COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13234

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13234

Uncovering the Structure of Semantic Representations Using a Computational Model of Decision-Making

Sonia Ramotowska,[a] Shane Steinert-Threlkeld,[b] Leendert van Maanen,[c] Jakub Szymanik[d]

[a]*Institute für Linguistik, Heinrich Heine Universität Düsseldorf*
[b]*Department of Linguistics, University of Washington*
[c]*Department of Experimental Psychology, Utrecht University*
[d]*Center for Mind/Brain Sciences and Department of Information Engineering and Computer Science, University of Trento*

Received 9 July 2022; received in revised form 29 October 2022; accepted 13 December 2022

Abstract

According to logical theories of meaning, a meaning of an expression can be formalized and encoded in truth conditions. Vagueness of the language and individual differences between people are a challenge to incorporate into the meaning representations. In this paper, we propose a new approach to study truth-conditional representations of vague concepts. For a case study, we selected two natural language quantifiers *most* and *more than half*. We conducted two online experiments, each with 90 native English speakers. In the first experiment, we tested between-subjects variability in meaning representations. In the second experiment, we tested the stability of meaning representations over time by testing the same group of participants in two experimental sessions. In both experiments, participants performed the verification task. They verified a sentence with a quantifier (e.g., "*Most* of the gleerbs are feezda.") based on the numerical information provided in the second sentence, (e.g., "60% of the gleerbs are feezda"). To investigate between-subject and within-subject differences in meaning representations, we proposed an extended version of the Diffusion Decision Model with two parameters capturing truth conditions and vagueness. We fit the model to responses and reaction times data. In the first experiment, we found substantial between-subject differences in representations of *most* as reflected by the variability in the truth conditions. Moreover, we found that the verification of *most* is proportion-dependent as reflected in the reaction time effect and model parameter. In the second experiment, we

Correspondence should be sent to Sonia Ramotowska, Institute für Linguistik, Heinrich Heine Universität Düsseldorf, Universitätstr. 1, 40225, Düsseldorf, Germany. E-mail: sonia.ramotowska@hhu.de

showed that quantifier representations are stable over time as reflected in stable model parameters across two experimental sessions. These findings challenge semantic theories that assume the truth-conditional equivalence of *most* and *more than half* and contribute to the representational theory of vague concepts. The current study presents a promising approach to study semantic representations, which can have a wide application in experimental linguistics.

Keywords: Quantifiers; Computational modeling; Semantics; Decision-making; Individual difference; Diffusion Decision Model

1. Introduction

Human language is an exceptionally complex phenomenon. The meaning of words and sentences has been studied from a variety of perspectives, ranging from formal semantics descriptions through computational simulations to experimental semantics (e.g., Carcassi & Szymanik, 2021; Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; Frank & Goodman, 2012; Hackl, 2009; Katsos et al., 2016; Pietroski, Lidz, Hunter, & Halberda, 2009). Formal semantics analyzes linguistic meaning at an abstract level and studies the logical form and composition of meanings. One of the challenges to formal semantics are cases when the meaning boundaries cannot be precisely defined. This is often referred to as vagueness of meaning. Vagueness is omnipresent in natural language, for example, in the formation of categories (e.g., Gruenenfelder, 2019; Verheyen & Storms, 2013), meaning of adjectives (see Solt, 2015, for review), and quantifiers (e.g., Denić & Szymanik, 2022; Solt, 2011).

Vague concepts are challenging to formalize because they apply to a given object only to a certain degree. Several attempts have been made to formalize vague concepts (e.g., Glöckner, 2006; Gruenenfelder, 2019; Kamp & Partee, 1995). One of the adopted strategies was to enrich the classical two-valued logic (e.g., Kamp & Partee, 1995) by assigning undefined truth value to vague concepts. Another (e.g., Glöckner, 2006) was to reject the classical logic completely and use fuzzy logic (Zadeh, 1983) instead. In this paper, we develop a proposal that is agnostic to specific logical framework. Moreover, our approach completes the existing proposals by considering one of the consequences of vagueness—the individual differences in meaning representations (Verheyen & Storms, 2013). When the members of the categories belong to them only to a certain degree, different language users can disagree about the classification criteria. For example, in the domain of categorization, swimming indisputably belongs to a category of sports, while sleeping does not. In contrast, playing chess or hiking may raise a controversy. In the domain of quantifiers, one can say "*many* first-year bachelor students came to the Introduction to Linguistics lecture" if 180 out of 200 students came and *not many*, when only five out of 200 students came. The group of 100 students, in turn, is more difficult to classify as *many* or *not many*.

The goal of the current paper is to propose a modeling approach to meaning representations of vague concepts. The psychologically realistic model of linguistic behavior must account for linguistic factors (e.g., the effect of context, vagueness, ambiguity, etc.) and individual differences in meaning representations. While semantic theories usually elaborate on the former aspect, they widely ignore the latter. The interpretation of the noisy linguistic behavioral

data (e.g., accuracy or reaction times) in the context of nuanced predictions of the semantics theories is often ambiguous. The problem of linking theory and data is a common problem in cognitive science (Guest & Martin, 2021). Computational modeling can be a solution because it helps to formalize our theoretical intuitions and make the theories transparent and testable.

Among the existing models of linguistic behavior, the Rational Speech Act (Frank & Goodman, 2012) has gained increasing attention. The Rational Speech Act model is a Bayesian model of communication between speaker and listener. It is particularly useful to explain *what* kind of linguistic choices speakers make (Martin, 2016). Martin (2016) argued that the Bayesian approach belongs to the *statistical what camp*. They explain the distribution of choices, but they do not explain how the specific choice was made. In this paper, we take a different perspective and focus on the processing aspects of vague concepts. We argue that linguistic behavior can be studied using computational models of decision-making.

To model processing aspects of linguistic behavior, we applied a formal model of decision-making to data from a quantifier verification task. This task is a commonly used paradigm to test meaning representations (Bott, Augurzky, Sternefeld, & Ulrich, 2017; Clark, 1976; Deschamps et al., 2015; Just & Carpenter, 1971; Szymanik, 2016; Zajenkowski & Szymanik, 2013). In this type of task, participants have to indicate whether a sentence is "true" or "false" given an example scenario. Participants base their judgment on the evidence provided (e.g., a sentence or picture). Verification is a process of collecting evidence in favor of all decision options. Once the individual has collected sufficient evidence for one of the decision options (e.g., the "true" or "false" answer), the decision can be made. Based on participants' responses, we can reverse engineer the truth conditions they assigned to a specific quantifier. Based on the choices and reaction time data, we can also infer how participants made their decisions. Following this observation, we show that the performance in the verification task can be analyzed using computational models of decision-making such as the Diffusion Decision Model (DDM; Ratcliff, 1978), which is widely used in many cognitive domains outside of language (Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016). We illustrate how this computational model allows us to disambiguate the role various aspects of meaning play in the subjects' behavior. Our approach applies to many psycholinguistic tasks including, but not limited to, lexical decision, inferences, and verification.

Vagueness and individual differences are especially challenging to formal theories specifying the meanings of logical words, such as quantifiers (e.g., *most*, *many*, *more than half*). The long tradition of formal semantics (Barwise & Cooper, 1981; Mostowski, 1957) specifies the meaning of quantifiers in a form of rigid truth conditions. However, some natural language quantifiers such as *many* and *few* and also *most* are vague (Denić & Szymanik, 2022; Solt, 2011, 2016). Moreover, one study (Yildirim, Degen, Tanenhaus, & Jaeger, 2016) has shown individual differences in meaning representations of quantifiers, namely, that speakers apply *many* and *some* to describe different quantities. As a case study, we test the differences between two natural language quantifiers: *most* and *more than half*. We chose these quantifiers because they have drawn much attention from semanticists in recent years (Coppock & Strand, 2019; Hackl, 2009; Kotek, Sudo, & Hackl, 2015; Lidz, Pietroski, Halberda, & Hunter, 2011; Pietroski et al., 2009; Register, Mollica, & Piantadosi, 2020; Solt, 2016). Traditionally, they are treated as truth-conditionally equivalent; however, the behavioral (e.g.,

Denić & Szymanik, 2022; Hackl, 2009; Ramotowska, Steinert-Threlkeld, van Maanen, & Szymanik, 2020) and corpus data (e.g., Solt, 2016) showed a number of differences between them. We show that our modeling strategy and experimental design allow us to disentangle different aspects of meaning, such as semantic representations, vagueness, pragmatics, or processing, usually confounded in experimental and corpus data. In this way, we exhibit a systematic approach to infer the meaning of vague concepts from linguistic data and validate the semantic theories.

Although in this paper we focus mostly on *most* and *more than half*, we will show that our approach can be generalized beyond these two examples by also analyzing data from three control quantifiers *fewer than half*, *few*, and *many* and investigating the differences between quantifiers in all DDM parameters. Moreover, we believe that our approach can have a wide application in the analysis of vague concepts in general. It can be generalized to other vague concepts because we treat the verification of quantifiers as a decision task, and in an analogous way, we could treat the categorization of other vague concepts as a decision task as well.

The current paper has the following structure. In the next sections of the introduction, we will present the linguistic discussion about *most* and *more than half*, introduce the DDM, and our experimental design. Next, we will propose an extended version of the DDM with incorporated meaning representations of quantifiers and vagueness. We will present results from two behavioral experiments and show how our modeling approach can test semantic differences between *most* and *more than half*. Finally, in the discussion section, we will show how our approach can contribute to the development of the theories of vague concepts. We will discuss our findings in the context of existing theoretical proposals and show the challenges to these approaches.

Semantic representations: The case of most *and* more than half

The meaning of quantifiers such as *most* and *more than half* can be expressed on the ground of logical theories of meaning, for example, the Generalized Quantifier Theory (Barwise & Cooper, 1981; Mostowski, 1957), in the form of truth conditions. For an illustration consider the following example. The sentence "*Most/More than half* of the students passed the exam" is true under one of two conditions: (1) if the number of students, who passed the exam ($|[\![\text{Students}]\!] \cap [\![\text{passed exam}]\!]|$) is greater than half of all students ($\frac{1}{2}|[\![\text{Students}]\!]|$), or (2) if the number of students, who passed the exam is greater than the number of students, who did not pass the exam ($|[\![\text{did not pass exam}]\!]|$). Following Hackl's (2009) linguistic analysis, we can formulate example truth conditions for *more than half* and *most*:

1. *More than half* of the students passed the exam. $\leftrightarrow |[\![\text{Students}]\!] \cap [\![\text{passed exam}]\!]| > \frac{1}{2}|[\![\text{Students}]\!]|$
2. *Most* of the students passed the exam $\leftrightarrow |[\![\text{Students}]\!] \cap [\![\text{passed exam}]\!]| > |[\![\text{Students}]\!] \cap |[\![\text{did not pass exam}]\!]|$

A brief reflection on these two examples leads to the conclusion that the truth conditions of *most* and *more than half* are logically equivalent. The truth value of the sentences with *most* or *more than half* will be the same in every situation. For example, both sentences will be

false if nine out of 20 students passed or true if 11 out of 20 students passed. Therefore, the logical theories of meaning predict that these quantifiers are used interchangeably.

Contrary to this claim, there is ample evidence that *most* and *more than half* lead to different linguistic behavior. These differences have been attributed to verification (Hackl, 2009) or pragmatics (Solt, 2016). Hackl (2009) postulated that both *most* and *more than half* are associated with different verification strategies, even though they are logically equivalent. The expression *most* (Example 2) is equivalent to the expression *more than half* (Example 1) in terms of truth conditions, but they have different linguistic representations, which yield different verification strategies. A verification strategy is an algorithm to determine the truth value of the sentence. The algorithm to verify *more than half(A, B)* requires the computation of the intersection of sets A and B (e.g., students who passed the exam) and half of the set of all As (e.g., half of all students). The algorithm for *most(A, B)*, in turn, is a vote-counting strategy, which involves tracking whether the number of As that are B is greater than the number of As that are not B. These two algorithms correspond to different cognitive processes and lead to different behaviors.

The idea that *most* and *more than half* differ in processing was further developed by associating *most* with the approximate number system (Lidz et al., 2011; Pietroski et al., 2009; Solt, 2016) and *more than half* with the precise number system (Solt, 2016). In the approximate number system, the representations of numbers overlap on the number mental line (Dehaene, 2007). Consequently, the numerocities close to each other are difficult to distinguish. For example, 49 is difficult to distinguish from 51, but 10 is easy to distinguish from 90. Because *most* is processed using the approximate number system, its verification is difficult when the to-be-compared proportions are close to each other on the number mental line. This means that participants may be unsure about the true value of the sentence "*Most* of the As are B" if the proportion of As that is B is close to half.

The difference in the processing of *most* and *more than half* has two consequences. The first one is related to the pragmatics of *most*, namely, it is used to describe proportions that are significantly different (Kotek et al., 2015; Solt, 2016). For example, a corpus study (Solt, 2016) showed that *most* is not suitable to use in some contexts when the proportion is close to 50% (e.g., "*Most* of the American population is female" vs. "*More than half* of the American population is female"; Solt, 2016, p. 67). Furthermore, Carcassi and Szymanik (2021) support a pragmatic explanation by embedding it into the computational framework of the Rational Speech Act (Frank & Goodman, 2012).

