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

Milica Denić and Jakub Szymanik

University of Amsterdam

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 pragmatic effect, 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 negative environments, which we argue to speak 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; Lidz et al., 2011; Steinert-Threlkeld et al., 2015; Talmina et al., 2017).

*Many thanks to Fausto Carcassi, Alexandre Cremers, Martin Hackl, Manuel Križ, Leendert van Maanen, Mora Maldonado, Sonia Ramotowska, Stephanie Solt and Benjamin Spector 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.

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 representation of these truth conditions.

In this work, we challenge the assumption that *most* and *more than half* are truth-conditionally equivalent. This is done by studying how the two quantifiers are interpreted in negative environments in an experimental setting. This challenge calls for 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 *semantic* threshold for truth of *most* and *more than half* according to the semantics in (1).

(1) $[\![\text{ most }]\!] = [\![\text{ more than half }]\!] = \lambda A.\lambda B.|A \cap B| > |A - B| = \lambda A.\lambda B.|A \cap B| > \frac{1}{2} \cdot |A|$

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

While the semantic thresholds of *most* and *more than half* are the same (i.e. 50%) according to the semantics in (1), the empirically *observed* thresholds of the two quantifiers differ. More specifically, there is empirical evidence that while the observed threshold for *more than half* is indeed 50%, the observed threshold for *most* is often higher than 50%: in other words, *most* is often interpreted as 'significantly more than half'.

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 contexts in which 50-55% of As are Bs. Beyond 55%, the usage of *most* steadily increases to reach its peak in contexts in which 80-85% of As are Bs.

The second 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%.

3 Accounting for the differences

Where do the differences in observed thresholds between *most* and *more than half* come from? The first possibility is that the semantics in (1) is not correct for *most*, and that the difference in thresholds between *most* and *more than half* reflects their difference at the level of truth-conditional semantics. This option, as discussed in detail in Section 4, is pursued by Solt (2011, 2016).

Much other work on *most* and *more than half* however assumes that the semantic threshold of both *most* and *more than half* is 50% (Ariel, 2004; Pietroski et al., 2009; Horn, 2006; Hackl, 2009; Lidz et al., 2011; Kotek et al., 2015, a.o.). In general, this work attributes the differences in observed thresholds between *most* and *more than half* to pragmatic factors. Strikingly, however, most of this work doesn't develop a proposal for what these pragmatic factors are. For instance, Ariel (2004) proposes that the observed threshold of *most* is higher than 50% due to a conversational implicature of noteworthiness, whereby (2a) pragmatically implicates that a significant amount of As are Bs. She however offers no formal account of this implicature. Kotek et al. (2015) suggest that, while "*most* is true in the same cases as *more than half*", the two quantifiers differ in the pragmatics of their usage resulting in differences in observed thresholds, without making more specific claims about the pragmatic mechanisms involved. Similarly, Pietroski et al. (2009) (footnote 1) suggest that the differences in observed thresholds should be treated as a pragmatic effect, again, without making more specific claims about the pragmatic mechanisms involved. The idea that the differences in observed thresholds between *most* and *more than half* are a pragmatic effect is thus left under-specified in the work by Ariel (2004); Pietroski et al. (2009); Kotek et al. (2015) and is as such difficult to evaluate empirically. A notable exception in this respect is Horn (2006), who offers a more concrete proposal about the pragmatic mechanism behind the differences in observed thresholds between *most* and *more than half*, which will be discussed in Section 5.1.

The central goal of this paper is to make progress on the question of where the differences in observed thresholds between *most* and *more than half* come from. To this end, the two aforementioned approaches will be contrasted: the approach according to which the differences in thresholds reflect the differences in truth-conditional semantics between *most* and *more than half* (Solt, 2011, 2016), and the approach according to which the differences in thresholds are a pragmatic effect (Horn, 2006). In addition to Horn's 2006 pragmatic approach, we develop two novel alternative proposals for what the pragmatic mechanism behind the differences in observed thresholds may be. These proposals are then evaluated experimentally. It will be seen that these three pragmatic proposals (Horn's + two novel ones) — and, more generally, proposals invoking a pragmatic strengthening mechanism — contrast with approaches postulating the difference in truth-conditional semantics between *most* and *more than half* in terms of their predictions for how the two quantifiers are interpreted in negative environments. Two experiments will be presented which test these predictions: their results favor accounts according to which *most* and *more than half* differ in their truth-conditional semantics.

A secondary goal of the paper is to explore, once the results of the two experiments are in place, yet another difference between *most* and *more than half*: vagueness (Solt, 2011). The results of the two experiments will be used to support Solt's proposal that *most* is vague, while *more than half* is not. Finally, the connection between differences in thresholds and differences in vagueness between the two quantifiers will be explored.

4 Semantic account: Solt (2011, 2016)

Let us start by presenting the account by Solt (2011, 2016), which grounds the differences in observed thresholds between *most* and *more than half* in the differences in truth-conditional semantics between the two quantifiers. By this it is meant that, according to her account, (2a) may be semantically false in certain cases when more than 50% of As are Bs. On the other hand, (2b) is never semantically false in such cases.

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

Solt (2011, 2016) adopts a degree-based semantic framework, in which the ontology is extended to include degrees as a basic type (type d) (Bartsch and Venneman, 1973; Cresswell, 1976; Bierwisch, 1989; Kennedy, 1999; Hackl, 2000; Heim, 2000, a.o.). Degrees are organized into measurement scales, which Solt (2011, 2016) argues to be central for understanding the differences in thresholds between *most* and *more than half*. Formally, a scale of measurement S is a structure which minimally involves a set of degrees D_S, an ordering relation $\succ_S$ on D_S, and a dimension of measurement DIM_S (e.g. weight, volume, number of elements of a set). Building on the proposal by Hackl (2009), according to which the semantic entry of *most* encodes a comparison between $A \cap B$ and $A - B$, while that of *more than half As are Bs* encodes a comparison between $A \cap B$ and $\frac{1}{2} \cdot A$, with A and B the first and the second argument of the quantifiers, Solt proposes that the semantics of *most* and *more than half* is in (3) and (4), respectively. According to (3) and (4), the meaning of *most* and *more than half* is evaluated with respect to a contextually supplied measurement scale S, with μ_S being a measure function mapping entities to degrees on S, and $\succ_S$ the ordering relation of S.

(3) $[\![\text{ most }]\!]^S = \lambda A.\lambda B.\mu_S(A \cap B) \succ_S \mu_S(A - B)$

(4) $[\![\text{ more than half }]\!]^S = \lambda A.\lambda B.\mu_S(A \cap B) \succ_S \frac{1}{2} \cdot \mu_S(A)$

According to the semantics in (3), (2a) true iff the degree $\mu_S(A \cap B)$ succeeds the degree $\mu_S(A-B)$ on the scale S. According to the semantics in (4), (2b) is true iff the degree $\mu_S(A \cap B)$ succeeds the degree $\frac{1}{2} \cdot \mu_S(A)$ on the scale S. It is crucial to note that, according to (3), the truth or falsity of (2a) is not determined by the objective 'greater than' relation of the objective measures of As that are Bs, and of As that are not Bs, but rather by the ordering relation $\succ_S$ of the degrees on the scale S which the As that are Bs and the As that are not Bs are mapped to by μ_S (similarly for (2b)). Importantly, the ordering relation $\succ_S$ may not correspond perfectly to the objective 'greater than' relation, which is, according to Solt (2011, 2016), the source of the differences in thresholds between *most* and *more than half*, as explained below.

More specifically, Solt (2016) argues that *most* and *more than half* differ in the measurement scales they accept. In the terminology of measurement theory (Stevens et al., 1946), evaluating sentences with *most* is compatible with the measurement of quantities on a semiordered scale, but evaluating sentences with *more than half* is not. A semiordered scale S is a scale whose ordering relation $\succ_S$ has properties of a semiorder. Simplifying somewhat, this means that the $\succ_S$ relation incorporates a **discriminability threshold**, defined informally in (5): two stimuli whose objective measures (e.g. two sets whose objective cardinalities) fall within the relevant threshold are not discriminable when these stimuli are mapped to degrees on the scale S. In other words, on a semiordered scale, two different degrees d, d' may be, but are not necessarily, discriminable: in the former case, they stand in $d \succ_S d'$ or in $d' \succ_S d$ relation; in the latter case, they stand in $d \sim_S d'$ relation.

(5) *Discriminability threshold* of a semiordered scale S is the minimal objective measurement difference of two stimuli being compared necessary for their discriminability on S; in other words, the two stimuli are discriminable when measured on the scale S iff their objective measurement difference is at least as large as the discriminability threshold.

For example, take two sets A and B, with objective cardinalities 22 and 21 respectively. The measure function μ_S maps these sets to degrees $\mu_S(A)$ and $\mu_S(B)$ on a semiordered scale S. Whether or not it is the case that $\mu_S(A) \succ_S \mu_S(B)$ will depend on the discriminability threshold of S. If the objective difference of 1 is sufficient for discriminability on S, $\mu_S(A) \succ_S$

$\mu_S(B)$ on that scale. On the other hand, if the objective difference larger than 1 is needed for discriminability on S, $\mu_S(A) \sim_S \mu_S(B)$ on that scale.

The reason why evaluating sentences with *most* but not those with *more than half* is compatible with measurement of quantities on a semiordered scale is because a semiordered scale supports the comparison of degrees, which is by assumption sufficient for evaluating sentences with *most*, but it doesn't support the operations of multiplication and division of degrees, which is by assumption necessary for evaluating sentences with *more than half*.

