(27)
$$P(test\ \%) = \frac{1}{1 + e^{\frac{(T - test\ \%)}{k}}}$$

$Test\ \%$ indicates the test % in a given trial, $P(test\ \%)$ indicates the probability that a participant answers 'true' for a given test %; T is the midpoint of the sigmoid which is interpreted as the participant's threshold, and k is the logistic growth rate.

To estimate T and k parameter for each participant, fitting was done using R *nls* self-starter

Table 2: **Experiment 1: Thresholds results**

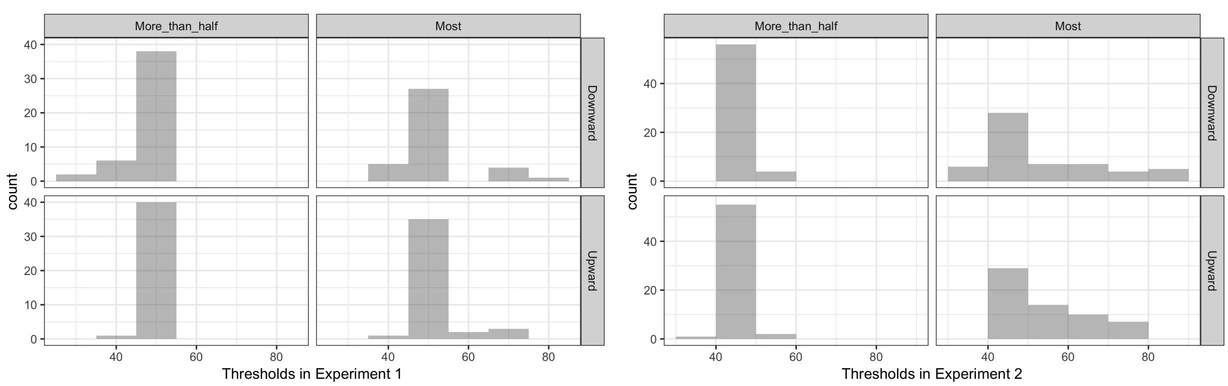
	More than half	**Most**
Upward monotonicity	$M = 48.4, SD = 2.1$	$M = 51.7, SD = 7.2$
Downward monotonicity	$M = 47.3, SD = 4.4$	$M = 51, SD = 8.9$

function: the starting value for T = 50; the starting value for $k = 4$ for the two experimental versions in which *most* and *more than half* were in UE environments, and $k = -4$ for the two experimental versions in which *most* and *more than half* were in DE environments.

For participants for whom this method could not converge to a threshold (80 participants)[3], the threshold was estimated as follows. For the two experimental versions in which *most* and *more than half* were in UE environments, for each participant, two test %s were determined: (1) the largest test % to which the participant answered 'False' and they answered 'False' with more than 80 % consistency for the smaller test %; and (2) the smallest test % to which the participant answered 'True' and they answered 'True' with more than 80 % consistency for the greater test %. The estimated participant's threshold was the average of these two test %. Parallel procedure was employed for the two experimental versions in which *most* and *more than half* were in DE environments, modulo the 'True'-'False' switch for determining the two relevant test %s.

5.5 Results

Because of the sampled test %, if the actual participants threshold is 50, the expected threshold in our experiment was 49.[4] Mean participants' thresholds and SDs for the four experimental versions are in Table 2. The distribution of thresholds from Experiment 1 can be visualized in Figure 2a.



(a) Experiment 1

(b) Experiment 2

Figure 2: Threshold distribution in Experiments 1 and 2 by Quantifier and Monotonicity.

Note that there is much more variability (as evidenced by their SDs) in thresholds for the quantifier *most* than for the quantifier *more than half*. This is compatible with the proposal that

[3]This was the case when percentages to which a participant responds 'True' and percentages to which a participant responds 'False' had very little or no overlap.

[4]This is because the largest test % below 50 was 46, and the smallest test % above 50 was 52; their mean is 49 (50 itself was not a test % in any of the target items trials).

most (but not *more than half*) is a vague quantifier (cf. also the discussion in Ramotowska et al. (2020)).

We turn now to the main question of interest: how do thresholds of *most* and *more than half* differ in *upward* and *downward* monotone environments?

Welch's unequal variances t-test confirmed that in *upward* monotone environments the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* ($t(46.7) = 2.8, p < .01$): this result replicates previous findings (Ariel, 2003, 2004; Solt, 2016; Ramotowska et al., 2020). The crucial novel finding is that the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* in *downward* monotone environments as well ($t(49.7) = 2.3, p = .03$).

Two-way ANOVA didn't reveal an interaction between Quantifier (Most vs. More than half) and Monotonicity (upward vs. downward) ($F(1) = 0.06, p = .8$). Because the homogeneity of variance assumption of classical two-way ANOVA was not satisfied in our data set, we also conducted Aligned-rank ANOVA, a non-parametric factorial analysis (Wobbrock et al., 2011), implemented in R in the ARTool package (Kay and Wobbrock, 2020). This analysis didn't reveal an interaction between Quantifier and Monotonicity either ($F(1) = 0.1, p = .7$).