The second consequence is that *most* is a vague quantifier (cf. Denić & Szymanik, 2022; Solt, 2011, 2016), while *more than half* has a sharp meaning boundary. Vagueness can relate to a threshold (Solt, 2015), a cut-off point (e.g., proportion) for which the responses change the truth value. Solt (2011) postulated that *most* is a vague quantifier with a fuzzy threshold, while *more than half* has sharp meaning boundaries and a 50% threshold. There are individual differences in thresholds with vague quantifiers (Ramotowska et al., 2020). Specifically, quantifiers such as *many* or *few* have varying usage depending on the speaker (Yildirim et al., 2016) and a flexible meaning, which can be adjusted to another speaker's usage (Yildirim et al., 2016) or learning criterion (Heim et al., 2015), and can be changed during the adapta-

tion process (Heim, Peiseler, & Bekemeier, 2020). We predict that if *most* is a vague quantifier, then we should observe individual differences in its representation.

To summarize, the proposed explanations of differences between *most* and *more than half* fall into two main categories: the processing explanation (e.g., differences in verification strategies) and the pragmatic explanation (e.g., the context in which *most* is used). In this paper, we show that while controlling for processing and pragmatic effects by means of a computational model and a specific experimental design, we can still observe different linguistic behaviors related to *most* and *more than half*. We argue that *most* and *more than half* in fact have different truth conditions. The processing and pragmatic explanations assume that *most* and *more than half* have equivalent truth conditions and that other factors cause different linguistic behaviors for these quantifiers. However, the previous studies that argued in favor of these explanations (e.g., Hackl, 2009; Kotek et al., 2015; Solt, 2016) did not explicitly test the truth-conditional equivalence of *most* and *more than half*.

The logical theories give a clear prediction about participants' behavior concerning *most* and *more than half*. Because we control for both processing and pragmatic factors, participants should apply the same truth conditions to classify the given proportion as *most* or *more than half*. Therefore, there should not be any variation between-participants or within-participants. On the other hand, if the meaning of *most* is vague, we could expect the between-participants or within-participants variation in representations of *most*. More specifically, if *most* is a vague quantifier, we can expect participants to differ in thresholds for *most* and to possibly change their truth conditions over time. Moreover, we also test an alternative hypothesis that *most* has the same truth conditions as *more than half*, but participants hesitate about the exact threshold for *most*. This would lead to observed uncertainty around thresholds rather than a shift in truth conditions. We operationalized uncertainty as longer reaction times around the threshold. We will show that we can test these hypotheses independently by mapping truth-conditional meaning and vagueness onto different model parameters.

1.1. Modeling, experiments, and predictions

In the previous section, we identified possible sources of differences between *most* and *more than half*. We argued that the linguistic discussion around these two quantifiers cannot be decided based on existing experimental (e.g., Hackl, 2009; Kotek et al., 2015; Pietroski, Lidz, Hunter, & Halberda, 2009) and corpus data (e.g., Solt, 2016). The goal of the current paper is to model linguistic behavior in a decision-making framework. For our purposes, we chose the DDM.

1.1.1. DDM

The DDM (Ratcliff, 1978) is a canonical evidence accumulation model (Ratcliff & Smith, 2004; Ratcliff et al., 2016). This class of models assumes that decisions are the outcome of a gradual accrual of evidence for the various choice options at hand. That is, over time, evidence accumulates until one of the options surpasses a critical amount of evidence, after which a decision-maker is assumed to commit to that option.

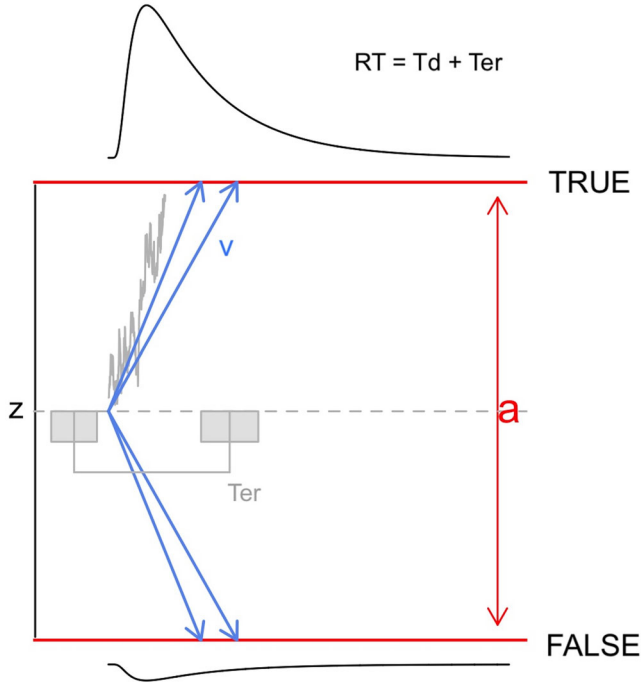

Fig. 1. Representation of the Diffusion Decision Model. The accumulation process (the gray zigzag line) starts in point z (starting point between the decision boundaries) and finishes when one of the decision boundaries (TRUE/FALSE response indicated by the red horizontal lines) is reached. The parameter a indicates the distance between decision boundaries (the red, vertical, two-sided arrow). T_{er} represents non-decision time needed to encode the stimuli (the first gray rectangle) and to provide a response (the second gray rectangle). v represents the drift rate. In the model proposed here, the drift rate depends on features of the task (blue arrows), predicting faster decisions (higher slopes) as well as slower decisions (lower slopes). Reaction times (RT) are the sum of non-decision time (T_{er}) and decision time (T_d).

The evidence accumulation models are routinely applied in the field of decision-making (e.g., Donkin & van Maanen, 2014; Miletić & van Maanen, 2019; van Maanen et al., 2012). The *choice alternatives* should be understood here as any set of mental representations that one can select from. The DDM can be used to jointly investigate the proportion of decision outcomes and reaction times in the tasks in which participants have to choose between two options (e.g., they have to assess whether the sentence is "true" or "false"). These possible decision outcomes (in Fig. 1 options TRUE and FALSE) are represented as two boundaries.[1] The decision-making process itself can be thought of as a noisy process of evidence accumulation toward one of the boundaries (one example decision process is shown in Fig. 1, indicated by the gray zigzag line). The model assumes that a decision is reached as soon as one of the boundaries is crossed, with the time required to reach the boundary called the decision time (Fig. 1, T_d). The time before and after the decision is called a non-decision time (Fig. 1, T_{er}), and it includes the time to encode the stimuli and execute the response after the decision is made. The reaction times are a sum of decision time and non-decision time.

Because of the noisy accumulation process, the DDM predicts that decision times vary from trial-to-trial, and consequently it models the full reaction time distribution. Given the specific shape of the reaction time distribution, the proportion of responses on each of the boundaries, and the relation between reaction times and choice, the DDM can be used to estimate a set of parameters that best explains the observed data (Voss, Nagler, & Lerche, 2013).

The parameters that specify the DDM are typically found to match specific cognitive processing components (Mulder, van Maanen, & Forstmann, 2014). In this way, various sources of variability of behavior can be separated. The parameter that expresses the speed of the accumulation process is called drift rate (v). It can be considered as a general measure of task performance, task difficulty, cognitive ability, or the speed of information processing (for an introduction to the DDM, see Voss et al., 2013). The drift rate can reflect multiple cognitive processes. For example, it could correspond to retrieval of the threshold from memory (Ratcliff, 1978), comparison between threshold and proportion given in the second sentence (Ratcliff & McKoon, 2018; 2020), or lexical decision (Ratcliff, Thapar, & McKoon, 2010). The evidence accumulation process starts at one point (z) and finishes when enough evidence is accumulated toward one of the decision boundaries. The decision boundaries are separated by parameter a. Moreover, two parameters, a decision time (T_d) and a non-decision time (T_{er}), model the duration of reaction times. Additionally, the model allows for the trial-to-trial variability in the starting point (s_z), drift rate (s_v) and non-decision time (s_{Ter}).

The DDM was successfully applied to model cognitive processes in two domains that are relevant to the current study: the domain of number cognition (Ratcliff & McKoon, 2018, 2020; Ratcliff, Thompson, & McKoon, 2015) and the domain of individual differences in linguistics tasks (Pexman & Yap, 2018; Ratcliff et al., 2010; Yap, Balota, Sibley, & Ratcliff, 2012). For example, Ratcliff et al. (2010) found that participants with a lower IQ (Intelligence quotient) score had lower drift rates in a lexical decision task than participants with a higher IQ. Pexman and Yap (2018) fitted DDM to data from a categorization task (concrete vs. abstract words) to test the individual differences in semantic processing and decision-making. They found differences in drift rates between participants with high versus low vocabulary knowledge. Yap et al. (2012) found an association between DDM parameters (drift rate, boundaries separation, and non-decision time), performance in a lexical decision task, and individual differences in vocabulary knowledge. In the domain of number cognition, Ratcliff et al. (2015) showed that individual differences in accuracy or reaction times in numeracy tasks are explained by different parameters.

To summarize, evidence accumulation models such as DDM seem to be an excellent tool to analyze response proportion and reaction time distribution data in linguistic and number cognition tasks. Moreover, the analysis of the participants' parameters gives a meaningful interpretation of individual differences between-subjects. Finally, the flexibility of the evidence accumulation model provides the opportunity to adapt them to different tasks. For these reasons, DDM also seems ideally suited to the analysis of semantic data as we will illustrate below.

1.1.2. Current experiment

To test the equivalence of the truth conditions of *most* and *more than half*, we chose a specific experimental paradigm in which we limited pragmatic and processing differences between these quantifiers. In our task, participants verified a sentence with *most* and *more than half* against the sentence with a proportion given as a number, for example, 55%. In Experiment 1, we tested whether the variation of individual thresholds differed between quantifiers. In Experiment 2, we tested the stability of individual thresholds over a 2-week period.

Previous studies in the domain of categorization (e.g., Verheyen, White, & Égré, 2019) have shown that participants change categorization criteria over time. Experiment 2 intends to investigate if similar effects can be observed for functional words such as quantifiers. This is the first study that tests the stability of semantic representations of functional words.

We tested a prediction that follows directly from logical theories. Logical Theories hypothesis: *Most* will have the same threshold (50%) as *more than half* for all participants, and the threshold for both quantifiers will be stable over time.

Moreover, we considered the effect of vagueness. Following the hypothesis that the meaning of *most* is vague (Denić & Szymanik, 2022; Solt, 2011), we formulated a hypothesis complementary to the Logical Theories hypothesis. Vagueness hypothesis: *Most* will have a greater variation in individual thresholds than *more than half*, meaning that some participants will have a higher threshold for *most* than *more than half*, and some participants will have a 50% threshold for both quantifiers. With regard to the stability of thresholds over time, we considered that *more than half* should have a stable threshold, and we could observe the variation in the thresholds for *most* over time.

Vagueness can result in different thresholds for *most*, but it could also lead to uncertainty about the threshold. Because thresholds can vary, participants may be uncertain about their decision when they verify *most* against the proportions close to the threshold. The uncertainty and difficulty of the decision would be reflected in longer reaction times. Therefore, we predict that under the Uncertainty hypothesis: The speed of verification of *most* but not *more than half*, will be proportion-dependent, meaning that the reaction times for proportions close to the individual threshold will be longer than for proportions further from the threshold. Moreover, we tested whether the variability in thresholds and vagueness interact. We considered the logical possibility that (1) *most* and *more than half* can have the same 50% threshold, and yet *most* is vague, and participants have longer reaction times around 50% proportion; (2) *most* can have greater variation in thresholds than *more than half* and participants have longer reaction times around the individual threshold; (3) or *most* can have greater variation in thresholds and the uncertainty about the threshold depends on the threshold choice. This analysis was exploratory in nature.

In addition, we investigated other model parameters and quantifiers. We predicted that also vagueness of *many* and *few* should be reflected in model parameters. Following Schlotterbeck, Ramotowska, van Maanen, and Szymanik (2020), we expected to replicate differences in non-decision time and drift rate for negative versus positive quantifiers (see the Appendix).

2. Experiment 1[2]

2.1. Methods

2.1.1. Participants

We tested 90 users of the Amazon Mechanical Turk platform (https://www.mturk.com/). We included 72 participants (48 male) in our analysis, age: $M = 35$, $SD = 11$, range: 22–59. The sample represented various educational backgrounds: high school graduates (24 participants), high school graduates who started college (22 participants), and college graduates (26 participants). The subjects received USD 4 in compensation for their participation. The study was approved by the Ethics Committee at the University of Amsterdam.

2.1.2. Exclusion criteria

We applied two exclusion criteria. First, we excluded fast-guessing participants (11 subjects), whose reaction times were faster than 300 ms for 50% or more responses. Additionally, we tested whether participants respected quantifier monotonicity, that is, for the quantifiers *most*, *more than half*, and *many*, we expected the increasing probability of saying "true" with increasing proportion. The opposite effect was expected for *few* and *fewer than half*. We tested this assumption by estimating random slopes for proportion for each participant (*glmer* function in R package *lemrTest*; Kuznetsova, Brockhoff, & Christensen, 2017). For positive quantifiers, we excluded participants with negative slopes, and for negative quantifiers, we excluded those with positive slopes. We excluded a total of six participants based on this criterion. Finally, we excluded one participant who had participated in a similar experiment previously.

2.1.3. Design

Participants verified a sentence with a quantifier of the form "Q of the As are B," where Q was one of the quantifiers: *most*, *more than half*, *fewer than half*, *few*, and *many* (within-subject) against the sentence with a proportion ("p% of the As are B") given as a number, for example, 55%. Because quantifiers differ in a number of properties (Katsos et al., 2016), we included *many*, *few*, and *fewer than half* for control and generalization purposes.