Solt (2016) provides a series of arguments supporting the claim that *most*, but not *more than half*, is compatible with a semiordered scale of measurement. One argument comes from the connection made between *most* and approximate number system (ANS), whereby online processing of sentences with *most* involves ANS representations and comparisons of quantities (Pietroski et al., 2009; Lidz et al., 2011). ANS is a cognitive system which supports estimatation and represention of quantities in an imprecise manner (Cantlon et al., 2009; Dehaene, 1997). This system is however known to influence the processing of exact cardinalities represented as number symbols as well (Moyer and Landauer, 1967; Dehaene, 1997). The ANS representations of quantities are imprecise in the sense that the representations of different quantities may overlap. In fact, the larger the quantity, the less precise its representation (size effect) — for example, the ANS representation of the quantity 20 is less precise than that of the quantity 10. Furthermore, the larger the distance between the two quantities, the better their discriminability (distance effect) — for instance, ANS representations of 20 and 10 are more discriminable than those of 20 and 15. Importantly, Solt (2016) draws a connection between the distance effect of ANS and the discriminability threshold of semiordered scales.[1]

Solt (2016) proposes that the compatibility of *most*, but not of *more than half*, with a semiordered scale of measurement is the key to understanding the differences in thresholds between the two quantifiers. More specifically, when $\mu_S(A \cap B)$ and $\mu_S(A - B)$ are represented and compared on a semiordered scale S, in order for it to be the case that $\mu_S(A \cap B) \succ_S \mu_S(A - B)$, degrees $\mu_S(A \cap B)$ and $\mu_S(A - B)$ need to be sufficiently discriminable. This can only be the case if the objective quantity of As that are Bs is sufficiently greater than the objective quantity of As that are not Bs: how much is *sufficient* depends on the discriminability threshold of S.

This essentially means that the sentence (2a) when evaluated with respect to a semiordered scale S will be ***semantically false*** in situations in which the quantity of As that Bs is objectively greater than the quantity of As that are not B, but $\mu_S(A \cap B)$ and $\mu_S(A - B)$ are not discriminable on S. Because of this, the semantic threshold of *most* is not always 50%: the value of the semantic threshold of *most* depends on the discriminability threshold of the scale with respect to which the sentence with *most* is evaluated. For instance, in the case of comparison of set cardinalities, if the objective difference of 1 is sufficient for discriminability on a semiordered scale S, the semantic threshold of *most* in (2a) will be 50% when the sentence is evaluated with respect to S. However, if the objective difference of 1 is not sufficient for discriminability on a semiordered scale S (i.e. if a larger difference is needed for the two quantities to be discriminable), the semantic threshold of *most* in (2a) will be ***higher*** than 50% when the sentence is evaluated with respect to S.

Solt (2016) goes on to propose that not only do sentences with *most* allow evaluation with respect to a semiordered scale, but they in fact may default to it via a pragmatic mechanism

[1]It should be noted however that there are empirical arguments that semiordered scales are too simplistic as a model of ANS: for instance, semiordered scales of measurement by definition do not support operations of multiplication and division, while ANS representations do (cf. Barth et al. (2009); Jacob et al. (2012); Matthews and Lewis (2016); McCrink and Spelke (2010, 2016); McCrink et al. (2017); Qu et al. (2021), a.o.).

even when a more precise scale of measurement is contextually available, such as for example when the relevant quantity information comes from sentences with digits (e.g. *'55% of As are Bs'*).[2] This is necessary to make her account compatible with the corpus data collected in her study, as well as with the data from the study of Ramotowska et al. (2020), both of which involve sentences where relevant quantity information is represented with digits. However, the pragmatic mechanism by which sentences with *most* default to an evaluation with respect to a semiordered scale is assumed to be optional; because of this, they can sometimes also be evaluated with respect to a more precise measurement scale S whose ordering relation $\succ_S$ precisely tracks the objective 'greater than' relation. In such cases, the semantic threshold of *most* would be 50%.

Importantly, sentences with *more than half* such as (2b) are never evaluated with respect to a semiordered scale. By assumption, they are necessarily evaluated with respect to a more precise measurement scale S whose ordering relation $\succ_S$ precisely tracks the objective 'greater than' relation. This is why the semantic threshold of *more than half* is always 50%.

Crucially, then, according to this proposal, *most* and *more than half* are not truth-conditionally equivalent. The truth conditions of a sentence with *most* vary with the contextually supplied measurement scale with respect to which the sentence is evaluated. When the domains of the two quantifiers are countable, an equivalent formulation of the semantics of *most* and *more than half* to that in (3) and (4) is in (6) and (7), respectively. For simplicity, we will limit ourselves to countable domains and adopt the latter formulation henceforth as it makes the differences in truth-conditional semantics between the two quantifiers more transparent. In (6), T varies with the contextually supplied measurement scale: if such a scale is semiordered, $T \geq 1$ and its value corresponds to the scale's discriminability threshold; on the other hand, if such a scale's ordering relation precisely tracks the objective 'greater than' relation, $T = 1$.

(6) $[\![\text{most}]\!]^T = \lambda A.\lambda B.\ |A \cap B| \geq |A - B| + T$ (with $T \geq 1$)

(7) $[\![\text{more than half}]\!] = \lambda A.\lambda B\ |A \cap B| > \frac{1}{2} \cdot |A|$

To summarize, according to the proposal in Solt (2011, 2016), sentences with *most* and *more than half* are interpreted with respect to contextually supplied measurement scales; the semantic threshold of the two quantifiers depends on these measurement scales. The quantifier *most* can be (and sometimes has to be) evaluated with respect to a semiordered scale; the quantifier *more than half* on the other hand requires a more precise measurement scale. Because of this, the semantic threshold of *most* may be higher than 50%, but that of *more than half* is always 50%.

5 Pragmatic accounts

In the previous section, the account by Solt (2011, 2016) was described, according to which the differences in thresholds between *most* and *more than half* are grounded in truth-conditional semantics. On the other hand, many authors assume that the two quantifiers are truth-conditionally equivalent, and suggest that differences in *observed* thresholds are due to pragmatics (cf. Section 3).

Let us start by discussing Horn's 2006 proposal for the pragmatic mechanism behind the differences in thresholds between *most* and *more than half*.[3]

[2]Solt suggests that the pragmatic mechanism behind the evaluation of *most* with respect to a semiordered scale even when a more precise scale of measurement is contextually available may be a (conventionalized) R-implicature, cf. Horn (1984).

[3]Note however that the proposal in question is not the main goal of Horn (2006), and is discussed very briefly in his paper, cf. Section 5.1.

We will then develop accounts of two alternative pragmatic mechanisms which may result in *observed* threshold of *most* being higher than 50%, even though its semantic threshold is 50%.

The first account will be a modified version of the account by Solt (2011, 2016); one in which measurement scales play a role not at the level of truth-conditional semantics, but at the (post-semantic) level of deciding whether to utter a sentence in a context and of determining the truth-value of a sentence in a context.

The second account assumes that *most* and *more than half* are truth-conditionally equivalent, but that *most* is subject to scalar implicatures leading to pragmatic strengthening. We also note that in parallel to the present work, another version of the scalar implicature proposal for the differences in the observed thresholds between *most* and *more than half* has been put forward by Carcassi and Szymanik (2021).

These accounts will now be introduced in turn.

5.1 Horn (2006): markedness considerations

While Horn (2006) mainly focuses on different accounts of the implication of (2) towards (8), he also briefly discusses differences in thresholds between *most* and *more than half.*

(2a) Most of As are Bs.

(8) Not all of As are Bs.

According to Horn (2006) (Section 4), "a speaker who goes out of her way to say *more than half*, eschewing the less marked and briefer *most*, must have a reason for so doing; marked forms are used in marked situations. One motive for using *more than half* may be precisely its compositional structure, which focuses on whether the proportion in question is less than, equal to, or greater than half of the set under consideration. This is likely to arise in situations involving ***near majorities or bare majorities***."

In other words, Horn (2006) assumes that *most* and *more than half* are truth-conditionally equivalent (ie. their semantic threshold is 50%), but suggests that *most* is less marked than *more than half.* Because of this, the use of *'more than half of As are Bs'* is only felicitous in marked situations, while *'most of As are Bs'* wouldn't be used in such marked situations. An example of such a marked situation is that involving bare majorities, such as, for example, the situation where 51% of As are B. This would result in differences in observed thresholds between *most* and *more than half*, whereby the observed threshold of *most* is higher than 50%.

Horn (2006) evokes his principle of division of pragmatic labor to account for this specialization of *most* to unmarked situations and specialization of *more than half* to marked situations (cf. Horn (1984)). According to this principle, unmarked forms specialize for unmarked situation via what he calls an R-implicature. According to Horn's account, then, the R-implicature drives the differences in thresholds between *most* and *more than half.* Subsequently, marked forms specialize for marked situations via a Q- (or scalar) implicature (cf. Horn (1984) for details).

5.2 ANS and confidence requirement

According to the account developed in this section, *most* and *more than half* are truth-conditionally equivalent. That is, both (2a) and (2b), repeated below, are true iff $|A \cap B| > |A - B|$ iff $|A \cap B| > \frac{1}{2} \cdot |A|$.

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

How do interlocutors decide whether sentences in (2) are true? Let us grant that they do so by comparing relevant quantities, such as $|A \cap B|$ and $|A - B|$ in the case of (2a), and $|A \cap B|$ and $\frac{1}{2} \cdot |A|$ in the case of (2b) (cf. Hackl (2009)). Let us further assume that such quantities in the case of *most* (but not *more than half*) by default involve ANS representations, even when more precise numeric representations are available: in other words, when (2a) is interpreted, $|A \cap B|$ and $|A - B|$ are by default represented and compared on the ANS scale of measurement, commonly referred to as the ANS mental number line (cf. also Pietroski et al. (2009); Lidz et al. (2011)). This aspect of the proposal is parallel to the aspect of Solt's proposal according to which sentences with *most* may by default be evaluated with respect to a semiordered measurement scale (cf. Section 4).