6 Experiment 2

Experiment 2 extends the investigation to another pair of UE and DE environments, using a similar experimental setting as Experiment 1. The motivation for conducting Experiment 2 is to ensure that the results of Experiment 1 generalize to other DE environments, and are not an artifact of the specific DE environment tested in Experiment 1. In Experiment 2, the UE environment is the scope of the quantifier *each boy*, as in (28a) and (29a). The DE environment is the scope of the quantifier *no boy*, as in (28b) and (29b).

(28) a. <u>Each boy</u> puggled **most** of his entands.
b. <u>No boy</u> puggled **most** of his entands.

(29) a. <u>Each boy</u> puggled **more than half** of his entands.
b. <u>No boy</u> puggled **more than half** of his entands.

6.1 Task

The task was the same as in Experiment 1. An example of a test item is in Figure 3.

6.2 Materials

Background information consisted of two sentences: one provided information about a boy named Ted, and the other about a boy named Dan. The background information had the format in (30), with A a pseudo-verb with (regular) past tense morphology, B a pseudo-noun with (regular) plural morphology, and $n, m \in [0, 100]$.[5]

(30) Ted A n% of his B.
Dan A m% of his B.

Test sentences had the format in (31), with A and B the same pseudowords as in the background

[5]Experiment 2 design file with As, Bs, and ns for all experimental items, as well as scripts used for the analyses, can be consulted at https://semanticsarchive.net/Archive/TIxOTBiY/.

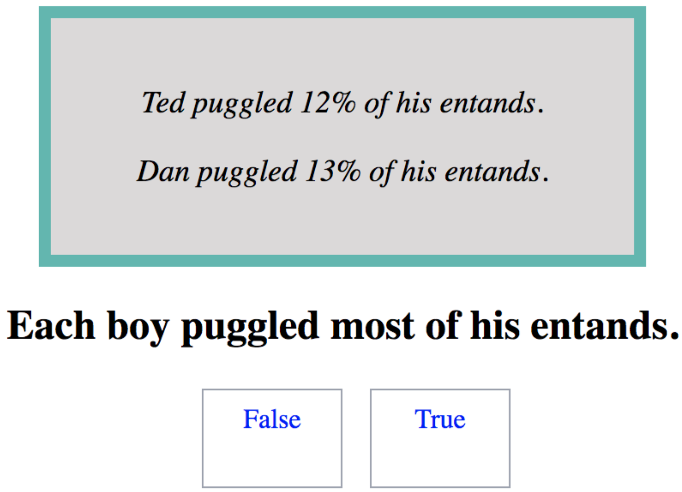


Figure 3: A test item from Experiment 2

information sentences. Q_1 reflects the level of the factor MONOTONICITY: Q_1 was *each* (*upward monotonicity*) or *no* (*downward monotonicity*). In the case of target items, Q_2 reflects the level of the factor TARGET QUANTIFIER: Q_2 was *most* or *more than half*. As in Experiment 1, with respect to crossing MONOTONICITY and TARGET QUANTIFIER, the experiment employed a between-subject design: each participant was exposed to only one of the four combinations of MONOTONICITY×TARGET QUANTIFIER. There were thus four experimental versions in Experiment 2 as well.

(31) Q_1 boy A Q_2 of his B.

One example target item for each combination of factors is in examples (32)-(35).

(32) MONOTONICITY: *upward*, TARGET QUANTIFIER: *most*
Each boy puggled most of his entands.

(33) MONOTONICITY: *downward*, TARGET QUANTIFIER: *most*
No boy puggled most of his entands.

(34) MONOTONICITY: *upward*, TARGET QUANTIFIER: *more than half*
Each boy puggled more than half of his entands.

(35) MONOTONICITY: *downward*, TARGET QUANTIFIER: *more than half*
No boy puggled more than half of his entands.

Control items also had the format as in (31). They differed from target items in that Q_2 reflects the level of CONTROL QUANTIFIER: Q_2 was existential *some/any* (*some* when MONOTONICITY was *upward*, *any* when MONOTONICITY was *downward*) or universal *all*. Control items were the same across four experimental versions.

One example control item for each combination of factors is in examples (36)-(39).

(36) MONOTONICITY: *upward*, CONTROL QUANTIFIER: *some*
Each boy puggled some of his entands.

(37) MONOTONICITY: *downward*, CONTROL QUANTIFIER: *any*
No boy puggled any of his entands.

(38) MONOTONICITY: *upward*, CONTROL QUANTIFIER: *all*
Each boy puggled all of his entands.

(39) MONOTONICITY: *downward*, CONTROL QUANTIFIER: *all*
No boy puggled all of his entands.

