Most quantifiers have *many* meanings

Sonia Ramotowska*[a], Julia Haaf[b], Leendert Van Maanen[c], and Jakub Szymanik[a]

[a]Institute for Logic, Language and Computation, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands
[b]Psychological Methods, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands
[c]Department of Experimental Psychology & Helmholtz Institute, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands

Corresponding author: Correspondence concerning this article should be addressed to Sonia Ramotowska (sonia.ramotowska@hhu.de or s.ramotowska@uva.nl).

Data availability statement: The data and analysis code is available at https://github.com/jstbcs/pling-quant.

*Institut für Sprache und Information, Gebäude 23.21. Etage 04 Raum 75, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

Abstract

Formal semantic theories of meaning assume that function words, such as natural language quantifiers, have a rigid meaning expressed by their truth conditions. In this study, we challenge this view by showing that there are systematic individual differences in semantic representations of quantifiers. We selected five natural language quantifiers (*few*, *fewer than half*, *many*, *more than half*, and *most*) and collected binary truth value judgement data in an online quantifier verification experiment. Using a Bayesian three-parameter logistic regression model, we separated three sources of individual differences: truth condition, vagueness, and response error. The k-means clustering of the parameters of the model revealed three subgroups of participants with different semantic representations of quantifiers and a different organization of the mental line of quantifiers. This finding supports the view that logical words, like content words, are sensitive to individual differences, and hence it challenges the traditional view on meaning.

Keywords

Quantifiers; vagueness; meaning representations; Hierarchical Bayesian Model

1 Introduction

Needless to say, humans differ in their cognitive abilities. Similar to other cognitive domains, individual differences are also present in natural language processing (?, ?). They have been studied extensively in many natural language categories, including gradable adjectives (?, ?, ?), semantic categorizations of nouns (?, ?, ?, ?), probability terms (?, ?, ?, ?), and presuppositions projection (?, ?). In this paper, we investigate individual differences in natural language quantifier representations. Quantifiers, such as *many*, *few*, *most*, *some*, and *at least 5*, are functional words used to express quantities. They have drawn the attention of researchers from different fields ranging from logic (?, ?, ?) to formal semantics (?, ?, ?) to cognitive science (?, ?; see ?, ? for review). They have been studied mostly in the verification paradigm (e.g., ?, ?, ?, ?, ?), in which participants have to decide if a quantified sentence is true in a given context. While the studies of quantifiers are common, individual differences in their use have gained somewhat less attention in the literature. The dominant perspective on quantity words considers meaning representations to be rigid logical forms (e.g., ?, ?, see also ?, ? for discussion). A growing body of evidence (?, ?, ?, ?) questions this traditional view and calls for incorporating individual differences in the domain of quantifiers.

Individual differences in quantifiers may come from three different sources. The first source are differences in general cognitive abilities, e.g., working memory (?, ?, ?) or executive functions (?, ?). For example, the accuracy and speed of verification of proportional quantifiers depend on working memory capacities (?, ?, ?, ?) and cognitive control (?, ?, ?). The second source of individual

differences could lie in the choice of verification strategies (?, ?). For example, ? (?) showed that some participants prefer to use a precise strategy while verifying *most*, and others choose an approximate strategy. Moreover, strategy preference depends on the context (?, ?). Finally, the third source of individual differences could be different meaning representations of quantifiers where individuals assign different truth values to the same sentence (?, ?). ? (?) divided participants into two groups based on their truth value evaluation of the underinformative sentence "*Some* As are B" when in fact *all* As were B. The group of so-called pragmatic responders judged the underinformative sentence as false and logical responders as true.

While the first two sources of individual differences are compatible with the formal semantics perspective on language, the last one is a conundrum for linguistic tradition. At first glance, it seems that rational subjects cannot assign different truth conditions to logical words such as quantifiers. Moreover, the alignment of meanings between speakers is a crucial requirement for successful communication. ? (?) referred to this principle as to *correlation of meanings*. This requirement puts constraints on how much meaning of the single word can vary across speakers.