Under the present approach, contrary to Solt (2011, 2016), a measurement scale does not enter into the truth-conditional semantics of *most*, but rather plays a role at the post-semantic level: the representation of quantities on the ANS mental number line is incorporated into speakers' decision of whether to utter a sentence with *most*, and into listeners' decision of whether a sentence with *most* is true. The key component of the present account is that speakers only produce sentences which they are confident to be true (cf. Gricean maxim of quality, (Grice, 1975)), and similarly, that listeners evaluate that a sentence is true of a given situation only when they can do so with confidence. This will be referred to in the continuation as the *confidence requirement*.

What does the confidence requirement entail for statements such as (9) about quantities A and B when quantity comparison is performed on ANS representations? Recall that ANS supports estimation and representation of quantities in an imprecise manner (Cantlon et al., 2009; Dehaene, 1997). In other words, different quantities represented on the ANS mental number line may overlap to a different extent — the larger their overlap, the less discriminable they are. When two ANS representations of quantities are not discriminable, it is not possible to establish with confidence whether any of the statements in (9) is true. Because of this, when two quantities A and B are not discriminable on the ANS mental number line, the confidence requirement entails (10).

(9) A is greater than B; A is smaller than B; A is equal to B

(10) *Confidence requirement and ANS:* For two quantities A, B represented and compared on the ANS mental number line, when A and B are not discriminable, statements in (9) are not uttered by speakers and they are evaluated to be false by listeners.

How does (10) accounts for the fact that the observed threshold of *most* is higher than 50%?

Given the confidence requirement and its interaction with ANS spelled out in (10), in order for a speaker to utter (2a), and for the listener to evaluate it to be true, it has to be the case that the objective difference between $|A \cap B|$ and $|A - B|$ suffices for discriminability on the ANS mental number line. When the objective difference of 1 is sufficient for discriminability on the ANS mental number line and thus for the satisfaction of the confidence requirement, the observed threshold of *most* in (2a) will be 50%. However, if the objective difference of 1 is not sufficient (i.e. if a larger difference is needed for the two quantities to be discriminable), the observed threshold of *most* will be higher than 50%. This means that a sentence such as (2a) will be evaluated as false in certain cases in which it is semantically true: when objectively $|A \cap B| > |A - B|$, but $|A \cap B|$ and $|A - B|$ are not discriminable on the ANS mental number line.

To summarize, according to the account described in this section, the fact that the observed threshold of *most* is higher than 50% follows from (10): it is a consequence of representation and comparison of quantities on the ANS mental number line, in conjunction with the confidence requirement.

5.3 Pragmatic strenghtening as scalar implicatures

The account developed in this section is that the pragmatic mechanism that results in thresholds of *most* being higher than 50% is a scalar implicature (SI)[4]. The idea is that *most* triggers SIs which lead to pragmatic strengthening: in a sentence such as (2a), the higher the threshold of *most*, the stronger the meaning of the sentence. The idea that *most* triggers pragmatic inferences which result in differences in thresholds between *most* and *more than half* connects the account developed here to Horn's account (cf. Horn (2006) and Section 5.1), as well as to Ariel's 2004 suggestion that the observed threshold of *most* may be higher than 50% due to a conversational implicature of noteworthiness.

According to most theories of SIs, these inferences are negations of more informative (i.e. logically stronger) alternative utterances a sentence may activate (Grice, 1975; Horn, 1972, a.o.). There may be various versions of the SI account for the differences in thresholds between *most* and *more than half* which may differ in the implicature computing mechanism and/or postulated alternative utterances; for concreteness, we develop one version of the SI account (see Carcassi and Szymanik (2021) for another). More specifically, it will be outlined 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 (11): (11) asymmetrically entails (2b). The SI of (2b) is thus the negation of (11).

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

In other words, (2b) enriched by its SI would be interpreted as (12):

(12) More than 50% and not more than 75% of As are Bs.

If (2a) were to compete with the pragmatically enriched meaning of (2b) in (12) (note that (12) asymmetrically entails (2a)), and thus as a consequence had the (higher-order) SI that (12) is false, (2a) would effectively be interpreted as (11). In other words, in this example, the threshold of *most* would be 75% as a consequence of a higher-order SI due to the competition with the pragmatically enriched meaning of *more than half*.[5]

6 Moving forward

In Sections 4 and 5, four accounts — one semantic and three pragmatic accounts — were

[4]We leave aside the debate of whether SIs are a pragmatic or a semantic phenomenon, as it is orthogonal to the present purposes (see Grice 1975; Sauerland 2004; Van Rooij and Schulz 2004; Schulz and Van Rooij 2006; Spector 2007; Franke 2011; Chierchia et al. 2012; Bergen et al. 2016, a.o. for various approaches to implicature derivation).

[5]It is further possible under this account that, due to contextual relevance considerations, in certain contexts *most* will compete not with the pragmatically enriched meaning of *more than half*, but with, say, the pragmatically enriched meaning of *more than two thirds*, or the like, which would affect the resulting threshold of *most*.

discussed which predict that the threshold of *most* should be higher than 50% in positive environments, while that of *more than half* should be 50%.

We propose to evaluate the plausibility of these accounts by studying how the two quantifiers are interpreted in negative environments: in particular, the question is how the thresholds of *most* and *more than half* compare in positive and negative environments. Two experiments will be reported which investigate this question.

(13) It is not the case that most of As are Bs.

For positive environments and a sentence such as (2a), the threshold can be thought of as the % x of As that are Bs, such that the sentence (2a) is (largely) evaluated as false for percentages lower than or equal to x, and (largely) evaluated as true for percentages higher than x; and vice versa for negative environments and a sentence such as (13).

Let us now examine in turn the predictions of the four accounts from Sections 4 and 5 for what the thresholds of *most* and *more than half* should be under negation.

6.1 Semantic account: Predictions for negative environments

In Section 4, the semantic account of *most* by Solt (2011, 2016) was described which explains why the threshold of *most* in sentences such as (2a) can be higher than 50%.

Recall that according to the semantic account, *most* incorporates a scale of degrees and its ordering relation into semantics, as in (3), repeated below, where $\succ_S$ is an ordering relation of the scale S.

(3) $[\![\text{most}]\!]^S = \lambda A.\lambda B.\mu_S(A \cap B) \succ_S \mu_S(A - B)$

When the scale with respect to which *most* is interpreted is a semiordered scale S, the predicted threshold for *most* in (2a) may be higher than 50%, and depends on the discriminability threshold of S.

What are the predictions of this account for the threshold of *most* under negation as in (13)? The truth conditions of (13) should be the complement of the truth conditions of (2a), i.e. (13) is true when (2a) is false, and (13) is false when (2a) is true. According to the semantics in (3), the truth of (13) simply requires that it is not the case that $\mu_S(A \cap B) \succ_S \mu_S(A - B)$. This means that according to the semantics of *most* in (3), (13) is true whenever $\mu_S(A \cap B) \sim_S \mu_S(A - B)$ or $\mu_S(A - B) \succ_S \mu_S(A \cap B)$. Crucially, this entails that the threshold of *most* shouldn't be affected by negation. The semantic account by Solt (2011, 2016) described in Section 4 therefore predicts that the threshold of *most* in negative sentence as in (13) may still be higher than 50%, in line with the discriminability threshold of the scale with respect to which (13) is evaluated.

Likewise, the threshold of *more than half* is predicted to not be affected by negation: as the threshold of *more than half* is 50% in positive environments, it should remain 50% in negative environments.

6.2 Horn's 2006 account: Predictions for negative environments

In Section 5.1, Horn's 2006 account for differences in thresholds between *most* and *more than half* was introduced. According to that account, *most* and *more than half* are truth-conditionally equivalent, but differ in terms of markedness: *most* is less marked than *more than half*. Because of this, *more than half* is only felicitous in marked situations, such as the one involving bare majorities, while in such marked situations *most* wouldn't be used (cf. SectiIt is not the case that on 5.1 and Horn (2006)).

What are the predictions of this account for the threshold of *most* in negative sentences such as (13)?

If one assumes together with Horn (2006) that situations involving near majorities are marked (ie. when e.g. 49% or 50% of As are Bs), just like the situations involving bare majorities are marked, the prediction is that (13) shouldn't be used in such marked situations and that (14) should specialize for them. Critically, the consequence of the latter is that the (observed) threshold of *most* in sentences such as (13) will be ***lower*** than 50%.

(14) It is not the case that more than half of As are Bs.

To summarize, according to this account, the semantic threshold of both *most* and *more than half* is 50%. However, while the observed threshold of *more than half* is expected to be 50% in both positive sentences such as (2b) and negative sentences such as (14), the expected threshold of *most* is higher than 50% in positive sentences such as (2a), but lower than 50% in negative sentences such as (13).

6.3 ANS and confidence requirement: Predictions for negative environments

According to the account presented in 5.2, the sentence (2a) is semantically true iff $|A \cap B| > |A - B|$. However, because of the confidence requirement and ANS involvement in the processing of sentences with *most*, participants may evaluate the sentence (2a) to be false in certain cases where $|A \cap B| > |A - B|$, but this cannot be established with confidence (cf. (10)). The consequence of this is that the observed threshold of *most* will sometimes be higher than 50%, even though its semantic threshold is 50%.