As in Experiment 1, each experimental version had 48 target items, and 96 control items (48 control items contained the CONTROL QUANTIFIER *some/any*, and the other 48 the CONTROL QUANTIFIER *all*; half of the occurrences of each control quantifier were in *upward*, and the other half in *downward* monotonic environment). Crucially, the truth evaluation of both target and control items depended on the two %s exhibited in the background sentences. Note that when MONOTONICITY is *upward*, the smaller of the two %s in the background information sentences is determining the truth of the test sentence; when MONOTONICITY is *downward*, the greater of the two %s is. In the context of Experiment 2, we will henceforth refer to the % which determines the truth of the test sentence as the test %, and to the % which does not as the non-test %. Test % of target and control items of Experiment 2 were the same as test % of target and control items of Experiment 1. In the cases of *upward* monotonicity, the non-test % was obtained by adding 1 or 2 to the test %; in the cases of *downward* monotonicity, the non-test % was obtained by subtracting 1 or 2 from the test %.

Note that the pronoun *his* in the quantifier phrase headed by Q_2 in target and control items was bound by Q_1. This ensured that the quantifier phrase headed by Q_2 is interpreted in the scope of the quantifier phrase headed by Q_1, which eliminates the possibility that these sentences have multiple readings due to scope ambiguities.

6.3 Participants and exclusions

280 participants were recruited at Prolific, and directed randomly to one of the four experimental versions. The experiment was a web-based binary truth value judgment task, hosted on Alex Drummond's Ibex platform for psycholinguistic experiments. Three participants were excluded for not being native speakers of English. Due to an experimenter's error, eight participants did two different versions of the experiment; their responses were excluded from the analysis. Furthermore, we excluded 23 participants who made more than 20% of errors on 'uncontroversial' control items, determined in the same way as in Experiment 1. Furthermore, for each experimental version we excluded responses whose response time was more than 2 SD away from the average (response times were log-transformed). Finally, two participants were excluded because their estimated threshold wasn't in the range (0,100). 235 participants remained for the analysis, with comparable number of participants across four experimental versions (60, 57, 58, 60).

6.4 Participants' threshold estimation

The procedure for participants' threshold estimation was the same as in Experiment 1.

6.5 Results

As in Experiment 1, because of the sampled test %s, if the actual participant's threshold is 50, the expected threshold in our experiment was 49. Mean participants' thresholds and SDs for the four experimental versions are in Table 3. The distribution of thresholds from Experiment 2 can be

Table 3: **Experiment 2: Thresholds results**

	More than half	**Most**
Upward monotonicity	$M = 48.7, SD = 2.3$	$M = 55.9, SD = 9.4$
Downward monotonicity	$M = 48.6, SD = 1.5$	$M = 54.8, SD = 14.3$

visualized in Figure 2b.

We again note that there is much more variability (as suggested by their SDs) in thresholds for the quantifier *most* than for the quantifier *more than half*.

We turn now to the main question of interest: how do thresholds of *most* and *more than half* differ in *upward* and *downward* monotone environments?

The results confirm the findings of Experiment 1. Welch's unequal variances t-test confirmed that in *upward* monotone environments the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* ($t(66.2) = 5.7, p < .01$). Crucially, again, the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* in *downward* monotone environments as well ($t(57.1) = 3.2, p < .01$).

Two-way ANOVA didn't reveal an interaction between Quantifier and Monotonicity ($F(1) = 0.2, p = .6$). However, aligned-rank ANOVA revealed a significant Quantifier-Monotonicity interaction ($F(1) = 7.4, p < .01$). We will come back to this interaction effect in the Discussion section.

7 Additional evidence for vagueness of *most*

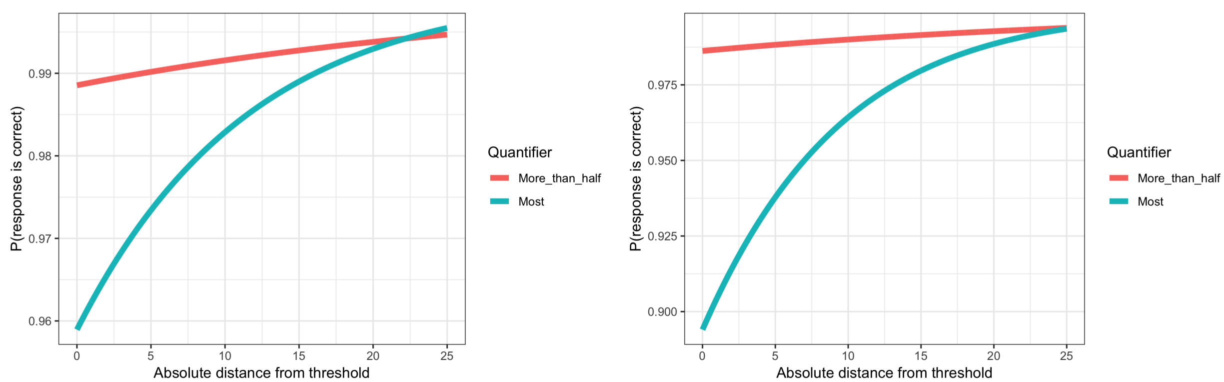
In the Introduction section, we have introduced arguments for the vague quantifier treatment of *most*. In this section, we provide an additional argument in favor of this hypothesis derived from our experimental results.