We decided to use pseudowords in order not to introduce any other variability in meaning beyond different quantifiers. We used *Wuggy* software (Keuleers & Brysbaert, 2010) to generate pseudowords from English nouns (As) and adjectives (Bs). We selected 50 pseudo-adjectives and 50 pseudo-nouns, which a native English speaker indicated sounded like plausible English words. We controlled for the frequency of the original English words. The words on both final lists had a *Zipf* value of 4.06 (SUBRLEX-US database, van Heuven, Mandera, Keuleers, & Brysbaert, 2014). All words on the final lists were six-letter words. We matched each quantifier with each pair of pseudowords. We presented the trials in random order.

To ensure that we studied the semantic properties of *most* and *more than half*, we made two important choices regarding experimental design. First, we decided to eliminate the potential differences in processing and verification strategies between the two quantifiers. We chose the presentation of proportion as a number, and we did not limit the time to make a decision in

order to enforce the precise processing of the numerosities. In this way, the results obtained cannot be attributed to the usage of the approximate number system (Dehaene, 1997) during the verification of *most*, compared to *more than half*. We also used pseudowords for As and Bs to limit pragmatic inferences (van Heuven et al., 2014). Therefore, our participants could only access the meaning of the quantifiers and proportions.

2.1.4. Procedure

Participants saw two sentences on separate screens. On the first screen, they saw a sentence with a quantifier and pseudowords A and B: "{*most/more than half/many/few/fewer than half*} of the As are B." To display the first sentence, participants had to press the down arrow key and keep it pressed as long as they wanted to read the sentence. When they released the down arrow key, the sentence disappeared. To display the second sentence, they had to press the down arrow key again (but they did not have to keep it pressed). On the second screen, participants saw a sentence with a proportion given as a percentage: "*p*% of the As are B," where As and B were the same pseudowords as in the first sentence. The proportion, *p*%, in the second sentence was randomly drawn from 1% to 99%, excluding 50%. On this screen, participants had to decide whether the first sentence was true or false based on the information from the second sentence by pressing the left or right arrow key (counterbalance between-subjects). We presented first the sentence with a quantifier and second the sentence with a proportion because quantifiers had different lengths. This factor could affect the reading times of sentences with quantifiers and, therefore, the reaction times and the estimation of DDM parameters.

We randomly chose 25 proportions above and 25 proportions below 50% for *more than half*, *fewer than half*, and *most* for each participant. Altogether, participants saw 250 trials, 50 trials for each quantifier. The experiment was preceded by a short training block (eight trials). In the training block, participants saw sentences with the quantifiers *some, all*, and *none*. At the end of the experiment, participants filled in a short demographic survey.

2.1.5. Preprocessing reaction times data

Apart from excluding participants, we also excluded reaction times that were too short or too long. Before we fitted the DDM, we excluded reaction times shorter than 300 ms and longer than the group mean + 2*SD* for each quantifier and each response type (true, false) separately. In this way, we excluded 2.7% of the trials.

2.2. Computational modeling of the verification process

2.2.1. Extended DDM

We fitted the simple DDM (without variability parameters s_z, s_v, and s_{Ter}) to reaction times and responses data for each participant. It is important to estimate parameters on an individual level for multiple reasons. The first obvious reason is that individuals might differ in the parameter estimates, yielding a group estimation procedure difficult to interpret. The second reason is that because of averaging over individuals, it might appear as if a concept has a vague

decision threshold, whereas in reality, it has a precise threshold, the location of which varies across participants. Modeling individuals allows us to disentangle these possible scenarios.

We did not include variability parameters in order to simplify the model and because we did not expect them to differ between quantifiers. We assumed that the differences in threshold and uncertainty between *most* and *more than half* would manifest during the verification process in the drift rate. We conceptualized the threshold as a proportion for which the drift rate is zero. A drift rate of zero indicates an indecision. This means that participants are equally likely to choose either option (e.g., "true" or "false"). At this point, a decision boundary can only be reached because of random fluctuations in the accumulation process and not because of a drift toward one of the boundaries. For positive quantifiers (*most*, *more than half*, and *many*), below the threshold, the drift rate will be negative, making a response on the lower boundary (representing "false") more likely; above the threshold, the drift rate will be positive, making a response on the upper boundary (representing "true") more likely. For negative quantifiers (*few* and *fewer than half*), the relationship between the drift rate direction (positive vs. negative) and response choice ("true" vs. "false") is reversed. The speed with which the drift rate changes from 0 to its maximum value represents the uncertainty between the response options. If this is low, then the drift rate function is essentially a step function, but if the uncertainty is large, the drift rate will remain about zero for proportions around the threshold (Fig. 2). We operationalized these concepts via the generalized logistic function presented in Fig. 2. The threshold is the midpoint of the logistic function (parameter p_0), and the uncertainty is the growth rate parameter of the logistic function (parameter s). Additionally, the logistic function defines a maximum drift rate (V_U) and a minimum (V_L).

2.2.2. Bayesian model averaging

Because one of the goals of our study was to capture individual differences between participants, we considered that there might also be individual differences between participants in terms of which model is best according to Akaike Information Criterion (AIC) values (Akaike, 1998). Therefore, we decided to use Bayesian model averaging for all DDM parameters (Hoeting, Madigan, Raftery, & Volinsky, 1999; Miletić & van Maanen, 2019; Wagenmakers & Farrell, 2004) rather than parameters from the winning model. Bayesian model averaging is a method to compute parameters for each participant, taking into account the weighted average of the parameters from each model.

The weight for model i (w_iAIC) is defined using the AIC values (Wagenmakers & Farrell, 2004):

$$w_iAIC = \frac{e^{-\frac{1}{2}\Delta_i(AIC)}}{\sum_{k=1}^{K} e^{-\frac{1}{2}\Delta_k(AIC)}} \quad (1)$$

where $\Delta_i\,(AIC) = AIC_i$- min(AIC) for each model i.

2.3. Mixed-effects regression modeling

For all linear mixed-effects regression models, we applied the individual Bayesian model averaged thresholds to each participant's response data from Experiment 1. For positive

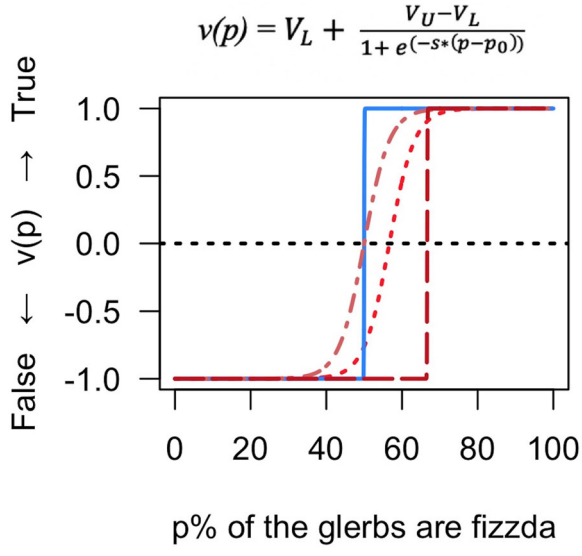

Fig. 2. Drift rate structure is given as generalized logistic function equation, where V_L is the lower asymptote of the logistic function (minimum drift rate), V_U is the upper asymptote of the logistic function (maximum drift rate), s is the growth rate the logistic function (uncertainty), p is the proportion in the second sentence of each trial, and p_0 is midpoint the logistic function (threshold). On the x-axis are shown the proportions that participants saw during the experiment. On the y-axis are the values of the drift rate $v(p)$. The proportion on the x-axis that corresponds to the drift rate equals zero is the threshold. The shape of the logistic lines indicates the uncertainty about the response around the threshold (sharper lines indicate greater certainty). The figure presents the predicted drift rate $v(p)$ for *more than half* (blue, solid line) and three possible drift rates $v(p)$ for *most*. The lightest red, dot-dashed line shows the threshold for *most* at 50% and greater uncertainty around the threshold, compared to *more than half*. The darkest red, dashed line shows the threshold higher than 50% for *most* and low uncertainty. The middle red, dotted line shows the combined effect: the threshold is higher than 50% for *most* and greater uncertainty around the threshold, compared to *more than half*.

quantifiers, we included "false" responses below the threshold and "true" responses above the threshold in the analyses (inversely for negative quantifiers). To test the Uncertainty hypothesis, we fitted a linear mixed-effects model (*lmer* function in *lmerTest* package in R; Kuznetsova et al., 2017) with log-transformed (log10) reaction times (in seconds) as the dependent variable and z-scored proportion, quantifier (*most*, *more than half*), response ("true" or "false"), and their interactions as predictors. We used the "true" response and the quantifier *most* as the baseline. For exploratory analysis, we fitted a linear mixed-effects models (*lmer* function in *lmerTest* package in R; Kuznetsova et al., 2017) with log-transformed (log10) reaction times (in seconds) as a dependent variable and distance from threshold (proportion centered on individual threshold), individual threshold centered on mean threshold, response ("true" or "false"), and their interactions as predictors for each quantifier. We used the "true" response as the baseline. We used log-transformation of reaction times to improve the compatibility of mixed-effects models with their assumptions: normality and homoscedasticity of residuals.

Table 1
Mean reaction times (RT) in seconds (*SD*) and proportion of response "true" versus "false" in Experiment 1

	True Response		False Response	
Quantifier	RT	Response	RT	Response
Few	1.201 (0.133)	0.39	1.087 (0.134)	0.61
Fewer than half	1.170 (0.186)	0.48	1.064 (0.115)	0.52
Many	1.000 (0.117)	0.57	1.107 (0.122)	0.43
Most	1.044 (0.395)	0.47	1.038 (0.160)	0.53
More than half	0.917 (0.086)	0.50	0.942 (0.092)	0.50

We applied the same procedure of testing the random effect structure for all models. We tried to maximize the random structure of the model (Barr, Levy, Scheepers, & Tily, 2013). Therefore, we always included by-subject random intercept and by-subject random slopes if they improved the model (determined by *anova* function in R). We included by-item random intercept if the model was not overfit.

2.4. Results

2.4.1. Descriptive statistics

Table 1 summarizes the mean reaction times and proportion of "true" and "false" responses for each quantifier. This summary already suggests the differences between *most* and *more than half*. First, the reaction times for *more than half* are shorter than for *most*. Second, *most* has a greater proportion of "false" than "true" responses, indicating the possible difference in the threshold. We applied DDM to further explain these effects.

2.4.2. Model fit and comparison

We used R package *rtdists* to fit DDM, and we estimated the maximum likelihood of DDM parameters using particle swarm optimization (Clerc, 2006, 2010) on the seconds scale of reaction times. To identify model parameters, we fixed the diffusion coefficient, which indicates the standard deviation of diffusion process random noise, to a scaling constant of 0.1.

To test which DDM parameters were identical across conditions, we systematically varied the model constraints. We evaluated each constrained model by assessing AIC values (Akaike, 1998). We used the AIC values to compute the number of participants for whom each model was the best model (*n* best) or one of the three best models (*n* top 3). Based on the individual AIC values of each participant, we assigned ranks to each model (1 for the best model; 9 for the worst model). Next, we computed the *mean* rank for each model (the lower the rank, the better the model). Model comparison descriptive statistics are summarized in Table 2.

For some parameters (T_{er}, z, V_L, and V_U), we observed differences between positive versus negative quantifiers or, in the case of parameter s, between *more than half*/*fewer than half* and the rest of the quantifiers. We started the parameter estimation process with an unconstrained model in which all parameters could differ for all quantifiers (Model 1). We then chose T_{er} to be the same across negative quantifiers (*few* and *fewer than half*) and positive quantifiers

Table 2
Model comparison in Experiment 1 (k is the number of free parameters in the model; *Mean* is the mean rank; n best is the number of participants for whom the given model was the best; n top 3 is the number of participants for whom the given model was one of the three best models)

	Parameters			Rank		
Model	Free	Fixed	k	*Mean*	n best	n top 3
1	T_{er}, a, z, p_0, s, V_L, V_U		35	7.72	1	4
2	a, z, p_0, s, V_L, V_U	T_{er}	32	6.60	3	11
3	z, p_0, s, V_L, V_U	T_{er}, a	28	5.53	4	16
4	p_0, s, V_L, V_U	T_{er}, a, z	25	5.10	4	22
5	p_0, V_L, V_U	T_{er}, a, z, s	22	3.99	14	30
6	p_0,	T_{er}, a, z, s, V_L, V_U	16	4.14	10	31
7	p_0,	T_{er}, a, z, s, V_L, V_U	14	4.39	3	27
8		T_{er}, a, z, s, V_L, V_U, p_0	12	**3.38**	16	**41**
9		T_{er}, a, z, s, V_L, V_U, p_0	11	4.17	**17**	34

(*many*, *most*, *more than half*; Model 2). In Model 3, we constrained a parameter to be the same across all quantifiers. The z parameter was the same across negative quantifiers (*few* and *fewer than half*) and positive quantifiers (*many*, *most*, *more than half;* Model 4). We constrained parameter s to be the same for *fewer than half* and *more than half* and the same for *most*, *many*, and *few* (Model 5). We also constrained asymptotes parameters V_L and V_U (Model 6), in the same way as the T_{er} and z parameters. Finally, we tested the model with symmetric V_L and V_U parameters (Model 7). In Model 8, we constrained p_0 parameters for *more than half* and *fewer than half* and in Model 9 also for *most*.