What are the predictions of this account for the threshold of *most* in negative sentences such as (13)? Truth-conditionally, (13) is true iff $|A \cap B| \leqslant |A - B|$. If participants represent and compare the quantities $|A \cap B|$ and $|A - B|$ on the ANS mental number line upon processing (13), under which circumstances they produce (13), or evaluate (13) to be true, while respecting the confidence requirement? Recall that, if two quantities A and B are not sufficiently discriminable on ANS mental number line, one cannot say with confidence whether $A > B$, $A < B$ or $A = B$ (cf. (10)). This means that, in order to say with confidence that (13) is true, quantities $|A \cap B|$ and $|A - B|$ need to be discriminable on the ANS mental number line in such a way that it can be established with confidence $|A \cap B| < |A - B|$. Therefore, the confidence requirement and its interaction with ANS spelled out in (10) entails that $|A - B|$ needs to be greater than $|A \cap B|$ by at least the amount necessary for their discriminability on the ANS mental number line in order for speakers to utter (13) or for listeners to evaluate (13) to be true. To put it differently, according to (10), the number of As that are Bs need to be sufficiently lower from the number of As that are not Bs for (13) to be uttered by speakers or evaluated to be true by listeners. The consequence of this is that the observed threshold of *most* in negative sentences such as (13) should in fact be ***lower*** than 50%.

On the other hand, to the extent that processing of *more than half* doesn't involve ANS representations, but rather precise quantity representations and that accordingly its threshold is 50% in positive environments, its threshold should remain 50% in negative environments as well.

6.4 SI account: Predictions for negative environments

Let us next examine what the predictions of the SI account presented in Section 5.3 is with respect to how *most* is interpreted in negative sentences such as (13).

To this end, we start by describing an uncontroversial case of a sentence which triggers a SI, and observe that the SI in question is absent when the sentence is embedded under negation.

A sentence such as (15a) is commonly assumed to activate (15b) as alternative; note that (15b) asymmetrically entails (15a). Because of this, (15a) triggers as its SI the negation of (15b). In other words, (15a) enriched by its SI is interpreted as (15c) (Grice, 1975, a.o.).

(15) 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 (15a) is embedded under negation, as in (16). If the embedded sentence 'John ate a cookie or an apple' had the same SI in (16) as in (15a), the interpretation of (16) with its SI would be in (17). Such an interpretation of (16) is not naturally available[6]; (16) is naturally interpreted to be true in the situation in which John ate neither a cookie nor an apple, but false in the situation in which he ate both a cookie and an apple. In other words, the SI is not derived under negation.

(16) It is not the case that John ate a cookie or an apple.

(17) It is not the case that John ate a cookie or an apple but not both. (=John ate neither a cookie nor an apple, or John ate both a cookie and an apple.)

Importantly, when SIs are derived in positive sentences, they result in a logically stronger meaning ((15c) asymmetrically entails (15a)). However, if SIs were derived under negation, they would result in a logically weaker interpretation of the sentence (the logical meaning of (16) asymmetrically entails (17)). More generally, SIs are known to not be derived when their derivation would result in meaning weakening (but see Fox and Spector (2018) for a recent discussion).

The consequence of this is that, if the threshold of *most* is higher than 50% due to strengthening SIs, such SIs should not be derived in negative sentences, because they would lead to weakening. In other words, according to the SI account, the threshold of *most* in negative sentences such as (13) should be 50%. Importantly, this prediction holds not only of the version of the SI account outlined in Section 5.3: in principle, the prediction holds of any account which derives differences in thresholds between *most* and *more than half* via SIs of *most*. Note that as the threshold of *more than half* of 50% in positive environments is a consequence of its truth-conditional semantics and not of SIs, the expectation of the present account is that the threshold of *more than half* in negative environments should remain 50%.

It is worthwhile to point out that other accounts for the differences in thresholds between *most* and *more than half* due to pragmatic inferences other than SIs may be conceivable (cf. also Horn (2006) and Section 5.1). Whenever the derivation of such inferences is conditioned on them being *strengthening* inferences, these accounts will predict that the threshold of *most* in negative sentences should be (at most) the same as the threshold of *more than half*, i.e. (at most) 50%, for the same reason as the SI account: namely, such inferences would lead to weakening and not to strengthening under negation.

6.5 Summary of predictions

In Sections 6.1, 6.2, 6.3 and 6.4, the predictions of four accounts for differences in thresholds between *most* and *more than half* have been worked out for what these thresholds should be under negation. The predictions of the semantic account (Section 4), Horn's 2006 account (Section 5.1), the ANS and confidence requirement account (Section 5.2) and the SI account

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

(Section 5.3) are summarized in Tables 1, 2, 3 and 4 respectively.

In what follows, two experiments are reported which investigate the thresholds of *most* and *more than half* in positive and negative environments.

Table 1: **Predicted thresholds: Semantic account**

	More than half	**Most**
Positive environment	50%	> 50%
Negative environment	50%	> 50%

Table 2: **Predicted thresholds: Horn's 2006 account**

	More than half	**Most**
Positive environment	50%	> 50%
Negative environment	50%	< 50%

Table 3: **Predicted thresholds: ANS + confidence requirement account**

	More than half	**Most**
Positive environment	50%	> 50%
Negative environment	50%	< 50%

Table 4: **Predicted thresholds: SI account**

	More than half	**Most**
Positive environment	50%	> 50%
Negative environment	50%	50%

7 Experiment 1

Experiment 1 investigates thresholds of quantifiers *most* and *more than half* in a positive environment, as in (18a) and (19a) (unembedded *most* and *more than half*), and in a negative environment, as in (18b) and (19b) (*most* and *more than half* in the complement clause of *It is not the case that*).

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

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

7.1 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

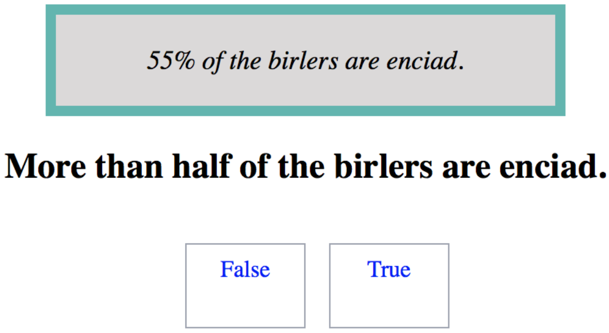


Figure 1: A test item from Experiment 1

'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.

7.2 Materials

Background information consisted of a single sentence which had the format in (20), with A a pseudo-noun with (regular) plural morphology, B a pseudo-adjective or a pseudo-noun with (regular) plural morphology , and $n \in [0, 100]$.[7]

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

Depending on the factor ENVIRONMENT, test sentences had the format in (21a) (*positive environment*), or (21b) (*negative environment*), 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 ENVIRONMENT and TARGET QUANTIFIER, the experiment employed a between-subject design: each participant was exposed to only one of the four combinations of ENVIRONMENT×TARGET QUANTIFIER. There were thus four experimental versions.

(21) 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 (22)-(25).

(22) ENVIRONMENT: *positive*, TARGET QUANTIFIER: *most*
Most of the birlers are enciad.

(23) ENVIRONMENT: *negative*, TARGET QUANTIFIER: *most*
It is not the case that most of the birlers are enciad.

(24) ENVIRONMENT: *positive*, TARGET QUANTIFIER: *more than half*
More than half of the birlers are enciad.

[7] 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://github.com/milicaden/most-mth-thresholds. Pseudowords used in Experiment 1 largely overlap with pseudowords used in the experiment reported in Ramotowska et al. (2020).

(25) ENVIRONMENT: *negative*, 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 (21a) and (21b). 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 (26)-(29).

(26) ENVIRONMENT: *positive*, CONTROL QUANTIFIER: *some*
Some of the birlers are enciad.

(27) ENVIRONMENT: *negative*, CONTROL QUANTIFIER: *some*
It is not the case that some of the birlers are enciad.

(28) ENVIRONMENT: *positive*, CONTROL QUANTIFIER: *all*
All of the birlers are enciad.

(29) ENVIRONMENT: *negative*, CONTROL QUANTIFIER: *all*
It is not the case that all of the birlers are enciad.

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 the *positive*, and the other half in the *negative* environment). Crucially, the truth evaluation of both target and control items depended on the % exhibited in the background sentence, which will be referred to 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 participants and across four experimental versions[8]. 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.

7.3 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, 29 participants were excluded who made more than 20% of errors on what will be referred to as uncontroversial control items. These were all control items apart from (i) those in which the control quantifier *some* was in the *positive* environment and test % $>$ 50 and (ii) those in which the control quantifier *all* was in the *negative* 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, responses whose response time was more

[8]The test %s in the target items were: 3, 4, 6, 7, 9, 10, 11, 12, 13, 17, 18, 20, 26, 28, 29, 30, 33, 34, 37, 38, 41, 42, 44, 46, 52, 53, 55, 59, 61, 65, 66, 67, 68, 70, 71, 73, 77, 81, 84, 85, 86, 88, 92, 93, 94, 96, 97, 99.

than 2 SD away from the average were excluded (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).

7.4 Participants' threshold estimation

To estimate participants' thresholds on target items, a method similar to that of Ramotowska et al. (2020) was used. For each participant, we fit the logistic function in (30):

(30)

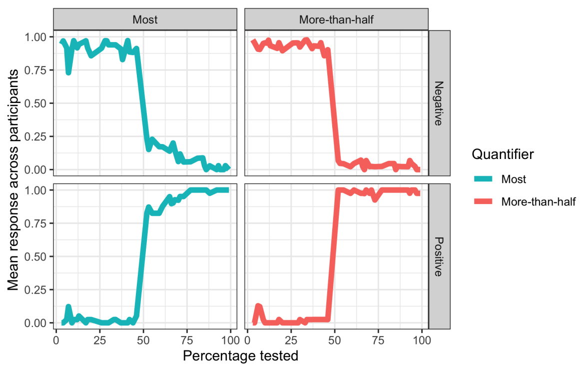
$$P(test\ \%) = \frac{1}{1 + e^{\frac{(T - test\ \%)}{k}}}$$

$Test\ \%$ 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.