If *most* is a vague quantifier, it should be susceptible to borderline cases: this is in fact one of the defining properties of vagueness (Kennedy, 2007). A borderline case of an expression is a case for which a competent speaker has difficulties making a judgment whether the expression is true or false. Take for example the vague predicate *tall*: while most people would agree that someone whose height is 2m is tall, and someone whose height is 1m50 is not tall, it may be more difficult to judge whether someone whose height is 1m75 is tall or not tall.

If *most* but not *more than half* is susceptible to borderline cases, one may expect that there will be a greater region of uncertainty around each participant's threshold in the case of *most* than in the case of *more than half*. Uncertainty may lead to an inconsistent behavior for test %s close to the individual thresholds. We thus asked the following question: is participants' behaviour more inconsistent with their estimated thresholds when test % are close to the estimated threshold in the case of *most* than in the case of *more than half*?

To answer this question, we proceed as follows. We subset the data for each participant to the target trials whose test % are within +/-25 range of that participant's estimated threshold. For each target trial, we compute the *absolute distance* to the participant threshold. Each participant's response is then coded as correct or not correct based on their estimated threshold. For instance, if a participant's estimated threshold is 60%, and they happen to respond 'True' on a trial with the test % 52 in an *upward monotone* environment, this response would be coded as incorrect. For a participant whose estimated threshold is 50%, the same response would be coded as correct. A more precise formulation of our question is thus as follows: are their more incorrect responses closer to the threshold (in terms of absolute distance) in the case of *most* than in the case of *more than half*?

To assess this, a mixed effects logistic regression model was fitted on participants' responses (*correct* vs. *incorrect*) separately for Experiment 1 and 2, with Quantifier (*Most* vs. *More than half*), Absolute distance (cf. above), and their interaction as fixed effects and random by-participant intercepts. A comparison of this model with a reduced model without the interaction term revealed that Quantifier-Absolute distance interaction has a significant effect on response correctness in both Experiment 1 ($\chi^2(1) = 4.96, p = .03$) and Experiment 2 ($\chi^2(1) = 12.36, p < .01$), with more errors being made closer to the threshold in the case of *most* than in the case of *more than half*. This effect can be visualized in Figures 4a and 4b.



(a) Experiment 1 (b) Experiment 2

Figure 4: Probability that a response is correct on target trials as a function of quantifier (*most* vs. *more than half*) and of absolute distance of test %s in the target trials from participants' estimated thresholds.

This result provides an additional argument that the semantic representation of *most*, but not of *more than half*, is vague.

8 Discussion

When it comes to thresholds of quantifiers *most* and *more than half*, the empirical picture is the following. In UE environments, thresholds of *most* are on average higher than the thresholds of *more than half*, and there is a lot of individual variation in the thresholds of *most*, in line with the results of Ramotowska et al. (2020).

Crucially, in Experiment 1 and 2 we demonstrate that *most* and *more than half* retain these properties in DE environments. This result cannot be explained by the pragmatic strengthening hypothesis. Our experimental results thus speak in favor of there being a difference in truth-conditional meaning between *most* and *more than half*.

How should we then model the truth conditions for *most*, so that they reflect the differences in thresholds between it and *more than half*? Given the arguments in favor of vagueness of *most*, a salient possibility is to model the truth conditions of *most* on a par with other vague quantifiers which may have proportional interpretation, notably *many* and *few* (Kennedy, 2007; Solt, 2011).

Importantly, *most* comes with a restriction on what values the threshold may take: the threshold needs to be at least 50%.[6] Such a semantic entry for *most* is in (40), in which r is a contextually supplied threshold.

(40) $[\![\text{most}]\!] = \lambda A.\lambda B.\lambda r.|A \cap B| > r \cdot |A|, \text{where } 0.5 \leq r < 1$

There is however a number of points to discuss in assigning to *most* the truth-conditional semantics in (40). We examine them in turn.

8.1 Verification and mental representations

Hackl (2009), Pietroski et al. (2009) and Lidz et al. (2011) studied verification procedures of the quantifier *most*; Hackl (2009) further investigated verification procedures for *more than half*. These authors assume that *most* and *more than half* are truth-conditionally equivalent, and draw interesting consequences from specifics of verification procedures of *most*.