In addition to model comparison, we also visually investigated the individual participant model fit and the aggregate Model 7 fit (Fig. 3) by means of the Vincentizing method (Ratcliff, 1979). Fig. 3 presents a cumulative probability of reaction time data separately for "true" and "false" responses. Each percentile of the data was scaled by the overall proportion of the "true" or "false" responses. The height of the cumulative distribution shows the proportion of "true" or "false" responses for each percentile. For example, the responses "true" for *few* make up 39% of all responses (see Table 1) and the 0.95 percentile of responses "true" has a cumulative probability of 0.37 as it covers 95% of "true" responses.

Finally, we used AIC values to compute weights for each model and weighted averaged parameters (Bayesian model averaging; Hoeting et al., 1999; Table 3 and Fig. 4). The large variation in best model fit (see Table 2) supports our choice to use Bayesian model averaged parameters. Fig. 5 shows the example drift rates for Model 7. The drift rates for *more than half* and *fewer than half* are steeper and have lower variability in thresholds (proportion for which $v(p) = 0$) than drift rates of other quantifiers.

2.4.3. Logical Theories and Vagueness hypotheses

We systematically constrained model parameters (see Table 2) and we compared the AIC (Hoeting et al., 1999) to determine which model had a better balance between goodness-of-fit and model complexity (Pitt & Myung, 2002). In order to test the Logical Theories and Vagueness hypotheses, we constrained parameter $p_0 = 50\%$ for *more than half*, *fewer than*

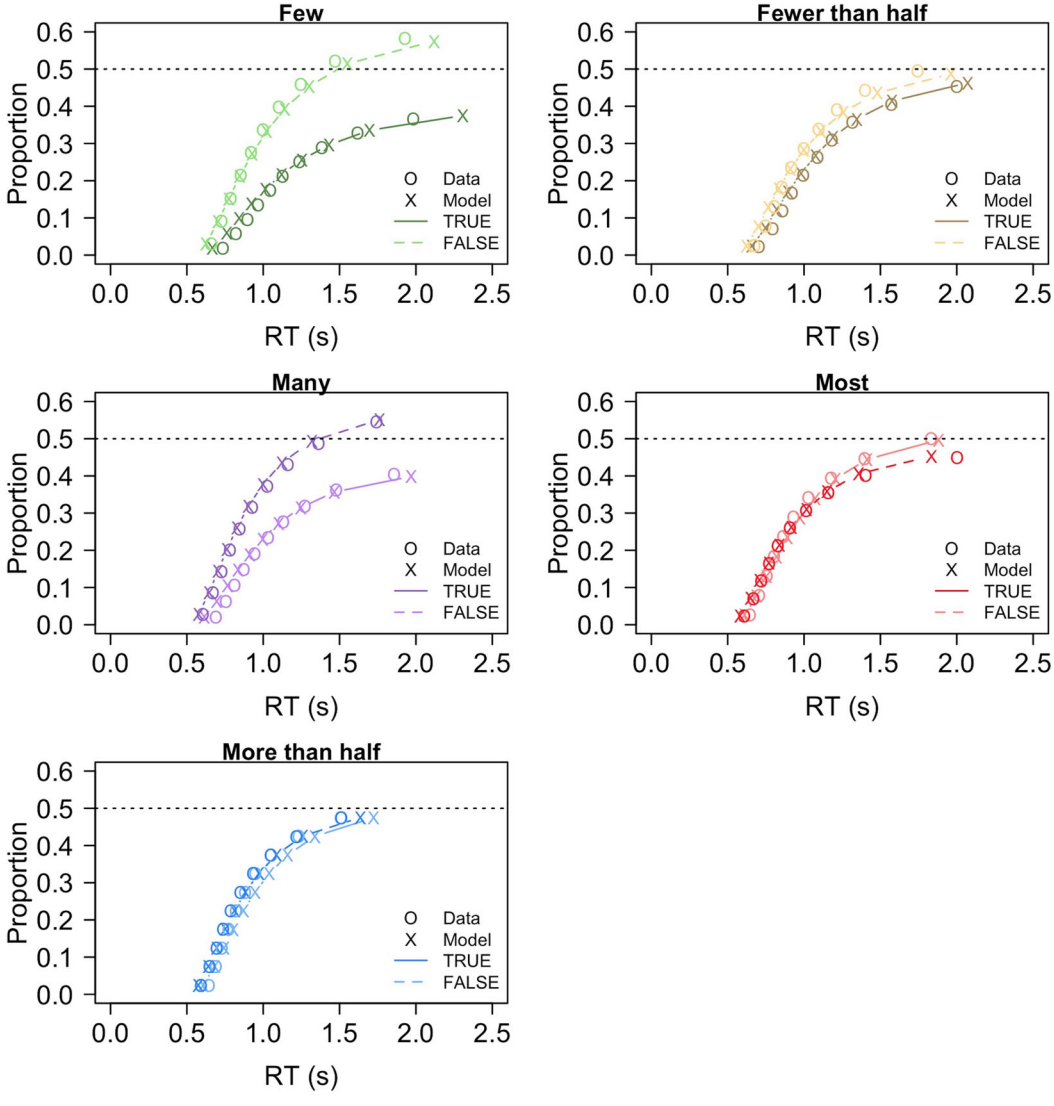

Fig. 3. Defective cumulative density plots show the average fit of Model 7. For this visualization, we plot the mean 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95% percentiles over participants, scaled by the proportion of "true" and "false" responses, separately for the data and Model 7 prediction.

half, and *most* and compared this model to the model with p_0 as a free parameter (Model 7, see Table 2). The Logical Theories hypothesis predicts that the constrained model will be preferred for all quantifiers, while the Vagueness hypothesis predicts this only for *more than half* and *fewer than half*.

In the first step, we constrained p_0 for *more than half* and *fewer than half*. We found that the constrained model was preferred over the model with p_0 as a free parameter by 57 out of 72

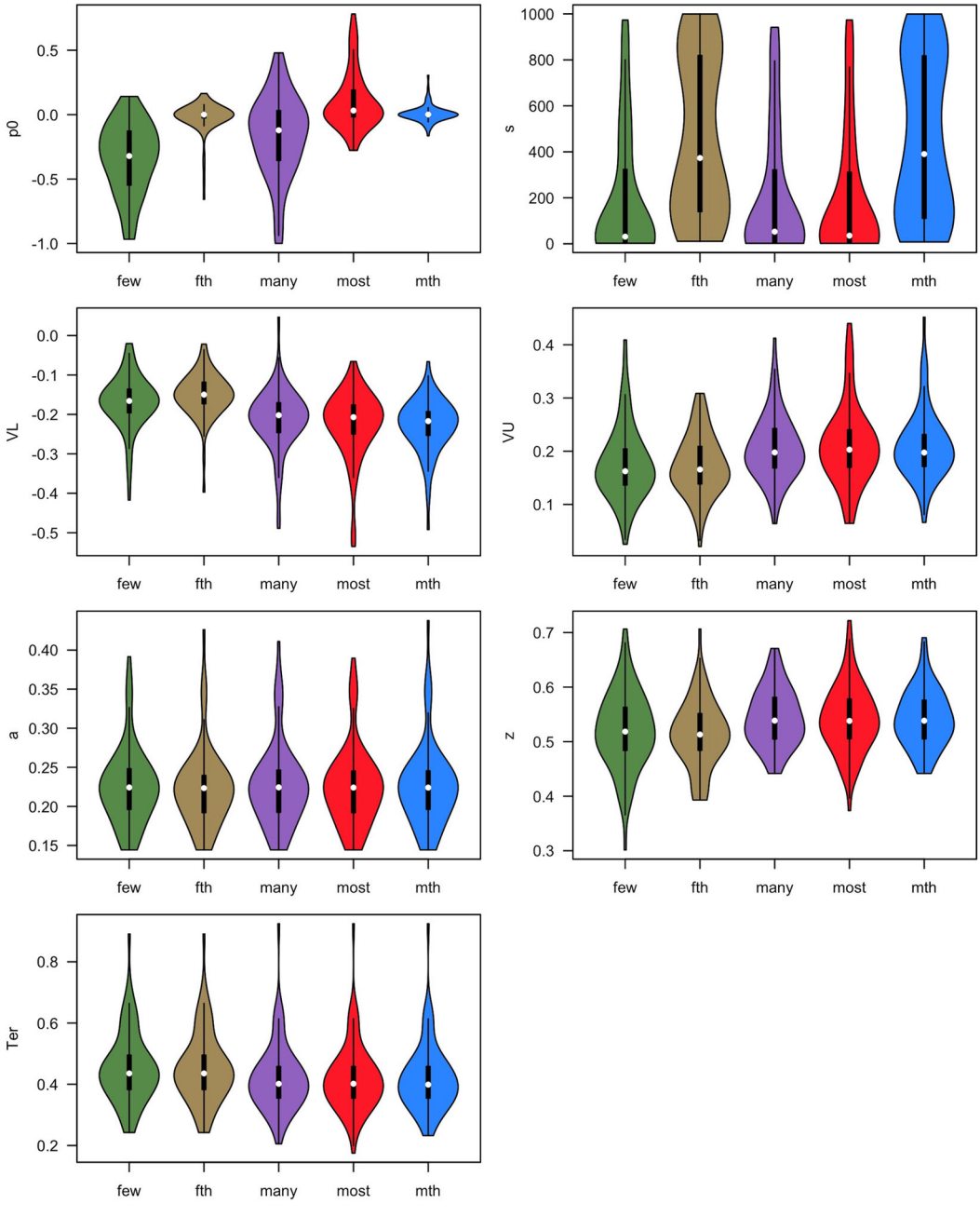

Fig. 4. Violin plots representing the distributions of Bayesian model averaged parameters (Model 1 to Model 9) for all five quantifiers (fth is *fewer than half* and mth is *more than half*).

Table 3
Summary for model mean (*SD*) parameters after Bayesian model averaged (Model 1 to Model 9) using Akaike Information Criterion

Quantifier	p_0	s	V_L	V_U	a	z	T_{er}
Few	−0.34	197	−0.17	0.18	0.23	0.52	0.45
	(0.27)	(283)	(0.07)	(0.07)	(0.05)	(0.07)	(0.12)
Fewer than half	−0.02	461	−0.15	0.17	0.22	0.51	0.45
	(.12)	(345)	(0.06)	(0.06)	(0.05)	(0.06)	(.12)
Many	−0.17	203	−0.21	0.21	0.23	0.54	0.42
	(0.32)	(286)	(0.08)	(0.06)	(0.05)	(0.05)	(0.11)
Most	0.10	196	−0.23	0.21	0.23	0.54	0.42
	(0.22)	(280)	(0.09)	(0.07)	(0.05)	(0.06)	(0.11)
More than half	0.006	456	−0.23	0.21	0.23	0.54	0.42
	(0.06)	(346)	(0.07)	(0.06)	(0.05)	(0.05)	(0.11)

Table 4
Summary of the model testing Uncertainty hypothesis in Experiment 1

Effect	Estimate	*t* value	*p*-value
intercept	−0.0002	−0.01	.99
prop	−05	−6.76	<.001
quant	−06	−5.66	<.001
resp	0.01	1.21	.23
prop:quant	0.04	3.80	<.001
prop:resp	0.11	10.42	<.001
quant:resp	0.02	1.12	.26
prop:quant:resp	−0.07	−4.69	<.001

Note. prop = proportion; quant = quantifier; resp = response.

participants. Next, we constrained the p_0 parameter for *most* as well. The model fitted better for 41 out of 72 participants. These results therefore support the Vagueness hypothesis: Only some participants had a 50% threshold for *most*, and the variation in thresholds was higher for *most* than *more than half*. This difference can be observed in Fig. 5, which shows a greater variability for *most* than *more than half* in drift rates for the model with p_0 as a free parameter (Model 7).