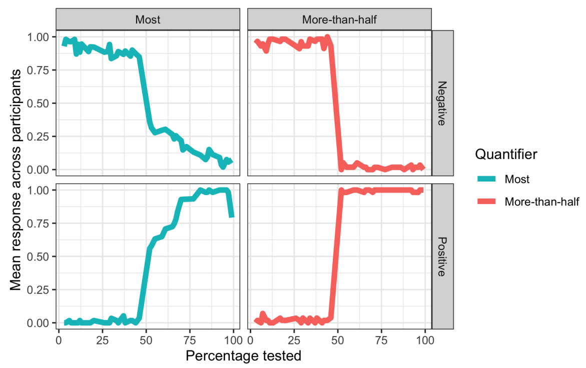
To estimate T and k parameter for each participant, fitting was done using R *nls* self-starter function with the following specification: the algorithm used was 'nl2sol' (Dennis Jr et al., 1981); the starting value for T = 50; the starting value for $k = 4$ for the two experimental versions in which *most* and *more than half* were in positive environments, and $k = -4$ for the two experimental versions in which *most* and *more than half* were in negative environments.

7.5 Results

Figure 2a represents responses across all tested percentages, averaged across participants. Prior to any analysis, it is interesting to observe the difference between the two quantifiers in the range 50-75%. While the average response drops gradually in that range in the case of *most* in the negative environment, it does so more sharply in the case of *more than half*. Similarly, while the average response rises gradually in that range in the case of *most* in the positive environment, it does so more sharply in the case of *more than half*. This is suggestive of the differences in thresholds between the two quantifiers.



(a) Experiment 1



(b) Experiment 2

Figure 2: Responses across percentages averaged across participants in Experiments 1 and 2 by Quantifier and Environment.

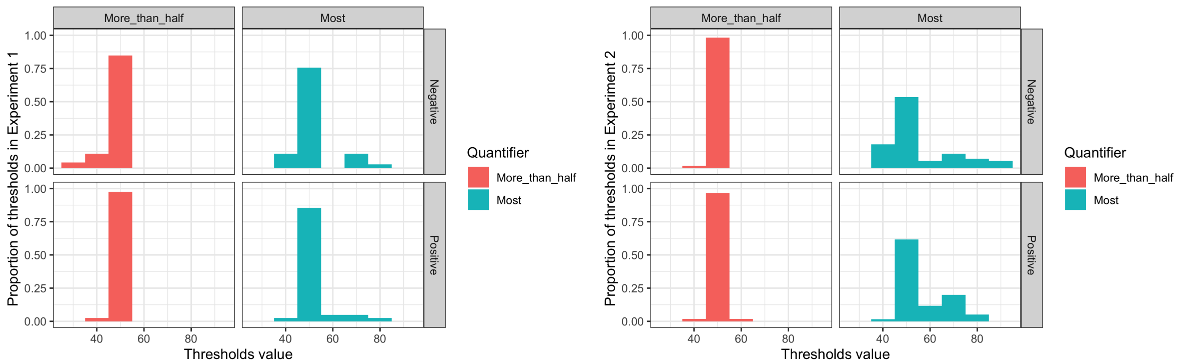
Because of the sampled test %s, if a participant's threshold was 50, their expected threshold in our experiment was 49.[9] Mean participants' thresholds and SDs for the four experimental

[9] 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).

Table 5: **Experiment 1: Thresholds results**

	More than half	**Most**
Positive environment	$M = 48.3, SD = 2.1$	$M = 51.6, SD = 7.2$
Negative environment	$M = 47.3, SD = 4.4$	$M = 51, SD = 9$

versions are in Table 5. The distribution of thresholds from Experiment 1 can be visualized in Figure 3a.



(a) Experiment 1 (b) Experiment 2

Figure 3: Threshold distribution in Experiments 1 and 2 by Quantifier and Environment.

Let us now turn now to the main question of interest: how do thresholds of *most* and *more than half* differ in *positive* and *negative* environments?

Welch's unequal variances t-test confirmed that in the *positive* environment the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* ($t(46.5) = 2.8, p < .01$): 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 the *negative* environment as well ($t(49.6) = 2.3, p = .03$).

Two-way ANOVA didn't reveal an interaction between Quantifier (Most vs. More than half) and Environment (positive vs. negative) ($F(1) = 0.05, p = .8$). 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 Environment either ($F(1) = 0.3, p = .6$).

Finally, it is interesting to note that there is much more variability (as evidenced by the SDs reported in Table 5) in participants' thresholds for the quantifier *most* than for the quantifier *more than half* (cf. also the results in Ramotowska et al. (2020)). This point will be discussed in Section 10.

8 Experiment 2

Experiment 2 extends the investigation to another pair of positive and negative 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 negative environments, and are not an artifact of the specific negative environment tested in Experiment 1. In Experiment 2, the positive environment is the scope of the quantifier *each boy*, as in (31a) and (32a). The negative environment is the scope of the quantifier *no boy*, as in (31b) and (32b).

(31) a. Each boy puggled **most** of his entands.
b. No boy puggled **most** of his entands.

(32) a. Each boy puggled **more than half** of his entands.
b. No boy puggled **more than half** of his entands.

For the purposes of Experiment 2, the threshold of *most* in the positive environment (31a) can be thought of as the lowest % x such that (31a) is evaluated as true if each boy puggled more than x% of his entands and false otherwise (similarly for *more than half* and (32a)). The threshold of *most* in the negative environment (31b) can be thought of as the highest % x such that (31b) is evaluated as true if no boy puggled more than x% of his entands and false otherwise (similarly for *more than half* and (32b)).

The predictions of the four accounts for the threshold of *most* (semantic account from Section 4, Horn's 2006 account from Section 5.1, ANS and confidence requirement account from Section 5.2 and SI account from Section 5.3) are largely the same for the scope of sentential negation and the scope of a negative quantifier. Let us examine the predictions of the four accounts in turn. Suppose that there are only two boys, Ted and Dan, with respect to whom the truth of (31b) is evaluated, as it will be the case in Experiment 2. In such a case, the sentence (31b) is truth-conditionally equivalent to 'Ted didn't puggle most of his entands and Dan didn't puggle most of his entands'. Because of this equivalence, it is easy to see that the predictions of the semantic account and the ANS and confidence requirement account for the threshold of *most* in the scope of a negative quantifier are the same as for the scope of sentential negation (cf. Sections 6.1 and 6.3). In other words, the semantic account predicts that the threshold of *most* in the scope of a negative quantifier will be higher than 50%, and the ANS and confidence requirement account predicts that it will be lower than 50%. Let us now discuss the predictions of Horn's 2006 account. The predictions will depend on what types of situations in which neither of the boys puggled more than 50% of their entands are marked, as according to the account, sentences with *most* shouldn't be used for marked situations. There are two options. The first one is that as soon as at least one of the boys puggled a near majority of his entands, the situation is marked. The second one is that only if both boys puggled a near majority of their entands, the situation is marked. In the first case, the threshold of *most* is predicted to be lower than 50%. In the second case, the threshold of *most* is predicted to be 50% if there is at least one boy who puggled less than a near majority of his entands, and lower than 50% if both boys puggled a near majority of their entands. The general prediction of Horn's 2006 account is thus that the threshold of *most* will be at most 50% in the scope of a negative quantifier. When it comes to the SI account for the threshold of *most*, SIs under a negative quantifier would lead to weakening instead of strengthening, and as such are predicted to not be derived in that environment, just like it was the case for the scope of sentential negation (cf. Section 6.4). In other words, the SI account predicts that the threshold of *most* will be 50% in the scope of a negative quantifier.

8.1 Task

The task was the same as in Experiment 1. An example of a test item is in Figure 4.

8.2 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 (33), with A a pseudo-verb with (regular) past tense morphology, B a pseudo-noun with

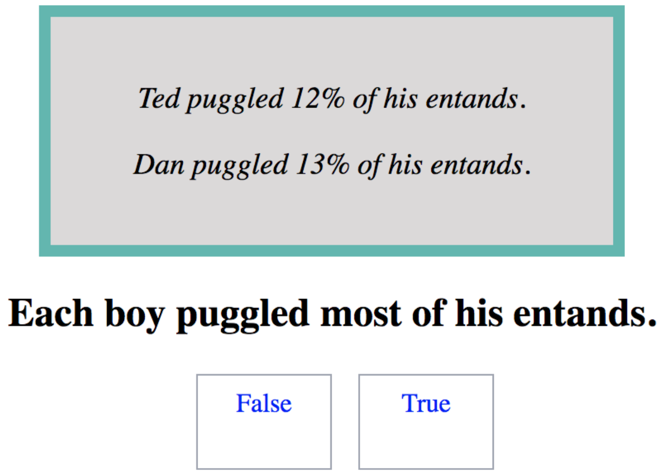


Figure 4: A test item from Experiment 2

(regular) plural morphology, and $n, m \in [0, 100]$.[10]

(33) Ted A $n\%$ of his B.
Dan A $m\%$ of his B.

Test sentences had the format in (34), with A and B the same pseudowords as in the background information sentences. Q_1 reflects the level of the factor ENVIRONMENT: Q_1 was *each* (*positive environment*) or *no* (*negative environment*). 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 ENVIRONMENT and TARGET QUANTIFIER, the experiment employed a between-subject design: each participant was exposed to only one of the four combinations of ENVIRONMENT×TARGET QUANTIFIER. There were thus four experimental versions in Experiment 2 as well.

(34) Q_1 boy A Q_2 of his B.

One example target item for each combination of factors is in examples (35)-(38).

(35) ENVIRONMENT: *positive*, TARGET QUANTIFIER: *most*
Each boy puggled most of his entands.

(36) ENVIRONMENT: *negative*, TARGET QUANTIFIER: *most*
No boy puggled most of his entands.