Hackl supports the truth-conditional equivalence assumption by the fact that *most* and *more than half* did not differ significantly in terms of accuracy and overall response times in his study (Hackl, 2009, p.89).[7] Pietroski et al. (2011) provide arguments that in the studies reported in Pietroski et al., 2009 and Lidz et al., 2011 *most* was interpreted as truth-conditionally equivalent to *more than half*.

The fact that *most* was truth-conditionally indistinguishable from *more than half* in Hackl, 2009, Pietroski et al., 2009 and Lidz et al., 2011 is clearly not incompatible with the semantic entry for *most* in (40): participants in these studies may have been prone to setting their individual threshold — r parameter in the entry in (40) — to $\approx 50\%$. It is of course an interesting question why the differences in truth-conditional behavior between *most* and *more than half* was not observed in Hackl, 2009, Pietroski et al., 2009 and Lidz et al., 2011, and they were in our study, as well as in the studies of Ariel, 2003, 2004; Talmina et al., 2017 and Ramotowska et al., 2020. One salient difference between these two sets of studies is that the former tested the interpretation of *most* and *more than half* against visual stimuli (e.g. images with colored dots), while the latter tested their interpretation against exact specified percentages. Whether it is indeed this difference in experimental setting or some other factor which influences the threshold of *most* that participants settle for and why remains to be explored.

Be that as it may, what consequences does the conclusion that *most* and *more than half* are not truth-conditionally equivalent after all have for the work of Hackl, 2009, Pietroski et al., 2009 and Lidz et al., 2011 in relation to mental representation and cognitive processing of *most*? At the very least, our results call for a great care when interpreting the results of these studies. More specifically, differences in behavioural verification patterns of *most* and *more than half*, such as variations in response time and acceptability, and/or tendency to involve approximate number system (ANS, cf. Dehaene (1997)), may be consequences of differences in truth conditions rather than of the format of mental representation of truth conditions. This cautionary remark holds even if participants settle for the threshold of $\approx 50\%$ throughout verification experiments, because (i) if participants more commonly assign the threshold to *most* that is greater than 50% outside of the

[6] It is not clear whether the truth conditions for *many* and *few* need to encode restrictions on what values the threshold may take.

[7] For the near replication and critical review of Hackl's study see also the work by Talmina et al. (2017). The results presented in this paper diverge in a few respects from the experiment of Hackl. In particular, contrary to Hackl, the authors have found significant overall differences in reaction times between *most* and *more than half*.

lab, they may nonetheless preserve some properties of their commonly used verification strategies when they decide to lower their threshold to $\approx 50\%$ inside the lab, and (ii) even with a threshold closer to 50% than usual, participants may still perceive *most* but not *more than half* as vague.

8.2 Domains

Recall that there is a contrast between *most* and *more than half* when their first argument is a vague or uncountable noun, such as 'sadness' (cf. (4), repeated below).

(4) Most/?? more than half of sadness diminishes over time.

To our knowledge, the only attempt to explain the contrast in (4) is put forward in Solt (2016). Her account however is not compatible with the semantics of *most* being that in (40). The reason is essentially the following. Solt (2016) assumes that (2a) is verified by checking that $|A \cap B|$ is greater than $|A - B|$, while (2b) is verified by checking that $|A \cap B|$ is greater than $\frac{1}{2} \cdot |A|$ (in line with the proposal by Hackl (2009)). In the terminology of measurement theory (Stevens et al., 1946), this presupposes that *more than half* involves measurement of quantities on a ratio scale, while *most* is compatible with measurement of quantities on the ordinal scale. Simplifying a lot, ordinal scales produces the order of measurements of quantities, but the distances between the measurements are not meaningful. Ratio scales are stronger than ordinal scales in that they produce not only the order of measurement of quantities, but they also make the distances and the ratios between measurements meaningful.

Now, Solt (2016) argues that, in the case of nouns such as 'sadness' in (4), (i) we are able to compute the quantity of sadness which, for instance, diminishes over time and quantity of sadness which doesn't, and compare these two quantities in terms of magnitude to establish which one is greater, but crucially (ii) we are not able to compute the quantity of half of sadness, or any proportion of sadness more generally. In other words, she argues that sadness can be measured on the ordinal, but not on the ratio scale. Because of this, using *more than half* in (4) is not an option, while using *most* is acceptable.

Crucially, if it is correct that sadness cannot be measured on a ratio scale and the semantics of *most* is that in (40), using *most* in (4) would not be an option either.

In what follows, we provide a novel empirical argument that challenges the idea that that sadness cannot be measured on a ratio scale.

The argument is based on the acceptability of the vague quantity expression *almost all* with nouns such as 'sadness': (41) is judged to be equally acceptable as its *most* counterpart.

(41) Almost all sadness diminishes with time.

Informally, the interpretation of (41) can be paraphrased as 'close to all, but not all, sadness diminishes with time' (cf. Horn (2002); Nouwen et al. (2006); Spector (2014); Crnič (2018) for various approaches to semantics of 'almost'). Arguably, 'close to all' is to be understood as 'some percentage less than all' rather than 'some number less than all'. To see why, compare (42a) and (42b).