2.4.4. Uncertainty hypothesis

Next, we tested the Uncertainty hypothesis (see model summary in Table 4 and the comparison of regression models in the Appendix). We applied the individual Bayesian model averaged thresholds to each participant's response data and included the "true" responses above the threshold and "false" responses below the threshold in the analysis for *most* and *more than half*. As expected, we found a significant effect of proportion ($\beta = -0.05$; $t = -6.76$; $p < .001$) and a significant proportion-quantifiers interaction ($\beta = 0.04$; $t = 3.80$;

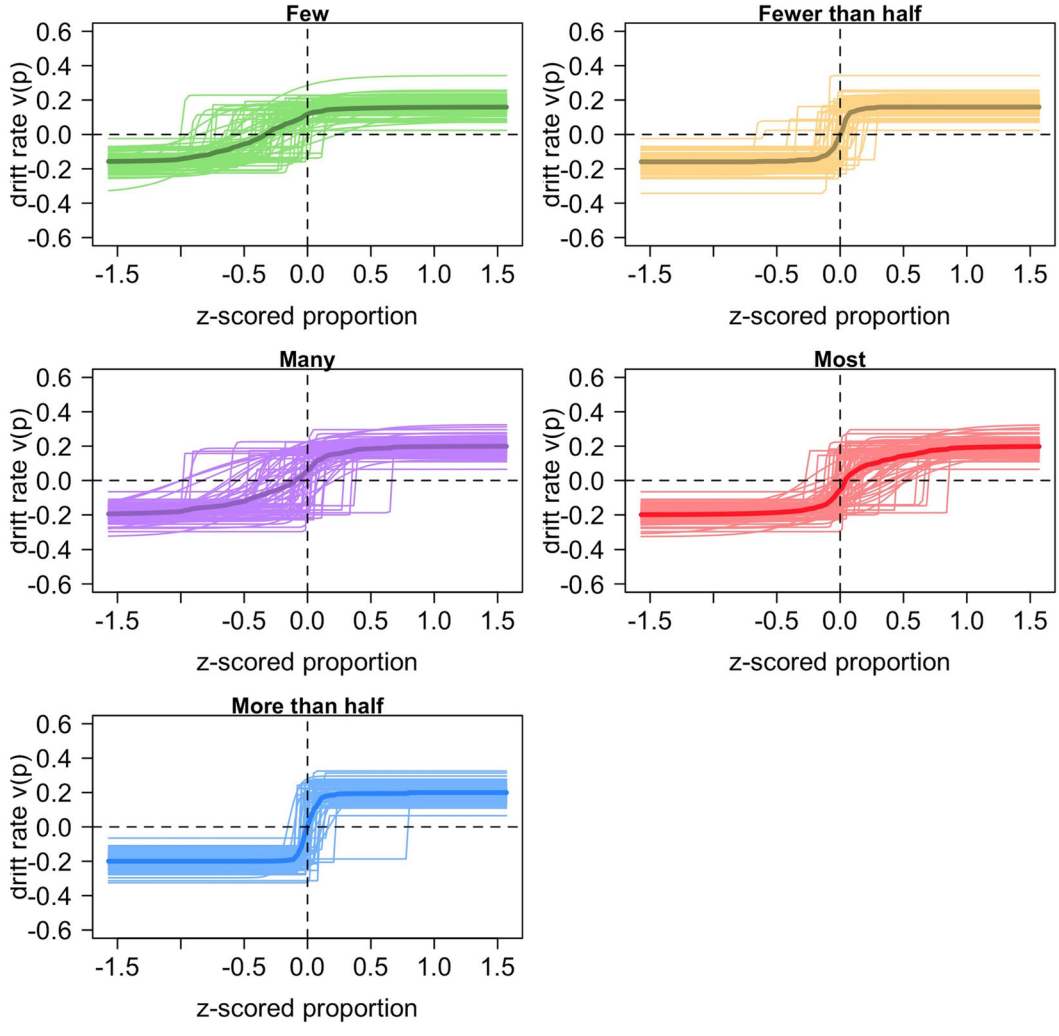

Fig. 5. Drift rates in Experiment 1 for Model 7. Darker lines indicate mean drift rate and lighter lines indicate individual drift rates. Individual participants' threshold p_0 is the proportion on the x-axis for which drift rate $v(p)$ on the y-axis is equal to zero. The scale parameter s indicates the steepness of the drift rate function.

$p < .001$). In addition, we found that *most* was verified more slowly than *more than half* ($\beta = -0.06$; $t = -5.66$; $p < .001$). This finding supports the Uncertainty hypothesis that verification of *most* is proportion-dependent. The relationship between proportion and reaction times is illustrated in Fig. 6. *Most* was verified slower when the proportion presented in the second sentence was close to the individual threshold of a participant. In contrast, no change in the speed of verification was observed for *more than half*.

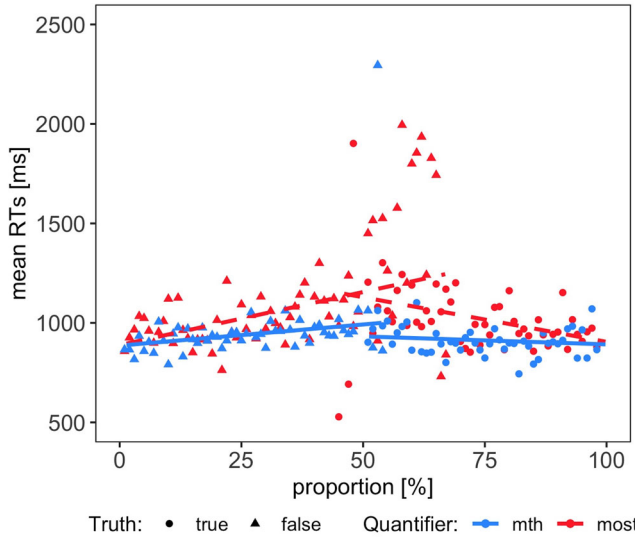

Fig. 6. RTs as a function of proportion for *most* versus *more than half* (mth). Each triangle represents mean RTs for proportions below the individual threshold ("false" response), and each dot represents proportions above the individual threshold ("true" response). Dashed, red lines represent regression lines for *most*, and solid, blue lines represent *more than half*.

2.4.5. *Effect of threshold on reaction times*

Finally, we tested whether the choice of the individual threshold affects the speed of the verification process. We applied the individual Bayesian model averaged thresholds to each participant's response data and included the "true" responses above the threshold and "false" responses below the threshold in the analysis for *most*, *many*, and *more than half* (opposite for *few* and *fewer than half*). This analysis was exploratory in nature. We expected the selection of the threshold to affect the verification process of quantifiers with various possible representations such as *many*, *few*, and *most*. We fitted linear mixed-effects models to data from each quantifier separately.

More than half and *fewer than half*: We tested the effect of the individual threshold on reaction times (see Table 5). For both quantifiers, we found that the effect of the threshold was not significant: *more than half* ($\beta = -0.009$; $t = -1.45$; $p = .15$), *fewer than half* ($\beta = 0.003$; $t = 0.77$; $p = .44$). As predicted, we did not find evidence that the choice of threshold affects the speed of verification in *more than half* and *fewer than half*.

Most: For *most*, we found a significant effect of threshold ($\beta = 0.006$; $t = 2.32$ $p < .05$) and threshold-distance interaction ($\beta = -0.0002$; $t = -3.03$; $p < .01$; see Table 5). As expected, the verification of *most* was affected by the choice of threshold.

Many and *few*: For *many*, we found close to significance level 0.05 effect of threshold ($\beta = -0.003$; $t = -1.93$; $p = .056$) and no significant threshold-distance interaction ($\beta = 0.0001$; $t = -0.25$; $p = .80$; see Table 5). For *few*, we found close to significance level 0.05 effect

Table 5
Summary of the models estimates testing the effect of threshold

Effect	*More than half*	*Fewer than half*	*Most*	*Few*	*Many*
intercept	−0.06***	0.02	0.003	0.10***	0.003
Dist	−0.0004.	−0.00003	−0.002***	0.004***	−0.002***
Thr	−0.009	0.003	0.006*	−0.004	−0.003
Resp	0.03**	−0.006	0.02	−0.05***	0.04**
Dist:thr			−0.0002**	−0.0002**	−0.00001
Dist:resp	0.001***	−0.001**	0.004***	−0.005***	0.004***
Thr:resp	0.009***		−0.002	0.0001	−0.0006
Dist:thr:resp			0.0002**	0.0002**	−0.0001*

Note. dist = distance; thr = threshold; resp = response
***$p < .001$; **$p < .01$; *$p < .05$.

of threshold ($\beta = -0.004$; $t = -1.90$; $p = .06$) and significant threshold-distance interaction ($\beta = -0.0002$; $t = -2.75$; $p < .01$; see Table 5).

To summarize, we found the effect of threshold for *most* and close to the significance level of 0.05 effect of threshold for *few* and *many*, but we did not find this effect for quantifiers with sharp meaning boundaries. This finding gives moderate support to the hypothesis that the speed of the verification depends on the choice of threshold.

3. Experiment 2

The goal of the second experiment was to test the within-subject stability of truth-conditional representations over time. We tested participants in two sessions separated by 2 weeks and estimated the DDM parameters from the data from both sessions. Next, we tested if the threshold parameter was the same across the sessions indicating the stable meaning representations. We predicted that meaning representations should be stable for *more than half* and *fewer than half*, and we considered that they might change for vague quantifiers. In addition, we tested other DDM parameters to test the stability of model estimates over time.

3.1. Methods

3.1.1. Participants

We recruited 89 participants via the Prolific platform (https://www.prolific.co/), and 72 participants completed both sessions. We included 64 participants (46 female, see Section 3.1.2.) in the final sample. Participants were 32 years old on average ($SD = 9$, range: 18–60) and represented the following educational backgrounds: attending high school or high school graduates (five participants), high school graduates, who started college (six participants), and college graduates (53 female). Participants were paid 7.5£ per h. The study was approved by the Ethics Committee at the University of Amsterdam.

3.1.2. Exclusion criteria

We used the same exclusion criterion as in Experiment 1. We excluded participants if they met one of the criteria in at least one testing session. We excluded three fast-guessing participants and four participants who failed to meet the monotonicity criterion. In addition, we excluded one participant who was not a native English speaker.

3.1.3. Design

We used the same design as in Experiment 1.

3.1.4. Procedure

The stability experiment had the same number of trials as Experiment 1, and it was also preceded by a short training block. After completing the experiment, participants filled in a brief demographic survey. We simplified the procedure from Experiment 1. Participants had to press the K key on their keyboard to move to the first screen, containing a sentence with a quantifier, but they did not have to keep the K key pressed. To move to the second screen, containing a sentence with proportion, participants had to press the K key again. To provide a response ("true" or "false") they had to press the J or L key (counterbalanced across participants). Participants did the experiment twice. The proportions presented in the second sentence and the order of the trials were randomized across sessions. The second session was held in the third week after the first session.

3.1.5. Preprocessing reaction time data

We used the same preprocessing procedure as in Experiment 1. We excluded 4.4% of the trials in the first session and 2% in the second session.

3.2. Computational model

We applied Model 7 from Experiment 1 to data from both sessions (Table 7 shows mean parameters). We chose this model because it included all the constraints, which highlighted differences between positive and negative quantifiers and between vague and quantifiers with sharp meaning boundaries quantifiers but had the threshold parameter-free.

3.3. Results

3.3.1. Descriptive statistics

We summarize the descriptive statistics in both sessions (Table 6). We observed similar patterns to those observed in Experiment 1: We found shorter reaction times for *more than half* than *most* and larger proportion of "false" responses for *most*. We also noticed that participants were faster in the second session, which indicates a learning effect.

3.3.2. Model fit and model parameters

Fig. 7 shows that Model 7 from Experiment 1 also has a good fit to the data from both sessions in Experiment 2. Table 7 shows that parameter estimates are also comparable to those found in Experiment 1.

Table 6
Mean RTs in seconds (*SD*) and proportion of "true" versus "false" responses in each session

Quantifier	Truth Value	Session 1		Session 2	
		RTs	Responses	RTs	Responses
Few	False	1.232 (0.143)	0.60	1.103 (0.121)	0.59
Few	True	1.286 (0.130)	0.40	1.168 (0.181)	0.41
Fewer than half	False	1.114 (0.104)	0.53	1.123 (0.280)	0.51
Fewer than half	True	1.179 (0.109)	0.47	1.141 (0.216)	0.49
Many	False	1.141 (0.145)	0.47	1.021 (0.136)	0.49
Many	True	1.082 (0.115)	0.53	0.978 (0.146)	0.51
Most	False	1.078 (0.112)	0.54	1.015 (0.206)	0.53
Most	True	1.043 (0.134)	0.46	0.961 (0.183)	0.47
More than half	False	1.016 (0.127)	0.51	0.919 (0.143)	0.50
More than half	True	0.965 (0.099)	0.49	0.909 (0.302)	0.50

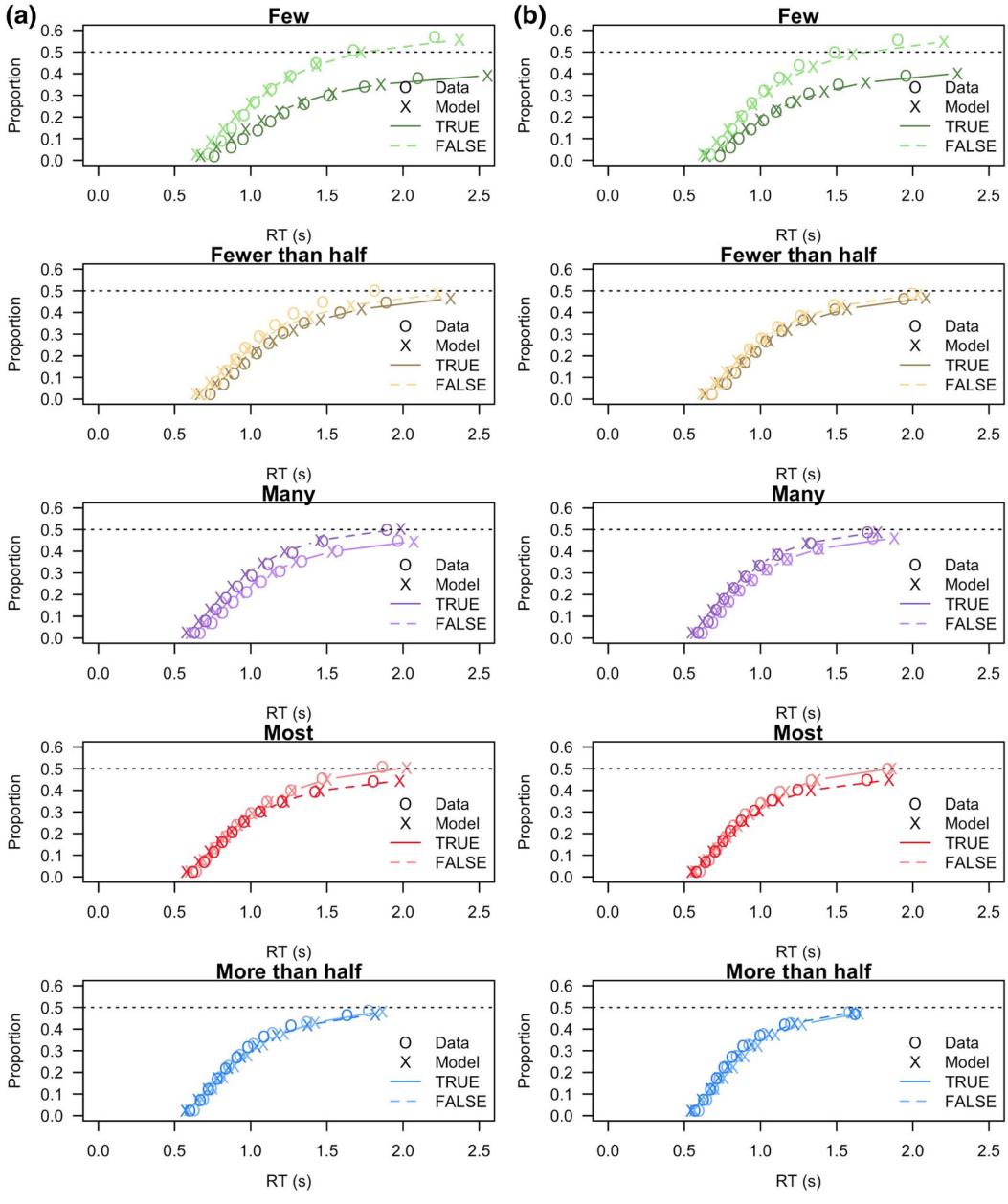

Fig. 7. Defective cumulative density plots show the average fit of Model 7 to data from Session 1 (a) and Session 2 (b). For this visualization, we plot the mean 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95% percentiles over participants, scaled by the proportion of "true" and "false" responses, separately for the data and Model 7 prediction.