(37) ENVIRONMENT: *positive*, TARGET QUANTIFIER: *more than half*
Each boy puggled more than half of his entands.

(38) ENVIRONMENT: *negative*, TARGET QUANTIFIER: *more than half*
No boy puggled more than half of his entands.

Control items also had the format as in (34). They differed from target items in that Q_2 reflects the level of CONTROL QUANTIFIER: Q_2 was existential *some/any* (*some* when ENVIRONMENT was *positive*, *any* when ENVIRONMENT was *negative*[11]) or universal *all*. Control items were

[10]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://github.com/milicaden/most-mth-thresholds.

[11]The existential *any* was used instead of *some* in negative environments because the positive polarity status of

the same across four experimental versions.

One example control item for each combination of factors is in examples (39)-(42).

(39) ENVIRONMENT: *positive*, CONTROL QUANTIFIER: *some*
Each boy puggled some of his entands.

(40) ENVIRONMENT: *negative*, CONTROL QUANTIFIER: *any*
No boy puggled any of his entands.

(41) ENVIRONMENT: *positive*, CONTROL QUANTIFIER: *all*
Each boy puggled all of his entands.

(42) ENVIRONMENT: *negative*, 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 the *positive*, and the other half in the *negative* environment). Crucially, the truth evaluation of both target and control items depended on the two %s exhibited in the background sentences. Note that when ENVIRONMENT is *positive*, the smaller of the two %s in the background information sentences is determining the truth of the test sentence; when ENVIRONMENT is *negative*, the greater of the two %s is. In the context of Experiment 2, the % which determines the truth of the test sentence will henceforth be referred to 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 *positive* environment, the non-test % was obtained by adding 1 or 2 to the test %; in the cases of *negative* environment, 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.

8.3 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, 23 participants were excluded 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, responses whose response time was more than 2 SD away from the average were excluded (response times were log-transformed). Finally, one participant was excluded because the estimation of their threshold failed to converge and two participants were excluded because their estimated threshold wasn't in the range (0,100). 234 participants remained for the analysis, with comparable number of participants across four experimental versions (60, 56, 58, 60).

8.4 Participants' threshold estimation

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

---

some precludes its interpretation in the direct scope of a negative quantifier, cf. Baker, 1970, a.o.

Table 6: **Experiment 2: Thresholds results**

	More than half	**Most**
Positive environment	$M = 48.6, SD = 2.3$	$M = 55.9, SD = 9.5$
Negative environment	$M = 48.6, SD = 1.5$	$M = 55.2, SD = 13.8$

8.5 Results

Figure 2b represents responses across all tested percentages, averaged across participants. The observations about the range 50-75% made in the context of Experiment 1 are applicable to Experiment 2 as well.

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 6. The distribution of thresholds from Experiment 2 can be visualized in Figure 3b.

Let us turn now to the main question of interest: how do thresholds of *most* and *more than half* differ in *positive* and *negative* environments?

The results confirm the findings of Experiment 1. Welch's unequal variances t-test confirmed that in the *positive* environment the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* ($t(65.9) = 5.8, p < .01$). Crucially, again, the participants' thresholds for the quantifier *most* are higher than the participants' thresholds for the quantifier *more than half* in the *negative* environment as well ($t(56.1) = 3.6, p < .01$).

Two-way ANOVA didn't reveal an interaction between Quantifier and Environment ($F(1) = 0.1, p = .8$). However, aligned-rank ANOVA revealed a significant Quantifier-Environment interaction ($F(1) = 6.9, p < .01$). We will come back to this interaction effect in the Discussion section.

Finally, note again that there is much more variability (as evidenced by the SDs reported in Table 6) in participants' thresholds for the quantifier *most* than for the quantifier *more than half*. A potential source of this variability will be discussed in Section 10.

9 Interim discussion

When it comes to thresholds of quantifiers *most* and *more than half*, the empirical picture is the following. In positive environments, thresholds of *most* are on average higher than thresholds of *more than half*, in line with the results of Ariel (2004); Solt (2011, 2016); Ramotowska et al. (2020).

Crucially, in Experiment 1 and 2 it was demonstrated that *most* and *more than half* retain this property in negative environments. This result is compatible with the semantic account of differences in thresholds between *most* and *more than half* (cf. Solt (2011, 2016), Section 4). On the other hand, it is at odds with the three pragmatic accounts — Horn's 2006 account, ANS and confidence requirement account, SI account — discussed in Section 5, each of which predicts that the threshold of *most* will be either 50% or lower in negative environments. Pending alternative pragmatic accounts, our experimental results thus speak in favor of there being a difference in truth-conditional meaning between *most* and *more than half*, in line with the proposal by Solt (2011, 2016) discussed in Section 4.[12]

[12] There are two further properties of the data pertaining to thresholds that deserve mention. First, the observed differences in thresholds between *most* and *more than half* are greater in quantificational contexts (Experiment 2) than in non-quantificational contexts (Experiment 1), as revealed by comparison of average thresholds (Tables 5

According to the semantic account, while the truth-conditional import of *more than half* is in (44), that of *most* is in (43), with T being a contextually supplied parameter determining the threshold of *most*. The semantic account of *most* put forward in Solt (2011, 2016) incorporates a specific hypothesis about the nature of the T parameter (namely, the T parameter is derived from the contextually supplied measurement scale). It's important to highlight however that our results would be compatible with alternative semantic accounts assigning to *most* and *more than half* the semantics in (43) and (44) respectively which may postulate a different source for the contextually supplied T parameter.

(43) $[\![\text{ most }]\!]^T = \lambda A.\lambda B.|A \cap B| \geq |A - B| + T$ (with $T \geq 1$)

(44) $[\![\text{ more than half }]\!] = \lambda A.\lambda B.|A \cap B| > \frac{1}{2} \cdot |A|$

10 Additional difference between *most* and *more than half*: Vagueness

The central goal of the present paper is to make progress on the question of where the differences in thresholds between quantifiers *most* and *more than half* come from.

There is however another important difference between the two quantifiers. 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).

Thresholds of vague expressions may vary across participants and contexts: this threshold variability is typically taken to be a hallmark of vague expressions (Kennedy, 2007). If *most*, but not *more than half*, is vague, that would explain the greater variability in participants' thresholds observed with *most* in Experiments 1 and 2.

There is also an experimental argument in favor of the vagueness of *most*. It is based on the 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 $n \in [0, 100]$, similarly to what was the case in our Experiments 1 and 2. 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* (but see footnote 13 for a concern with this interpretation).

In this section, the data collected in Experiments 1 and 2 is used to provide additional converging evidence for the vagueness of *most* coming from the differences in error pattern of *most* and *more than half.*

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

and 6), and of thresholds distributions (Figures 3a and 3b). To the extent that this contrast between quantificational and non-quantificational contexts is confirmed in future studies, it would be interesting to understand why and how such fine properties of linguistic contexts affects participants' thresholds. Second, there is in general more variability in individual thresholds in negative environments than in positive environments (cf. SDs in Tables 5 and 6). This is possibly due to negative environments being cognitively more difficult to process (Clark, 1976; Deschamps et al., 2015; Just and Carpenter, 1971, a.o.), leading to more participants' errors, and thus to less precise threshold estimates. Note however that as our main analyses involve comparing *most* and *more than half* in the same environment (be it positive or negative), the effect of environment on the precision of threshold estimates does not affect our conclusions.

is true or false of. 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 vague and thus 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 ask the following question: is participants' behaviour more inconsistent with their estimated thresholds when test %s are close to the estimated threshold in the case of *most* than in the case of *more than half*?

Before we proceed to the analysis, a caveat is described that needs to be addressed, coming from what is known about cognitive processing of numbers and ANS (cf. Sections 4 and 5.2).

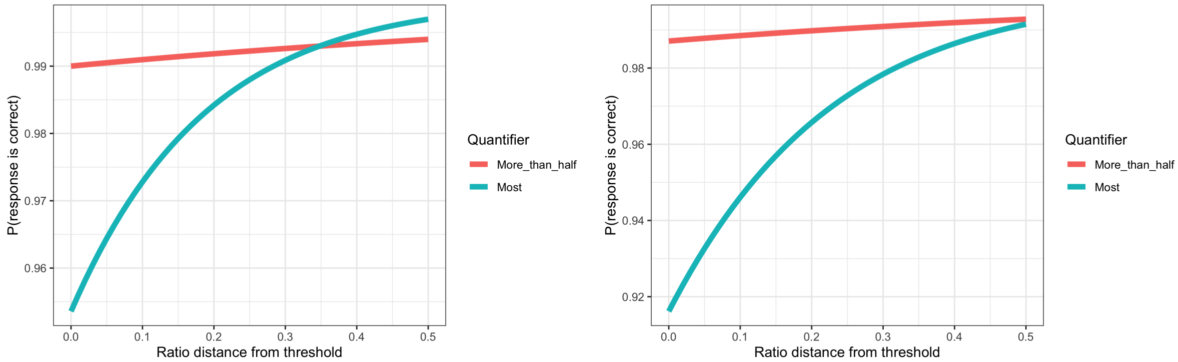
Recall that ANS representations of quantities are imprecise, and that whether or not ANS representations of two stimuli can be discriminated depends on the distance of their objective measures. In the case of ANS, the discriminability is function of the *ratio* of the objective measures of two stimuli: the larger the ratio, the better the discriminability (Dehaene, 1997). For instance, ANS representations of 20 and 10 are equally discriminable as those of 30 and 15, and more discriminable that those of 30 and 20. The consequence of this is that there may be more inconsistent behaviour around higher thresholds because of the ANS interference. As thresholds of *most* are on average higher than thresholds of *more than half*, participants' behaviour is expected to be more inconsistent with their thresholds in the case of *most* than in the case of *more than half* even if the semantics of *most* weren't vague.[13]

Because of this, in investigating whether the representation of *most* is vague, one needs to control for the ANS interference. This is achieved in the following way: upon examining how inconsistent participants' behavior is around their estimated thresholds, ratios of test %s relative to the participants' estimated thresholds are relied on as indicators of the distance of the test % from the threshold, rather than on the difference between the two.