(42) a. Almost all of my 100 friends came to my thesis defense.
b. Almost all of my 6 friends came to my thesis defense.

(42a) would plausibly be judged true in the situation where 95 out of my 100 friends (i.e. 5 less than all) came to my thesis defense. On the other hand, (42b) would plausibly be judged false in a situation where only 1 of my 6 friends (again, 5 less than all) came to my thesis defense. This shows that indeed what matters for the truth of *almost all A are B* is not what number less than all A are B, but what percentage less than all A are B.

In other words, the quantifier *almost all* has a proportional interpretation. Its semantics, leaving aside its compositional derivation, is plausibly the one in (43), in which r is a contextually supplied threshold.

(43) $[\![$ almost all $]\!] = \lambda A.\lambda B.\lambda r.|A \cap B| > r \cdot |A|$, where r is close to, but smaller than, 1

If the semantics of *almost all* is in (43), then in order to interpret (41), one must evaluate what proportion of sadness diminishes with time. Finally, if one is able to evaluate what proportion of sadness diminishes with time, the contrast in (4) is not an argument against the semantics of *most* in (40).

How to reconcile the empirical argument based on (41) with the contrast in (4)?

A direction to explore is thus that all three *most*, *almost all*, *more than half* have semantics which requires measurement on a ratio scale, but that there is another property behind the contrast in (40). Potential factors may be vagueness and generic uses of *most* and *almost all*.

Another direction would be to preserve Solt's claim that *more than half* requires measurement on a ratio scale, while *most* (and *almost all*) accept measurement on a weaker scale, but reject the claim that this scale can be as weak as an ordinal scale. One such candidate scale may be a *log-interval scale* (cf. Roberts (1979) for a discussion of log-interval scales). Interval scales are stronger than ordinal but weaker than ratio scales: they produce the order between measurement of quantities and make the distances between measurements meaningful, but the ratios between measurements are not meaningful. Log-interval scale corresponds to an exponential transformation of an interval scale. On a log-interval scale, a distance between two values on the scale thus corresponds to their ratio, rather than their difference: for instance, numbers 80 and 81 would be closer on a log-interval scale than numbers 1 and 2. An example of a log-interval scale is the Richter magnitude scale for the measurement of intensity of the earthquakes. If semantics of *most* and *almost all* were relying on log-interval scales, one may be able to simultaneously account for their proportional interpretation and for their difference with *more than half* in the case of (4), as the semantic entry of *most* and *almost all* would no longer need to make reference to the operation of division and multiplication. For instance, one could modify the proposed semantic entry for *most* as in (44), where μ is a measure function mapping entities to degrees on a log-interval scale, and r is a contextually supplied distance, measured in degrees, on the log-interval scale.

(44) $[\![$ most $]\!] = \lambda A.\lambda B.\lambda r.\mu(A \cap B) - r \geq \mu(A - B)$

Needless to say, developing a full proposal along any of these two lines requires further work.

8.3 Superlative interpretation of *most*

In this paper, we have focused on the *proportional* interpretation of *most* as in (45a). Proportional interpretation contrasts with the *superlative* interpretation of *most* as in (45b) (example (45) adapted from Kotek et al. (2015)).

(45) a. John talked to most of the students.
$\approx$ John talked to a proportion of students that surpasses certain threshold.

b. John talked to the most students.
$\approx$ John talked to more students that anybody else.

The two interpretations have been often treated as a case of lexical ambiguity. *Most* in (45a) is commonly treated as a lexical determiner (Barwise and Cooper, 1981), while *most* in (45b) as the superlative form of *many/much* (Heim, 1985; Hackl, 2009; Szabolcsi, 2012).

Whether the lexical ambiguity is the right approach to the two interpretations of *most* has been questioned. Most notably, Hackl (2009) proposes a derivation of *most* as in (45a) by treating it as the superlative form of *many/much* as well. The appeal of Hackl's proposal is two-fold. First, the two interpretations of *most* are accounted for in a unified manner, without positing lexical ambiguity. Second, Hackl demonstrates how this can account in a principled way that the contrary of *most* — **least* or **fewest*, with an interpretation *less than half* — is systematically absent from the category of quantifiers cross-linguistically: the explanation is, according to Hackl, that the superlative form of *few* would not result in the interpretation *less than half*, while the superlative form of *many* does result in the interpretation *more than half*. Crucially however, according to Hackl's proposal, *most* ends up truth-conditionally equivalent to *more than half*, which is at odds with our findings, and with the semantic entry proposed in (40).

We do not have an alternative proposal that would derive the proportional use of *most* from the superlative form of *many/much*, while encoding its vagueness and threshold properties. We leave the question of whether such a proposal should be put forward, or whether the proportional and superlative interpretations of *most* are ultimately a case of lexical ambiguity, open for future work.