Table 7
Mean parameters (SD) for each quantifier in both sessions

Session	Quantifier	p_0	s	V_L	V_U	a	z	T_{er}
1	Few	−0.28	284	−0.14	0.14	0.24	0.52	0.43
		(0.33)	(362)	(0.04)	(0.04)	(0.06)	(0.06)	(0.11)
1	Fewer than half	0.01	457	−0.14	0.14	0.24	0.52	0.43
		(0.15)	(355)	(0.04)	(0.04)	(0.06)	(0.06)	(0.11)
1	Many	−0.06	284	−0.19	0.19	0.24	0.52	0.38
		(0.33)	(362)	(0.04)	(0.04)	(0.06)	(0.06)	(0.09)
1	Most	0.15	284	−0.19	0.19	0.24	0.52	0.38
		(0.21)	(362)	(0.04)	(0.04)	(0.06)	(0.06)	(0.09)
1	More than half	0.02	457	−0.19	0.19	0.24	0.52	0.38
		(0.15)	(355)	(0.04)	(0.04)	(0.06)	(0.06)	(0.09)
2	Few	−0.27	196	−0.16	0.16	0.23	0.51	0.43
		(0.28)	(296)	(0.05)	(0.05)	(0.05)	(0.04)	(0.09)
2	Fewer than half	0.001	514	−0.16	0.16	0.23	0.51	0.43
		(0.11)	(352)	(0.05)	(0.05)	(0.05)	(0.04)	(0.09)
2	Many	−0.03	196	−0.22	0.22	0.23	0.53	0.37
		(0.31)	(296)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)
2	Most	0.13	196	−0.22	0.22	0.23	0.53	0.37
		(0.23)	(296)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)
2	More than half	0.03	514	−0.22	0.22	0.23	0.53	0.37
		(0.09)	(352)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)

Table 8
Bayes factors for paired *t* tests for parameter estimates between sessions

Quantifier	p_0	s	V_L	V_U	a	z	T_{er}
Few	0.14	0.41	71	71	0.63	0.39	0.15
Fewer than half	0.16	0.19	71	71	0.63	0.39	0.15
Many	0.17	0.41	1106	1106	0.63	0.15	0.16
Most	0.22	0.41	1106	1106	0.63	0.15	0.16
More than half	0.17	0.19	1106	1106	0.63	0.15	0.16

Note. The Bayes factors were the same if the parameters were constrained across quantifiers: s was the same for *more/fewer than half* and the same for *most*, *many*, and *few*; V_L and V_U were symmetric ($V_L = -V_U$); a was the same for all quantifiers; z and T_{er} were the same for positive and the same for negative quantifiers.

3.3.3. Stability of thresholds

The Logical Theories hypothesis predicts that thresholds should be stable over time. The Vagueness hypothesis allows variation in thresholds for *most*. To test these hypotheses, we performed Experiment 2 in which participants repeated the same experiment in two sessions 2 weeks apart.

To compare each parameter between the sessions, we used a Bayesian paired *t* test (R function *ttestBF* from *BeyesFactor* library; Morey, 2018). We chose to use a Bayesian statistic because we wanted to quantify evidence for the null hypothesis. The Bayesian *t* test indicates the relative likelihood of a difference in parameter estimates between sessions, expressed by the Bayes factor. A large Bayes factor suggests that there is a systematic parameter difference between the sessions, whereas a Bayes factor less than one provides evidence for the absence of a difference, suggesting a stable parameter across the sessions. We started by testing the stability of the threshold parameters (Table 8). We predicted that the threshold for *more than half* and *fewer than half* should be stable over time at 50%. We considered that thresholds for other quantifiers might differ between sessions.

We found that the Bayes factor for *most* was 0.22 and for *more than half* 0.17, which indicates evidence in favor of the null hypothesis (no difference between parameters between sessions). In particular, the Bayes factor for individual threshold parameters was below 0.33 for all quantifiers, which indicates substantial evidence in favor of the null hypothesis (Jeffreys, 1961). These results speak in favor of thresholds stability.

We also tested the stability of other DDM parameters between two experimental sessions (Table 8). In general, the model parameters were stable across sessions, apart from V_L and V_U parameters. We tested whether participants accumulated evidence faster in the second session, by computing the maximum speed of evidence accumulation in both sessions, which was operationalized as the distance between asymptotes ($V_U - V_L$) and tested the difference in distance between sessions (Bayesian paired *t* test, R function *ttestBF* from *BayesFactor* library; Morey, 2018). We found substantial evidence in favor of the hypothesis that participants speed up the evidence accumulation process in the second session for positive quantifiers (BF = 1106 ± 0%) and for negative quantifiers (BF = 71 ± 0%). This difference can be explained in terms of the training effect, consistent with previous literature that found

Table 9
Correlations of Diffusion Decision Model parameters between Sessions 1 and 2.

	p_0	s	V_L	V_U	a	z	T_{er}
Few	0.48***	0.03	0.55***	0.55***	0.77***	0.04	0.53***
Fewer than half	−0.05	−0.17	0.55***	0.55***	0.77***	0.04	0.53***
Many	0.47***	0.03	0.51***	0.51***	0.77***	0.31*	0.46***
Most	0.63***	0.03	0.51***	0.51***	0.77***	0.31*	0.46***
More than half	0.002	−0.17	0.51***	0.51***	0.77***	0.31*	0.46***

Note. Some parameters were constrained between quantifiers (see the modeling section 3.2), and therefore the correlations were the same, $df = 62$.

***$p < .001$; **$p < .01$; *$p < .05$.

that training effects are reflected in increased drift rates (Dutilh, Krypotos, & Wagenmakers, 2011; Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Petrov, van Horn, & Ratcliff, 2011).

3.3.4. Correlation of parameters between sessions

In addition, we correlated the thresholds parameters across sessions (Table 9 summarizes all correlations). The correlations between thresholds for vague quantifiers (*most, many*, and *few*) were moderate and significant (Fig. 8). The correlations between thresholds for quantifiers with sharp meaning boundaries (*more than half* and *fewer than half*) were close to zero (Fig. 8). Although this pattern of results seems counterintuitive, the very low correlations for *more than half* and *fewer than half* between thresholds reflect the very low variation in thresholds for these quantifiers. The correlations for *a,* T_{er}, V_L, V_U, and z (for positive quantifiers) parameters were moderate or high (Table 9).

4. Discussion

In this paper, we modeled the processing of vague concepts by means of a computational model of decision-making. We chose the quantifier verification task and two widely discussed quantifiers—*most* and *more than half*—for our case study. We used the DDM to distinguish different meaning aspects such as truth-conditional representation and vagueness. Our model accounted for vagueness by assuming a noisy decision process and individual differences in meaning representations by estimating parameters for each participant. We tested the predictions of logical theories about truth conditions of *most* and showed that these theories need an extension to account for individual differences in representations of *most*. In the discussion, first, we will summarize the main findings (Section 4.1) and interpret the revealed differences in representations of *most* and *more than half* (Section 4.2). Next, we will attempt to link our findings to a broad discussion about vague concepts (Section 4.3). Finally, we will discuss the limitations of this study and suggest future directions (Section 4.4).

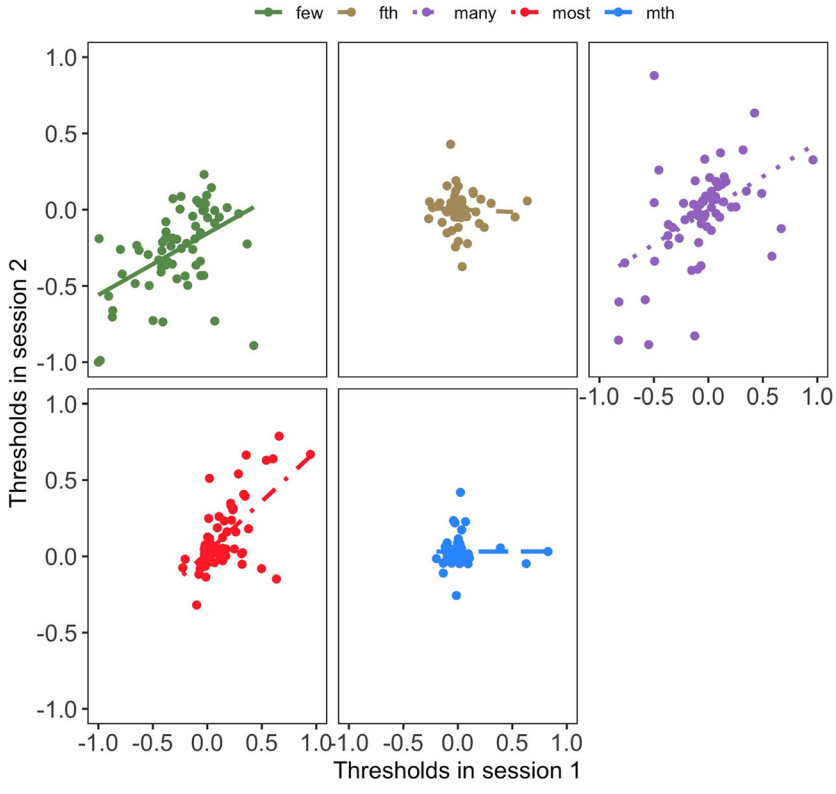

Fig. 8. Scatterplots from Experiment 2 showing correlations of thresholds between Sessions 1 and 2: *more than half* (mth) ($r(62) = .002$; $p = .99$), *fewer than half* (fth) ($r(62) = -.05$; $p = .72$), *most* ($r(62) = .63$; $p < .001$), *many* ($r(62) = .47$; $p < .001$), and *few* ($r(62) = .48$; $p < .001$).

4.1. Main findings

In Experiment 1, we tested two competing hypotheses: Logical Theories hypothesis and Vagueness hypothesis. The former hypothesis proposes that *most* has a truth-conditionally equivalent representation to *more than half*. The latter, in turn, claims that *most* is a vague quantifier with variable thresholds. Our modeling results support the Vagueness hypothesis and reveal substantial individual differences in the thresholds for *most*. The model with the threshold parameter p_0 constrained to 50% for *most* fitted better than the model with unconstrained p_0 only for some participants. This finding suggests that *most* is vague and sensitive to individual differences. In the next step, we tested the Uncertainty hypothesis. Vague quantifiers are more difficult to verify because participants may not be sure about the exact threshold. We observed that the verification of *most* is proportion-dependent, and participants verify *most* slower around the 50% proportion. This finding supports the hypothesis that *most* is a vague concept. In the exploratory analysis, we tested whether for vague quantifiers (*most*, *many*, and *few*) participants verified the proportions around the threshold slower (distance effect) because they were uncertain about the threshold. We found support for this prediction

in the reaction times analysis and in modeling, as reflected by a difference in the *s* parameter between quantifiers. Moreover, we investigated if the speed of verification depends on the choice of the threshold (threshold effect). The results were equivocal, and the threshold effect did not fully replicate (see Supplementary Materials).

In Experiment 2, we tested the stability of the threshold parameter and other parameters of the DDM. The thresholds were stable over time. In addition, we found strong evidence for a learning effect.