We thus proceed as follows. The data for each participant p with an estimated threshold T_p is subsetted to the target trials whose test %s are within the range $[T_p - \frac{1}{2} \cdot T_p, T_p + \frac{1}{2} \cdot T_p]$. 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 a positive 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 there more incorrect responses closer to the participants' thresholds in the case of *most* than in the case of *more than half*? As indicated above, in order to control for the ANS interference confound, the closeness to the threshold of a given test % is measured in terms of the ratio between the test % and the threshold. That is, for each target trial with a test % x for each participant p, the *ratio distance* to p's threshold T_p, defined as the $|\frac{x}{T_p} - 1|$, is computed.

To assess the question of interest, a mixed-effects logistic regression model was fitted to participants' responses (*correct* vs. *incorrect*) separately for Experiment 1 and 2, with Quantifier (*Most* vs. *More than half*), Ratio 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-Ratio distance interaction has a significant effect on response correctness in both Experiment 1 ($\chi^2(1) = 9.89, p < .01$) and Experiment 2 ($\chi^2(1) = 9.08, p < .01$), 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 5a and 5b.

[13]This confound complicates the interpretation of the response time analysis as favoring the vagueness of *most* in Ramotowska et al. (2020) .



(a) Experiment 1 (b) Experiment 2

Figure 5: Probability that a response is correct on target trials as a function of quantifier (*most* vs. *more than half*) and of ratio 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.

11 *Most*: Two populations?

In the previous section, evidence was provided that *most*, but not *more than half*, is semantically vague. In this section, it will be explored whether differences in vagueness and differences in thresholds correlate.[14] More specifically, we ask whether there are two populations in terms of how *most* is interpreted: a population which assigns to *most* a non-vague interpretation with the threshold 50%, and a population which assigns to *most* a vague interpretation with the threshold possibly higher than 50%.

We address this question by looking at the error rate around the threshold and how it relates to the threshold value.

To this end we separate participants into those whose threshold is in the range [48,50], and those whose threshold is outside of that range (recall that the expected threshold in our experiments is 49; the range [48,50] is thus a narrow range centered on the expected threshold). As in the analysis reported in Section 10, for both groups of participants, we subset the data for each participant p with an estimated threshold T_p to the target trials whose test %s are within the range $[T_p - \frac{1}{2} \cdot T_p, T_p + \frac{1}{2} \cdot T_p]$. Similarly, each response of each participant is coded as correct or incorrect as a function of that participant's threshold. The proportion of correct responses in the aforementioned range is then computed. Welch's unequal variances t-test reveals that the proportion of correct responses is significantly higher in the group whose threshold is in the range [48,50] than those whose threshold is not in that range both in Experiment 1 ($M_1 = 0.99, M_2 = 0.89, t(33.3) = 4.5, p < .01$), and in Experiment 2 ($M_1 = 0.98, M_2 = 0.9, t(83) = 4.5, p < .01$). For reference, 45 participants in Experiment 1 have their threshold in the range [48,50] and 33 outside that range; 48 participants in Experiment 2 have their threshold in the range [48,50] and 68 outside that range.

One may worry however that the above result is driven by the fact that some thresholds estimates may be inaccurate. More specifically, if there happens to be fewer inaccurate threshold estimates in the range [48, 50] than outside of that range, one would expect less noise (i.e. more correct responses) with former thresholds as opposed to the latter even if neither former nor latter were associated with a vague meaning. While this alternative explanation cannot be

[14]Many thanks to an anonymous reviewer for raising this question.

eliminated entirely, one way to reduce its plausibility is to re-do the aforementioned analysis on a subset of participants for whom there is very high confidence that their threshold estimate is accurate. To this end, all participants p with an estimated threshold T_p whose proportion of correct responses in the range $[T_p - \frac{1}{2} \cdot T_p, T_p + \frac{1}{2} \cdot T_p]$ is lower than 0.9 were excluded. This is a strict criterion; the participants who are not excluded gave very few, if any, responses inconsistent with their estimated threshold. Crucially, the above result is entirely replicated even in this restricted group. In other words, even in this group the proportion of correct responses is significantly higher in the group whose threshold is in the range [48,50] than those whose threshold is not in that range both in Experiment 1 ($M_1 = 0.99, M_2 = 0.96, t(26) = 3.3, p < .01$), and in Experiment 2 ($M_1 = 0.98, M_2 = 0.97, t(83.5) = 2.4, p = .02$). In this restricted group, 45 participants in Experiment 1 have their threshold in the range [48,50] and 21 outside that range; 45 participants in Experiment 2 have their threshold in the range [48,50] and 45 outside that range.

This result suggests that there are two populations in terms of how *most* is interpreted: one which interprets it with a non-vague, more-than-half meaning, and the other, which interprets it with a vague meaning with the threshold possibly higher than 50%.

12 Discussion

The main focus of the present paper are differences in thresholds between quantifiers *most* and *more than half*. In Experiments 1 and 2, the thresholds of *most* and *more than half* are compared in positive and negative environments. The threshold of *most* is found to be higher than that of *more than half* in both positive and negative environments. As discussed in Section 9, this speaks against the three pragmatic accounts presented in Section 5 and in favor of a difference in truth-conditional semantics of the two expressions: the semantics of *most* is in (43), with T a contextually supplied parameter, while the semantics of *more than half* is in (44).

(43) $[\![\text{ most }]\!]^T = \lambda A.\lambda B.|A \cap B| \geq |A - B| + T$ (with $T \geq 1$)

(44) $[\![\text{ more than half }]\!] = \lambda A.\lambda B.|A \cap B| > \frac{1}{2} \cdot |A|$

As explained in Section 9, these semantic entries are in line with those put forward in Solt (2011, 2016). Importantly however, her account incorporates a specific hypothesis about the nature of the T parameter in (43) coming from the contextually supplied measurement scale. Again, it is important to note that our results would also be in line with alternative semantic accounts assigning to *most* and *more than half* the semantics in (43) and (44) respectively which may postulate a different source for the contextually supplied T parameter.

There is however a number of points to discuss in assigning to *most* the truth-conditional semantics in (43). Let us examine them in turn.

12.1 Vagueness

In addition to establishing that *most* and *more than half* differ in their truth-conditional semantics, evidence was presented that (i) *most* is vague, but *more than half* isn't, or, rather, that *most* is more vague than *more than half* (Section 10), and (ii) that there is a correlation between differences in vagueness and differences in thresholds between the two quantifiers (Section 11).

That *most* is vague is compatible with Solt's semantics of *most*. Here is one possible connection between the discriminability threshold of semiordered scales and vaguenesss. While it is natural to think of discriminability thresholds of semiordered scales as rooted in perceptual discriminability, this needn't be the case: the discriminability threshold of a semiordered scale may also be rooted in what's discriminable in terms of, e.g., noteworthiness in some context, which is a stronger criterion than perceptual discriminability. To put it differently, contextual considerations may play a role in determining which objective differences between quantities are to be considered discriminable on some semiordered scale supplied in that context. In such cases, the vagueness of *most* may be a consequence of participants's uncertainty about the exact value of the contextual discriminability threshold.

How does the semantics of *most* in (43) fit together with the finding that there is a correlation between differences in thresholds and differences in vagueness? While we do not have a conclusive answer at this point, there are three plausible options one could pursue. The first is that *most* is in fact lexically ambiguous between $most_1$, whose semantic entry would be the same as that of *more than half* in (44) (i.e. non-vague, with threshold 50%), and $most_2$, whose semantic entry would be that in (43) (i.e. vague, with threshold possibly higher than 50%). If this is correct, an open question for future work is whether each individual speaker has two entries for *most*, such that they may access different entries in different contexts, or rather whether each individual speaker has a unique entry for *most*, but which entry it is may vary between speakers (for the variation of the latter kind, see Han et al. (2016)). A related option, in line with Solt's account, is that this is not a case of lexical ambiguity, but rather, a case of ambiguity brought about by association with different measurement scales. In particular, according to Solt's account, when *most* associates with a semiordered scale, it may be vague and its threshold may be higher than 50%; when it associates with a more precise measurement scale, it isn't vague and its threshold is 50%. According to this approach, the two populations that have been identified differ in the measurement scales with respect to which they interpret sentences with *most*. Finally, the third option is that there is a unique lexical entry for *most*, with semantics as in (43), but that uncertainty about the T parameter varies together with participants' estimate of T: when participants' estimated T is 1 (i.e. when the semantic threshold of *most* is 50%), they are confident about their estimate of T and the resulting interpretation of *most* isn't vague; on the other hand, when participants' estimate T is greater than 1 (i.e. when the semantic threshold of *most* is higher than 50%), they are less confident about their estimate of T, and the resulting interpretation of *most* is vague. Of course, more work is needed to develop the latter option into a full proposal; an important question to pursue in this respect is how and why participants' uncertainty about the threshold varies with their estimate of it.

12.2 Quantifier-Environment interaction in Experiment 2

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

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

Given that this effect was not observed in Experiment 1, we are not in a position to draw strong conclusions from it. To the extent that future work establishes that this effect is robust, however, here are two potential explanations for it.