8.4 Quantifier-Monotonicity interaction in Experiment 2

Our results established that thresholds of *most* are on average higher than thresholds of *more than half* in both UE and DE environments.

There is however another aspect of our findings that merits attention: the significant quantifier-monotonicity interaction in Experiment 2. Roughly put, this interaction suggests the following: the difference between thresholds of *most* and *more than half* is smaller in DE environments than in UE environments.

Given that this effect was not observed in Experiment 1, it is difficult to draw strong conclusions from it. To the extent that future work establishes that this effect is robust, however, we would like to point out to two potential explanations for it.

The first is that pragmatic strengthening of *most* may ultimately exist in addition to the truth-conditional meaning difference from *more than half*: the contribution of pragmatic strengthening to the threshold of *most* in UE environments is what is lost in DE environments, leading to the significant interaction described above.

There is a related alternative explanation: *most* may be subject to the so-called negative strengthening inferences. The two explanations 'co-vary', at least to some extent: the pragmatic strengthening hypothesis leads to increased thresholds of *most* in UE environments as compared to the baseline, while negative strengthening would lead to decreased thresholds of *most* in negative, and possibly DE environments more broadly, as compared to the baseline.

Let us examine more closely what negative strengthening inferences are. They are triggered by vague expressions when these appear in the scope of negation[8] (Horn, 1989; Krifka, 2007;

[8]To our knowledge, whether negative strengthening inferences are derived when vague expressions are embedded in DE environments other than the scope of negation is an open question.

Ruytenbeek et al., 2017; Gotzner et al., 2018). Consider for example (46a) and (46b):

(46) a. Not many people came.
 b. John isn't tall.

Due to negative strengthening inferences, (46a) may sometimes suggest that rather few people came, and (46b) that John is rather short. This interpretation is stronger than the complement of the interpretation of the positive versions of (46a) and (46b): in other words, the thresholds of *many* and *tall* are lower under negation than unembedded.

If *most* is indeed a vague expression, it may sometimes give rise to such negative strengthening inferences. These inferences would lead to lowering of the threshold of *most* in DE environments: if our participants were deriving them in Experiment 2, this can explain the significant interaction as well.

We stress again however that the robustness of this interaction effect is to be confirmed by future work: if it is confirmed, one may ask which of the two options outlined above — pragmatic strengthening in UE environments or negative strengthening in DE environments — characterizes *most*, in addition to the truth-conditional differences from *more than half*.

9 Conclusion

In this paper, we report the results of two experiments demonstrating that participants' thresholds of *most* are higher than thresholds of *more than half* in both UE and DE environments. This fact cannot be explained by pragmatic strengthening hypothesis, and suggests that these two quantifiers are not truth-conditionally equivalent.

In addition to being informative for the line of research studying mental representations and cognitive processing of *most* and *more than half*, our results also contribute to what we know about natural language quantifiers. We discuss the possibility to model the truth-conditional meaning of *most* on a par with other vague expressions whose semantics involves a threshold. In other words, contrary to common assumptions, the quantifying determiner *most*, and plausibly its cross-linguistic counterparts, may incorporate as their semantic ingredients a scale and a non-deterministic threshold, as well as constraints on what this threshold may be. We have also outlined some of the challenges of this approach, and pointed out what we believe to be interesting directions for future work.

This conclusion also raises a more general question: if such semantics is appropriate for *most*, should some of the other quantifying determiners be (re-)analyzed using similar semantic ingredients? We leave this question open for future work.

References

Ariel, M. (2003). Does most mean 'more than half'? In *Annual Meeting of the Berkeley Linguistics Society*, Volume 29, pp. 17–30.

Ariel, M. (2004). Most. *Language 80*(4), 658–706.

Barwise, J. and R. Cooper (1981). Generalized quantifiers and natural language. In *Philosophy, Language, and Artificial Intelligence*, pp. 241–301. Springer.

Beghelli, F. and T. Stowell (1997). Distributivity and negation: The syntax of each and every. In *Ways of Scope Taking*, pp. 71–107. Springer.

Braine, M. and D. P. O'Brien (1998). *Mental logic*. Psychology Press.

Cecchetto, C. (2004). Explaining the locality conditions of qr: Consequences for the theory of phases. *Natural Language Semantics 12*(4), 345–397.

Chater, N. and M. Oaksford (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology 38*(2), 191–258.

Cohen, L. J. (1971). Some remarks on grice's views about the logical particles of natural language. In *Pragmatics of Natural Languages*, pp. 50–68. Springer.

Crnič, L. (2018). A note on connected exceptives and approximatives. *Journal of Semantics 35*(4), 741–756.

Cummins, C., U. Sauerland, and S. Solt (2012). Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy 35*(2), 135–169.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Farkas, D. (1981). Quantifier scope and syntactic islands. In *Papers from the Seventeenth Regional Meeting of Chicago Linguistics Society*, pp. 59–66.