4.2. Most *versus* more than half

The logical theories make a strong and clear prediction about the meaning representation of *most*, which should be truth-conditionally equivalent to *more than half*. The empirical data (Kotek et al., 2015; Solt, 2016) showed that *most* is used and verified differently than *more than half*. These patterns of results have been explained in multiple ways (Carcassi & Szymanik, 2021; Hackl, 2009; Pietroski et al., 2009; Solt, 2016). None of these explanations, however, considered the difference in truth conditions between *most* and *more than half*. After controlling for pragmatic and processing effects, and accounting for vagueness and individual differences, we observe a difference in threshold between *most* and *more than half*, which is attributable to an inequivalence in truth conditions. Our results question the typical formulation of truth conditions for *most* on the basis of logical theories. Recall the example in the Introduction section. If *most* has a truth condition formulated as in Example 2, then it should also have a 50% threshold, like *more than half*. Solt (2016) claimed that *most* has a truth condition as in Example 2, but it is preferably used with higher proportions. In order to be judged as *most*, the proportion of As that is B has to be "significantly" higher than a proportion of As that is not B. This happens because *most* is represented on a semi-ordered scale, on which the proportions are compared using the approximate number system. *More than half* is represented on a ratio scale, on which proportions can be compared precisely. For the approximate comparison, the proportions have to be "significantly" greater to be distinguished from each other.

In the current study, we used an experimental paradigm, which makes this explanation unlikely. First, in our experiment, participants had to compare precise proportions given as a number. The scale on which they represented proportions was the same for both quantifiers. It was a ratio scale. Moreover, as the numbers were precise and there was no time pressure, there was no reason for participants to use the approximate number system for *most*. Taking this into account, we endorse the conclusion that the difference between *most* and *more than half* in our experiment is due to differences in meaning representations rather than processing strategies. More specifically, we claim that for both quantifiers, participants compared the given proportion to their internal threshold, but for *more than half*, this threshold was 50%, while for *most*, it varied between participants. The meaning representation of *most* is, therefore, similar to the representation of the proportional *many*. When participants verify the proportional *many*, they compare the given proportion to the threshold (Partee, 1988). For both *most* and *many*, the threshold varied between-participants.

In addition to testing the between-participants variability in truth conditions of *most*, we also evaluated their stability over time. Previous studies (Verheyen et al., 2019) showed that semantic categories change over time. Our study is the first study that addressed the stability of semantic representations of functional words such as quantifiers. We showed that participants' truth conditions are stable over a short period. Verheyen et al. (2019) suggested that the inter-individual differences relate to stable differences between groups that apply different categorization criteria. For example, the level of education (Verheyen & Storms, 2018) or individual traits (Verheyen, Dewil, & Égré, 2018) affect the choice of category criteria. Further studies need to explain the cause of the inter-individual variation in thresholds for vague quantifiers.

In both Experiment 1 and its replication (see Supplementary Materials), we found that *most*, but not *more than half*, was verified slower when the proportions were close to 50% proportion. Our modeling data reflected the effect of proportion on verification of *most*. We discovered that *most* and *more than half* differ in the *s* parameter (see the Appendix), which we interpreted as a measure of uncertainty about the threshold. The *s* parameter models the steepness of the drift rate curve. For *more than half*, the drift rate had the shape of a step-like function, which indicates that the evidence accumulation process was equally fast for all proportions. For *most*, in turn, the drift rate has a smoother shape indicating a slower evidence accumulation process around the threshold. Our data support the predicted drift rates in Fig. 2: the blue, solid line for *more than half* and higher threshold and greater uncertainty for *most* (dotted, red line). The modeling of the *s* parameter shows that *most* is more akin to *many* than *more than half*.

Our exploratory analysis showed that the choice of threshold might affect reaction times for vague quantifiers (*most, many*, and *few*). In the case of vague quantifiers, this analysis also showed that participants were faster when the verified proportion was further from their threshold than when it was close. We should consider these findings with caution because the effect did not fully replicate (see Supplementary Materials).

By presenting proportion as a number, we ruled out the processing explanation of differences between *most* and *more than half*. By using pseudowords, we limited the pragmatic effects. Nonetheless, one could argue that participants constructed a context for the experiment, especially because they had to verify the context-dependent quantifiers such as *many* and *few*. Although we cannot completely rule out this possibility, we argue that this is not very likely. First, participants did not assign any meaning to pseudowords (indicated by a very small amount of variance explained by by-item random intercepts, see the Appendix). Second, the second experiment shows stable threshold parameters for all quantifiers indicating that participants probably did not construct contexts *ad hoc*.

We believe that our experimental paradigm controlled well for the contextual effects. However, the question remains if the individual differences in the threshold for *most* could be due to individual differences in the pragmatics skills of participants, for example, in the sensitivity to pragmatic strengthening. Previous studies have shown (e.g., Spychalska, Kontinen, & Werning, 2016) individual differences in the availability of the pragmatically enriched representations in the case of the quantifier *some*. The logical responders interpret *some* according to its truth conditions (*some* can possibly mean *all*), while pragmatic responders interpret

some as *some but not all*. Contrary to this example, we think the differences between *most* and *more than half* are unlikely caused by pragmatic strengthening. In a similar experimental paradigm, Denić and Szymanik (2022) tested the pragmatic strengthening hypothesis directly and showed that it cannot explain the higher threshold for *most*. Our findings are consistent with this result.

4.3. Theories of vague concepts

Having discussed the main findings, in this section, we will take stock of how our results fit the literature about vague concepts. First our modeling results revealed two sources of vagueness: individual variability in thresholds and uncertainty around thresholds. We separated their effect by using different model parameters p_0 and s and estimating these on an individual participant level. Following the literature on vagueness and categorization, we suggest that these two types of vagueness could be interpreted as criterial and degree vagueness (Verheyen, Droeshout, & Storms, 2019). Criterial vagueness is a disagreement between participants about the criteria or conditions to classify a given object into a given category (Verheyen et al., 2019). Degree vagueness, in turn, means that individuals agree about the classification criterion but disagree to what extent the given object satisfies this criterion (Verheyen et al., 2019). For example, one can believe that an abstract painting cannot be considered a piece of art because art should imitate reality (criterial vagueness). In contrast, this person can consider two realistic paintings as art but to a different extent depending on how well they imitate reality (degree vagueness). Both types of vagueness can cause between-participants variation in semantic categorization (Verheyen & Storms, 2013, 2018). We suggest that the variability in thresholds may be interpreted as criterial vagueness and uncertainty as degree vagueness. The application of different thresholds resembles the usage of different categorization criteria. The uncertainty about the threshold shows that the decision of whether the given proportion is above or below the threshold is harder for vague concepts.

Several theoretical frameworks have been developed to study the structure of mental representations of vague concepts. The multiple definitions model (Gruenenfelder, 2019) proposes that participants store in memory multiple definitions of a vague concept. In the domain of quantifiers, this would mean that participants store in memory multiple thresholds for each quantifier. During the verification process, one of the thresholds is sampled and compared to the proportion given in the experiment. What is challenging for this theory is to provide a sampling mechanism for the thresholds. For example, in our study, participants may have sampled the thresholds uniformly or they may have had a specific probability distribution over thresholds. The former scenario would result in a variability in thresholds from trial-to-trial and the latter would guarantee the more stable threshold across the trials. Another challenge to the multiple definitions model would be to account for our findings in the parameter s. The value of s parameter for vague quantifiers indicated that the reaction times around the threshold are slower. However, the multiple definitions model does not directly predict the reaction time difference. Under the assumption that participants store a set of thresholds in their memory and sample one threshold in each trial, there is no reason why they would verify slower the proportions closer to the threshold. Finally, our results on the stability of thresholds

over time could be explained by the multiple definitions model if we assume that participants sample certain thresholds with higher probability or if they sample uniformly from a large set of thresholds.

An alternative account, the partial model, was proposed by Kamp and Partee (1995). They used partial models to formally account for vagueness. Considering this account in the domain of quantifiers, we suggest that participants have a single threshold (a proportion) and intermediate ranges of proportions around the threshold with an undefined truth value. For example, let us consider a participant with a threshold equal to 50% and an intermediate region between 45% and 55%. The definition of the threshold can be extended in the range of the intermediate region by using partial models. The partial models are defined as follows: downward from the threshold by subtracting 1 in each step (49–50; …; 45–50), upward from the threshold by adding 1 in each step (50–51; …; 50–55), and both sides from the threshold by subtracting 1 and adding 1 (49–51; …; 45–55). Next, for a specific threshold and each proportion in the intermediate region, we can compute the proportion of completions, meaning the ratio of partial models that contain this proportion to the models that do not contain this proportion. In our example, all partial models contain the threshold, but only two out of five partial models contain the proportion (47%). The proportion of completions provides a mechanism explaining the uncertainty about choices that are missing in the multiple definitions model. However, the partial model does not provide any explicit prediction about the reaction times. Moreover, this approach originated from studies on categorization and assumes the existence of prototypes. Our modeling approach is agnostic to the porotype theory.

To summarize, in this paper, we provided a novel framework to study vague concepts that made it possible to separate different sources of vagueness. We suggested interpreting the two sources of vagueness as criterial and degree vagueness. Moreover, we sketched the proposal on how to link our model with existing accounts to vagueness in a domain of categorization: the multiple definitions model and the partial model. We also showed that both theories need further specifications to make predictions about the reaction times data. The contribution of this paper is to encourage further development of both theories that would enable to test them against each other.

4.4. Limitations and future directions

We also note a few limitations of this study. First, we did not find a single model that would fit best for all participants. In our analysis, we account for this fact by including Bayesian model averaged parameters. However, we cannot conclude that differences between representations of *most* and *more than half* in threshold (p_0 parameter) and uncertainty (s parameter) are the only possible sources of inter-individual variations. Second, although we obtained a good model fit, we noticed that the model sometimes failed to predict long reaction times. The worse model fit for long reaction times is not surprising because long reaction times are rare and, therefore, difficult to predict accurately; however, in our experiment, long reaction times mostly drove the proportion effect. Third, we notice that our regression models did not meet all mixed-effects model assumptions even after log-transformation of the reaction time vari-

able. Finally, we did not replicate all the thresholds effects on reaction times. Therefore, we can only draw a limited conclusion about the relationship between the speed of verification and truth-conditional representation.

It goes without saying that further studies are needed to better understand the nature of mental representations of vague concepts. Here, we only outline a few possible future directions in the context of our results. First, our findings suggest that *most* and *more than half* have different truth-conditional representations. This finding is challenging for logical theories of meaning. A new, full-fledged proposal of the semantics of *most* goes beyond the scope of this paper. Denić and Szymanik (2022) made an initial attempt to formulate semantics of *most* as vague quantifier.

Second, we argued in this paper for individual differences in semantic representations of vague quantifiers. Our modeling results support this claim but do not explain the origin of the individual differences. Participants may have different thresholds for vague quantifiers because of differences in linguistic experience or education (Verheyen & Storms, 2018). Individual differences in linguistic behavior may be caused by individual differences in other cognitive abilities (Kidd, Donnelly, & Christiansen, 2018), for example, working memory (Zajenkowski, Szymanik, & Garraffa, 2014) or executive functions (Zajenkowski & Szymanik, 2013). Moreover, in a different experimental paradigm, for example, picture sentence verification task, participants could also use different verification strategies (Talmina, Kochari, & Szymanik, 2017), which, in turn, could affect the uncertainty and threshold parameters of our model. For example, we would predict that participants who use less precise strategy (e.g., based on the approximate number system) have a lower value of the *s* parameter. We suggest investigating the mechanistic explanations (e.g., in terms of verification strategies) of the differences in semantic representations of *most* and *more than half* and correlation explanations (e.g., in terms of other cognitive functions and abilities) in future experiments.

5. Conclusion

The current study shows that computational modeling is necessary to understand complex linguistic behavior. Our modeling data showed apparent differences between *most* and *more than half* and also between negative and positive quantifiers (see the Appendix). This finding indicates that we can formulate testable hypotheses about the DDM parameters to answer other linguistic questions beyond the current case study. For example, Schlotterbeck et al. (2020) linked the difference in non-decision time with an extra step in the verification of negative quantifiers and drift rate with difficulties in processing negative quantifiers. Furthermore, the starting point can model the response bias in different contexts. The DDM with implemented approximate number system model (Ratcliff & McKoon, 2018) could be used to test how quantifiers interact with different cognitive systems (e.g., approximate and precise number systems) and how verification changes with changing task demands (Register et al., 2020). To conclude, we presented a novel approach to systematically study linguistic phenomenon as the decision-making process.

Note

1 In literature, decision boundaries are sometimes called decision thresholds. In this paper, we reserve the notion of threshold to semantic representations of quantifiers, and to avoid confusion, we use the term "decision boundaries" to refer to options between which participants choose.

2 Data and analysis scripts are available at: osf.io/d5xm8

Acknowledgments

Sonia Ramotowska, Shane Steinert-Threlkeld, and Jakub Szymanik have 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. We thank Fabian Schlotterbeck, Milica Denić, and Fausto Carcassi for the discussion and comments on the early draft of this manuscript.

Open access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In K. G. Parzen E. & K.Tanabe (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). New York: Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219. https://doi.org/10.1007/BF00350139

Bott, O., Augurzky, P., Sternefeld, W., & Ulrich, R. (2017). Incremental generation of answers during the comprehension of questions with quantifiers. *Cognition*, *166*, 328–343. https://doi.org/10.1016/j.cognition.2017.05.023

Carcassi, F., & Szymanik, J. (2021). 'Most' vs 'More Than Half': An alternatives explanation. *Proceedings of the Society for Computation in Linguistics*, *4*, 334. https://scholarworks.umass.edu/scil/vol4/iss1/30

Clark, H. H. (1976). *Semantics and comprehension*. The Hague: Mouton & Co. B. V.