The first is that pragmatic strengthening inferences (e.g. SIs which result in an increased threshold in positive environments) 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 positive environments is what is lost in negative 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 positive environments as compared to the baseline, while negative strengthening would lead to decreased thresholds of *most* in negative environments 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 (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018). Consider for example (45a) and (45b):

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

Due to negative strengthening inferences, (45a) may sometimes suggest that rather few people came, and (45b) that John is rather short. This interpretation is stronger than the complement of the interpretation of the positive versions of (45a) and (45b): 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 negative environments: if our participants were deriving them in Experiment 2, this can explain the significant interaction as well.

Note 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 positive environments or negative strengthening in negative environments — characterizes *most*, in addition to the truth-conditional differences from *more than half*.

12.3 Effect of experimental setting on differences in thresholds?

Not all experimental work on *most* and *more than half* has found that there are differences in thresholds between the two quantifiers. Hackl (2009) and Pietroski et al. (2009) studied the interpretation of sentences with *most* in a picture verification task. For instance, their participants had to evaluate whether a sentence such as *'Most of the dots are blue'* is true with respect to an image containing a certain number of blue and a certain number of non-blue dots. These authors argue that *most* and *more than half* are truth-conditionally equivalent. 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).[15] Pietroski et al., 2009 also discuss various properties of their data which suggest that in their study *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 and Pietroski et al., 2009 is clearly not incompatible with the semantic entry for *most* in (43): participants in these studies may have been prone to setting the T parameter in the entry in (43) to 1, resulting in the threshold of $\approx 50\%$. 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 and Pietroski et al., 2009, while such differences were observed in our study, as well as in the studies of Ariel, 2003, 2004 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 the experimental setting or some other factor which influences the threshold of *most* that participants settle for remains to be explored in future research. In particular, if the former is the case, and if the discriminability threshold of a contextually supplied semiordered

[15]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*.

scale with respect to which *most* is interpreted is sometimes rooted in differences which are contextually noteworthy, one may imagine various ways in which talk about percentages may lead to different expectations of noteworthiness than talk about the number of dots.

The possibility that participants select different thresholds in different experimental settings raises a more general question: how do participants decide on the threshold of *most* in different contexts? The same question has been raised in the literature on vague adjectives and vague quantifiers *many* and *few* (Qing and Franke, 2014a,b; Schöller and Franke, 2017). For the case of *most*, this question remains open for future work.

12.4 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 the specifics of verification procedures of *most*.

For instance, Pietroski et al. (2009) discuss two possible strategies one may employ upon a verification of a sentence *'Most of the dots are blue'* with respect to an image, assuming that such a sentence is semantically true as soon as the number of blue dots exceeds the number of non-blue dots. One may (i) check that each non-blue dot can be paired with a different blue dot, with some blue dots remaining unpaired — call this *pairing with remainder strategy*; or (ii) one may compute the total number of blue dots, the total number of non-blue dots, and compare these two quantities to establish that the former is greater than the latter — call this *cardinality comparison strategy*. They find that even in situations in which the pairing with remainder strategy would be more efficient, in the sense that it would lead to greater accuracy and faster response times than the cardinality comparison strategy, participants nonetheless seem to persist in computing and comparing cardinalities of blue and non-blue dots. They interpret this finding to suggest that the mental representation of *most* encodes cardinality comparison.

What consequences does the conclusion that *most* and *more than half* are not truth-conditionally equivalent have on conclusions by Pietroski et al., 2009 about mental representation of *most*? At the very least, our results call for a great care when drawing conclusions about the mental representation of *most* from its verification procedures, as in Hackl, 2009; Szymanik and Zajenkowski, 2010; Lidz et al., 2011; Pietroski et al., 2009; Steinert-Threlkeld et al., 2015; Talmina et al., 2017. For instance, in the context of the study by Pietroski et al. (2009), if participants interpret *most* as having a threshold higher than 50%, the pairing with remainder strategy would not always yield a correct result — in such cases, some cardinality computation is necessary to ensure that the result of the verification is in alignment with the interpretation of *most*. This raises a question of whether strong conclusions should be drawn from the fact that participants in the study by Pietroski et al. (2009) did not use the pairing with remainder strategy to verify sentences with *most*. More generally, behavioral verification patterns which would be surprising if *most* were truth-conditionally equivalent to *more than half* may become entirely unsurprising once the difference in truth conditions is taken into account. This cautionary remark holds even if participants settle for the threshold of *most* of $\approx 50\%$ throughout verification experiments as it is argued to be the case by Pietroski et al. (2009) in their study (cf. Section 12.3). More concretely, suppose that participants more commonly assign the threshold to *most* that is greater than 50% outside of the lab, and that for such an interpretation, the most efficient verification strategy typically involves cardinality comparison, and is thus most commonly used. It is conceivable that participants would preserve some properties of their most commonly used verification strategy when they decide to lower their threshold of *most* to $\approx 50\%$ inside the lab: this may ultimately be less cognitively expensive and therefore more efficient than to search for the most optimal verification algorithm at each verification instance.

13 Conclusion

In this paper, the results of two experiments were reported demonstrating that participants' thresholds of *most* are higher than their thresholds of *more than half* in both positive and negative environments. This result is at odds with the three pragmatic accounts discussed in Section 5, and suggests that these two quantifiers are not truth-conditionally equivalent, in line with the semantic account by Solt (2011, 2016) which is discussed in Section 4.

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 more generally to what is known about quantification in natural language. They strengthen the case that the truth-conditional meaning of *most* should be modelled on a par with other vague quantification expressions whose semantics involves a contextually supplied threshold, such as *many* and *few*. At this point, many interesting questions remain open for future work in relation to how thresholds of such expressions are determined in different contexts, as well as in relation to the connection between thresholds' values and vagueness.

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.

Baker, C. L. (1970). Double negatives. *Linguistic inquiry 1*(2), 169–186.

Barth, H., A. Baron, E. Spelke, and S. Carey (2009, aug). Children's multiplicative transformations of discrete and continuous quantities. *Journal of Experimental Child Psychology 103*(4), 441–454.

Bartsch, R. and T. Venneman (1973). *Semantic Structures: A study in the relation between syntax and semantics*. Athenium, Frankfurt am Main.

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.

Bergen, L., R. Levy, and N. Goodman (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics 9*.

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

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

Cantlon, J. F., M. L. Platt, and E. M. Brannon (2009). Beyond the number domain. *Trends in cognitive sciences 13*(2), 83–91.

Carcassi, F. and J. Szymanik (2021). An alternatives account of 'most' and 'more than half'. Manuscript under review.

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.

Chierchia, G., D. Fox, and B. Spector (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Semantics: An international handbook of natural language meaning 3*, 2297–2332.

Clark, H. H. (1976). *Semantics and comprehension*. Mouton.

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.

Cresswell, M. J. (1976). The semantics of degree. In *Montague grammar*, pp. 261–292. Elsevier.

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.

Dennis Jr, J. E., D. M. Gay, and R. E. Walsh (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software (TOMS) 7*(3), 348–368.

Deschamps, I., G. Agmon, Y. Loewenstein, and Y. Grodzinsky (2015). The processing of polar quantifiers, and numerosity perception. *Cognition 143*, 115–128.

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.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics 4*, 1–1.

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. (2000). *Comparative quantifiers*. Ph. D. thesis, Massachusetts Institute of Technology.

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

Han, C., J. Musolino, and J. Lidz (2016, jan). Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences 113*(4), 942–947.

Heim, I. (2000). Degree operators and scope. In *Semantics and linguistic theory*, Volume 10, pp. 40–64.

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. (1972). *On the semantic properties of logical operators in English*. Ph. D. thesis, University of California, Los Angeles.

Horn, L. R. (2006). The border wars: A neo-gricean perspective. In K. von Heusinger and K. Turner (Eds.), *Where semantics meets pragmatics*, pp. 21–48. Oxford: Elsevier.

Jacob, S. N., D. Vallentin, and A. Nieder (2012). Relating magnitudes: the brain's code for proportions. *Trends in cognitive sciences 16*(3), 157–166.

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.

Just, M. A. and P. A. Carpenter (1971). Comprehension of negation with quantification. *Journal of verbal learning and verbal behavior 10*(3), 244–253.

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. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. Routledge.

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.

Matthews, P. G. and M. R. Lewis (2016, oct). Fractions we cannot ignore: The nonsymbolic ratio congruity effect. *Cognitive Science 41*(6), 1656–1674.

McCrink, K., P. Shafto, and H. Barth (2017, apr). The relationship between non-symbolic multiplication and division in childhood. *Quarterly Journal of Experimental Psychology 70*(4), 686–702.

McCrink, K. and E. S. Spelke (2010, aug). Core multiplication in childhood. *Cognition 116*(2), 204–216.

McCrink, K. and E. S. Spelke (2016, feb). Non-symbolic division in childhood. *Journal of Experimental Child Psychology 142*, 66–82.

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

Moyer, R. S. and T. K. Landauer (1967). Time required for judgements of numerical inequality. *Nature 215*(5109), 1519–1520.

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

Qing, C. and M. Franke (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and linguistic theory*, Volume 24, pp. 23–41.

Qing, C. and M. Franke (2014b). Meaning and use of gradable adjectives: Formal modeling meets empirical data. In *Proceedings of the annual meeting of the cognitive science society*, Volume 36.

Qu, C., E. Szkudlarek, and E. M. Brannon (2021, jul). Approximate multiplication in young children prior to multiplication instruction. *Journal of Experimental Child Psychology 207*, 105116.

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*.

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*.

Sauerland, U. (2004). Embedded Implicatures. *Journal of Cognitive Science 5*(1).

Schöller, A. and M. Franke (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many. *Linguistics Vanguard 1*(open-issue).

Schulz, K. and R. Van Rooij (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and philosophy 29*(2), 205–250.

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.

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.

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.

Van Rooij, R. and K. Schulz (2004). Exhaustive interpretation of complex sentences. *Journal of logic, language and information 13*(4), 491–519.

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.