Fox, D. and B. Spector (2018). Economy and embedded exhaustification. *Natural Language Semantics 26*(1), 1–50.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition 86*(3), 223–251.

Geurts, B. and F. van Der Slik (2005). Monotonicity and processing load. *Journal of Semantics 22*(1), 97–117.

Gotzner, N., S. Solt, and A. Benz (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology 9*, 1659.

Grice, H. P. (1975). Logic and conversation. In *Speech Acts*, pp. 41–58. Brill.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics 17*(1), 63–98.

Heim, I. (1985). Notes on comparatives and related matters. *Ms., University of Texas, Austin*.

Higginbotham, J. (1994). Mass and count quantifiers. *Linguistics and Philosophy 17*(5), 447–480.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, Form, and Use in Context: Linguistic Applications 11*, 42.

Horn, L. (1989). A natural history of negation.

Horn, L. R. (2002). Assertoric inertia and npi licensing. In *Chicago Linguistic Society*, Volume 38, pp. 55–82.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

Johnson-Laird, P. N. and B. G. Bara (1984). Syllogistic inference. *Cognition 16*(1), 1–61.

Kay, M. and J. O. Wobbrock (2020). Package 'artool'.

Keenan, E. L. and J. Stavi (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy 9*(3), 253–326.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy 30*(1), 1–45.

Kotek, H., Y. Sudo, and M. Hackl (2015). Experimental investigations of ambiguity: the case of most. *Natural Language Semantics 23*(2), 119–156.

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. In *Presupposition and Implicature in Compositional Semantics*, pp. 163–177. Springer.

Lidz, J., P. Pietroski, J. Halberda, and T. Hunter (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics 19*(3), 227–256.

Lindström, P. (1966). First order predicate logic with generalized quantifiers. *Theoria 32*, 186–195.

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae 44*, 12–36.

Nouwen, R. et al. (2006). Remarks on the polar orientation of almost. *Linguistics in the Netherlands 23*(1), 162–173.

Pietroski, P., J. Lidz, T. Hunter, and J. Halberda (2009). The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language 24*(5), 554–585.

Pietroski, P., J. Lidz, T. Hunter, D. Odic, and J. Halberda (2011). Seeing what you mean, mostly. In *Experiments at the Interfaces*, pp. 181–217. Brill.

Ramotowska, S., S. Steinert-Threlkeld, L. van Maanen, and J. Szymanik (2020). Individual differences in semantic representations affect quantifier processing. In *Proceedings of the 24th Sinn und Bedeutung*.

Roberts, F. S. (1979). *Measurement Theory with Applications to Decision-making, Utility, and the Social Sciences*. Reading: Addison-Wesley.

Ruytenbeek, N., S. Verheyen, and B. Spector (2017). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: A Journal of General Linguistics*.

Solt, S. (2011). Vagueness in quantity: Two case studies from a linguistic perspective. *Understanding Vagueness. Logical, Philosophical and Linguistic Perspectives*, 157–174.

Solt, S. (2016). On measurement and quantification: The case of most and more than half. *Language 92*(1), 65–100.

Spector, B. (2007). Aspects of the pragmatics of plural morphology: On higher-order implicatures. In *Presupposition and implicature in compositional semantics*, pp. 243–281. Springer.

Spector, B. (2014). Global positive polarity items and obligatory exhaustivity. *Semantics and Pragmatics 7*, 11–1.

Steinert-Threlkeld, S., G.-J. Munneke, and J. Szymanik (2015). Alternative representations in formal semantics: A case study of quantifiers. In N. T. T. Brochhagen, F. Roelofsen (Ed.), *Proceedings of the 20th Amsterdam Colloquium*, pp. 368–377.

Stevens, S. S. et al. (1946). On the theory of scales of measurement. *Science 103(2684)*, 677–680.

Szabolcsi, A. (2012). Compositionality without word boundaries:(the) more and (the) most. In *Semantics and Linguistic Theory*, Volume 22, pp. 1–25.

Szymanik, J. (2016). *Quantifiers and Cognition. Logical and Computational Perspectives*. Studies in Linguistics and Philosophy. Springer.

Szymanik, J. and M. Zajenkowski (2010). Comprehension of simple quantifiers: empirical evaluation of a computational model. *Cognitive Science 34*(3), 521–532.

Talmina, N., A. Kochari, J. Szymanik, et al. (2017). Quantifiers and verification strategies: connecting the dots. In *21 st Amsterdam Colloquium*, pp. 465.

Van Benthem, J. (1986). *Essays in logical semantics*. Dordrecht: Reidel.

Westerståhl, D. (1985). Logical constants in quantifier languages. *Linguistics and Philosophy 8*(4), 387–413.

Wobbrock, J. O., L. Findlater, D. Gergle, and J. J. Higgins (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 143–146.