Clerc, M. (2006). *Particle swarm optimization*. New York: Wiley. https://doi.org/10.1002/9780470612163

Coppock, E., & Strand, L. (2019). Most vs. the most in languages where the more means most. In A. Aguilar-Guevera, J. Pozas Loyo, & V. Vázquez Rojas Maldonado (Eds.), *Definiteness across languages* (pp. 271–417). Berlin: Language Science Press.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition* (pp. 526–574). Oxford, England: Oxford University Press. https://doi.org/10.1093/acprof:oSo/9780199231447.003.0024

Denić, M., & Szymanik, J. (2022). Are *most* and *more than half* truth-conditionally equivalent? *Journal of Semantics*, *39*(2), 261–294.

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–128. https://doi.org/10.1016/j.cognition.2015.06.006

Donkin, C., & van Maanen, L. (2014). Piéron's Law is not just an artifact of the response mechanism. *Journal of Mathematical Psychology*, *62–63*, 22–32. https://doi.org/10.1016/J.JMP.2014.09.006

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task-related versus stimulus-specific practice a diffusion model account. *Experimental Psychology*, *58*, 434–442. https://doi.org/10.1027/1618-3169/a000111

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E. J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin and Review*, *166*, 1026–1036. https://doi.org/10.3758/16.6.1026

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. https://doi.org/10.1126/science.1218633

Glöckner, I. (2006). *Fuzzy quantifiers*. Berlin, Heidelberg: Springer.

Gruenenfelder, T. M. (2019). A multiple definitions model of classification into fuzzy categories. *Frontiers in Psychology*, *10*, 944. https://doi.org/10.3389/FPSYG.2019.00944/BIBTEX

Guest, O., & Martin, A. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 1-14. https://doi.org/10.31234/osf.io/rybh9

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *Most* versus *more than half*. *Natural Language Semantics*, *17*(1), 63–98. https://doi.org/10.1007/s11050-008-9039-x

Jeffreys, H. (1961). *The theory of probability*, New York: Oxford University Press.

Heim, S., McMillan, C. T., Clark, R., Golob, S., Min, N. E., Olm, C., Powers, J., & Grossman, M. (2015). If so many are "few," how few are "many"? *Frontiers in Psychology*, *6*, 441. https://doi.org/10.3389/fpsyg.2015.00441

Heim, S., Peiseler, N., & Bekemeier, N. (2020). "Few" or "Many"? An adaptation level theory account for flexibility in quantifier processing. *Frontiers in Psychology*, *11*, 382. https://doi.org/10.3389/fpsyg.2020.00382

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology (2006)*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–401. https://doi.org/10.2307/2676803

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*(3), 244–253. https://doi.org/10.1016/S0022-5371(71)80051-8

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kuvač Kraljević, J., Hrzica, G., Grohmann, K. K., Skordi, A., Jensen de López, K., Sundahl, L., van Hout, A., Hollebrandse, B., Overweg, J., Faber, M., van Koert, M., Smith, N., Vija, M., Zupping, S., Kunnari, S., Morisseau, T., Rusieshvili, M., Yatsushiro, K., Fengler, A., Varlokosta, S., Konstantzou, K., Farby, S., Guasti, M. T., Vernice, M., Okabe, R., Isobe, M., Crosthwaite, P., Hong, Y., Balčiūnienė, I., Ahmad Nizar, Y. M., Grech, H., Gatt, D., Cheong, W. N., Asbjørnsen, A., Torkildsen, J., Haman, E., Miękisz, A., Gagarina, N., Puzanova, J., Anđelković, D., Savić, M., Jošić, S., Slančová, D., Kapalková, S., Barberán, T., Özge, D., Hassan, S., Chan, C. Y., Okubo, T., van der Lely, H., Sauerland, U., & Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(33), 9244–9249. https://doi.org/10.1073/pnas.1601341113

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Kotek, H., Sudo, Y., & Hackl, M. (2015). Experimental investigations of ambiguity: The case of most. *Natural Language Semantics*, *23*(2), 119–156. https://doi.org/10.1007/s11050-015-9113-0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256. https://doi.org/10.1007/s11050-010-9062-6

van Maanen, L., Grasman, R. P. P. P., Forstmann, B. U., Keuken, M. C., Brown, S. D., & Wagenmakers, E. J. (2012). Similarity and number of alternatives in the random-dot motion paradigm. *Attention, Perception, and Psychophysics*, *74*(4), 739–753. https://doi.org/10.3758/S13414-011-0267-7/FIGURES/12

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, *7*, 120. https://doi.org/10.3389/FPSYG.2016.00120/BIBTEX

Miletić, S., & van Maanen, L. (2019). Caution in decision-making under time pressure is mediated by timing ability. *Cognitive Psychology*, *110*, 16–29. https://doi.org/10.1016/j.cogpsych.2019.01.002

Morey, R. (2018). *Package "BayesFactor" Title Computation of Bayes Factors for Common Designs*. https://richarddmorey.github.io/BayesFactor/

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae*, *44*(1), 12–36. https://doi.org/10.4064/fm-44-1-12-36

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences—A model-based review. *Neuroscience*, *277*, 872–884. https://doi.org/10.1016/j.neuroscience.2014.07.031

Partee, B. (1988). Many quantifiers. *Proceedings of ESCOL 5*, Ithaca, NY (pp. 383–402). https://doi.org/10.1002/9780470751305.ch12

Petrov, A. A., van Horn, N. M., & Ratcliff, R. (2011). Dissociable perceptual-learning mechanisms revealed by diffusion-model analysis. *Psychonomic Bulletin and Review*, *18*(3), 490–497. https://doi.org/10.3758/s13423-011-0079-8

Pexman, P. M., & Yap, M. J. (2018). Individual differences in semantic processing: Insights from the Calgary Semantic Decision Project. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(7), 1091–1112. https://doi.org/10.1037/xlm0000499

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of "Most": Semantics, numerosity and psychology. *Mind and Language*, *24*(5), 554–585. https://doi.org/10.1111/j.1468-0017.2009.01374.x

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. https://doi.org/10.1016/S1364-6613(02)01964-2

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2020). Most, but not more than half, is proportion-dependent and sensitive to individual differences. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, & J. Schwab (Eds.), *Proceedings of Sinn und Bedeutung* (Vol. 24 pp. 165–182). Osnabrück: Osnabrück University.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461. https://doi.org/10.1037/0033-2909.86.3.446

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review*, *125*(2), 183–217. https://doi.org/10.1037/rev0000085

Ratcliff, R., & McKoon, G. (2020). Decision making in numeracy tasks with spatially continuous scales. *Cognitive Psychology*, *116*, 101259. https://doi.org/10.1016/j.cogpsych.2019.101259

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*, 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127–157. https://doi.org/10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition*, *137*, 115–136. https://doi.org/10.1016/j.cognition.2014.12.004

Register, J., Mollica, F., & Piantadosi, S. T. (2020). Semantic verification is flexible and sensitive to context.

Schlotterbeck, F., Ramotowska, S., van Maanen, L., & Szymanik, J. (2020). Representational complexity and pragmatics cause the monotonicity effect. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 3398–3404).

Solt, S. (2011). Vagueness in quantity: Two case studies from a linguistic perspective. *Understanding Vagueness. Logical, Philosophical and Linguistic Perspectives*, *36*, 157–174.

Solt, S. (2015). Vagueness and imprecision: Empirical foundations. *Annual Review of Linguistics*, *1*(1), 107–127. https://doi.org/10.1146/annurev-linguist-030514-125150

Solt, S. (2016). On measurement and quantification: The case of most and more than half. *Language*, *92*(1), 65–100. https://doi.org/10.1353/lan.2016.0016

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *31*(6), 817–840. https://doi.org/10.1080/23273798.2016.1161806

Szymanik, J. (2016). *Quantifiers and cognition: Logical and computational perspectives*. Cham: Springer. https://doi.org/10.1007/978-3-319-28749-2

Talmina, N., Kochari, A., & Szymanik, J. (2017). Quantifiers and verification strategies: Connecting the dots. *Proceedings of the 21st Amsterdam Colloquium*, Amsterdam, The Netherlands (pp. 465–473).

Verheyen, S., Dewil, S., & Égré, P. (2018). Subjectivity in gradable adjectives: The case of *tall* and *heavy*. *Mind & Language*, *33*(5), 460–479. https://doi.org/10.1111/mila.12184

Verheyen, S., Droeshout, E., & Storms, G. (2019). Age-related degree and criteria differences in semantic categorization. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.74

Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PLoS One*, *8*(5), e63507. https://doi.org/10.1371/journal.pone.0063507

Verheyen, S., & Storms, G. (2018). Education as a source of vagueness in criteria and degree. In G. W. Sassoon, L. McNally, & E. Castroviejo (Eds.), *Gradability, scale structure and vagueness: Experimental perspectives* (pp. 149–167). Cham: Springer. https://doi.org/10.1007/978-3-319-77791-7_6

Verheyen, S., White, A., & Égré, P. (2019). Revealing criterial vagueness in inconsistencies. *Open Mind*, *3*, 41–51. https://doi.org/10.1162/opmi_a_00025

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. *Experimental Psychology*, *60*(6), 385–402. https://doi.org/10.1027/1618-3169/A000218

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. https://doi.org/10.1016/j.jml.2015.08.003

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, *9*(1), 149–184. https://doi.org/10.1016/0898-1221(83)90013-5

Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent. Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence. A Multidisciplinary Journal*, *41*(5), 456–466. https://doi.org/10.1016/j.intell.2013.06.020

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research*, *43*, 839–853. https://doi.org/10.1007/s10936-013-9281-3

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material

Appendix

Uncertainty hypothesis—Regression model comparison

To test the effect of proportion on reaction times for *most* in comparison to *more than half*, we tested the random structure of the mixed-effects regression model. In the first step, we included by-subject and by-item random intercepts. While participants were likely to vary in reaction times, we expected the small random effect of item because we used pseudowords. We found that the model with by-subject and by-item random intercepts did not significantly improve model fit, compared to a model with only by-subject random intercept ($\chi^2(1) = 0.64$; $p = .43$). This supports our assumption that participants did not analyze the meaning of pseudowords.

The model comparison revealed that the best model includes by-subject random slope for quantifier ($\chi^2(2) = 24.76$; $p < .001$). The by-subject random slope for response ($\chi^2(2) = 19.90$; $p < .001$), and by-subject random slope for proportion ($\chi^2(2) = 12.95$; $p = .002$) improved model fit less. In the next step, we included by-subject random slope for response ($\chi^2(3) = 20.94$; $p = .0001$). The model with three random slopes was overfit.

Effect of threshold on reaction time—Regression model comparison

More than half and *fewer than half*: For *more than half*, we included only by-subject and by-item random intercepts because a random slope for a response did not improve the model ($\chi^2(2) = 4.24$; $p = .12$), and a model with a random slope for distance did not converge. For *fewer than half*, we included by-subject and by-item random intercepts and by-subject random slope for response ($\chi^2(2) = 14.42$; $p = .001$). The model with by-subject random slope for distance did not converge.

Most: For *most*, we included by-subject random slope for response ($\chi^2(2) = 18.05$; $p = .0001$). The model with by-subject random slope for distance did not converge. We did not include by-item random intercept because the model was overfitted and the intercept did not explain any variance.

Many and few: For *many*, the best random effect structure included by-subject random intercept only (the models with by-subject random slopes for response did not improve fit ($\chi^2(2) = 3.37$; $p = .19$) and for distance did not converge). We did not include by-item random intercept because the model was overfitted and the intercept did not explain any variance.

For *few*, we included by-subject random slope for response ($\chi^2(2) = 16.55$; $p = .0003$). The model with by-subject random slope for distance did not converge. We did not include by-item random intercept because the model was overfitted and the intercept did not explain any variance.

Additional findings

In addition to testing the main hypotheses, we also found differences in other DDM parameters between quantifiers. We tested these differences using Bayesian model averaged parameters. We compared only the pairs of parameters that were not constrained between quantifiers.

First, we found that *more than half* had a higher *s* parameter than *most* (Fig. 4, $t(71) = -5.08$; $p < .001$; mean difference -259.39). This finding was expected. We found a proportion effect on reaction times for *most* but not *more than half*. This effect is also reflected in our modeling results in the difference in the *s* parameter.

Second, we also found a difference between positive and negative quantifiers in the Bayesian model averaged non-decision times parameters. We found that non-decision time was longer for *fewer than half* than *more than half* ($t(71) = 5.81$; $p < .001$; mean difference 0.03) and longer for *few* than *many* (Fig. 4, $t(71) = 5.70$; $p < .001$; mean difference 0.03). Furthermore, we found significant difference between the Bayesian model averaged starting point for *more than half* and *fewer than half* ($t(71) = -2.61$; $p = .01$; mean difference -0.03) and approached significance for *many* and *few* ($t(71) = -1.97$; $p = .053$; -0.03). The starting point for positive quantifiers (*more than half, many*) was higher than the starting point for corresponding negative quantifiers (*fewer than half, few*).

Finally, we also tested the distance between drift rate asymptotes (distance $= V_L - V_u$). The distance was greater for positive quantifiers than negative quantifiers: *more than half* versus *fewer than half* (Fig. 4, $t(71) = 11.87$; $p < .001$; mean difference 0.11), *many* versus *few* (Fig. 4, $t(71) = 5.51$; $p < .001$; mean difference 0.07). These findings indicate that the DDM can be useful to capture many properties of natural language quantifiers.