To summarise, on the one hand, we have evidence from many different linguistic domains (?, ?, ?, ?, ?, ?, ?, ?, ?, ?) for between-participants variability in meaning representations. On the other hand, we have a constraint on variation in meanings that arises from the communication pressure. In this paper, we show that individual differences in quantifier representations are tangible. We aim to answer three questions regarding individual differences. First, we tested the ranges of variability in meaning representations of quantifiers. We argue that even though there is individual variation meaning representations, this variation is constrained. Previous studies (?, ?, ?) showed that participants can form subgroups with different meanings. We propose to investigate the subgroups that would align the meaning of a quantifier within-group but have different meanings between groups. We will test *how many subgroups of participants with different meanings we can identify*.

Second, quantifiers convey information about quantities, and therefore they could be ordered with respect to the indicated amount on a mental line (e.g., ?, ?, ?). *None* is a lower bound on a mental line and expresses quantity of zero and *all* is an upper bound that expresses quantity of 100%. *Few* is intuitively more than *none* and less than *many*, while *many* could be less than *all*. The variability in meanings between participants poses a question on *how the meanings of quantifiers are interrelated at the subject level*. Given the correlation of meanings principle (?, ?), we hypothesize that participants should share the meanings of quantifiers at least to a certain extent. While they can choose various positions of a quantifier on a mental line, their choice for one quantifier constraints their choices for other quantifiers (?, ?). In this paper, we will test *whether the order of quantifiers on the mental line is consistent among participants*.

Third, we want to separate potential sources of individual variation in quantifier representations from variability in task performance. For example, we considered that participants may assign different truth conditions to the same

quantifier. However, they can also make mistakes while performing the task (e.g., due to attentional lapses) that are unrelated to meaning representation. The third question regarding individual differences is, therefore: *How can we separate the source of individual differences related to quantifier representation from variability in task performance*?

To answer these questions, we analyzed data from a quantifier verification task, in which participants were asked to judge the truth of a quantified sentence based on information about proportion. We modeled the choices using a logistic regression model to separate the individual differences meaning representations from variability in performance of the verification task. Then, we clustered participants based on their truth conditional representations to establish subgroups with aligned meanings. Our work continues the tradition of using computational modeling to better understand cognitive representations. Computational modeling has previously been successfully applied to test competing semantic theories (e.g., ?, ?) and to distinguish between different sources of individual differences in language processing (?, ?, ?). Moreover, computational modeling allows the investigation of qualitatively different effects in experimental data (?, ?, ?, ?, ?, ?). In the following section, we explain how we operationalized the meaning representations of quantifiers on model parameters.

1.1 Individual differences in truth conditional quantifier representations

Traditionally, formal semantics analyses the meaning of quantifiers in terms of truth conditions (e.g., Generalized Quantifier Theory, ?, ?, ?). The truth condition of a natural language quantifier specifies a threshold above or below which the quantifier is true[1]. For example, the quantifier *most* in the sentence "*Most* of the As are B" is true ($most(A, B) = 1$), if the intersection of sets A and B ($|A \cap B|$) is greater than the intersection of sets A and not B ($|A \cap \neg B|$). Example **??** shows truth conditions for quantifiers: *most*, *more than half*, *fewer than half*, *many*, and *few*.

Example 1.1.

1. *Most* $(A, B) = 1$ iff $|A \cap B| > |A \cap \neg B|$
2. *More than half* $(A, B) = 1$ iff $|A \cap B| > |A|/2$
3. *Fewer than half* $(A, B) = 1$ iff $|A \cap B| < |A|/2$
4. *Many* $(A, B) = 1$ iff $|A \cap B| > n$, where n is some cardinality or proportion
5. *Few* $(A, B) = 1$ iff $|A \cap B| > n$, where n is some cardinality or proportion

[1] In this paper, we focus only on quantifiers with one threshold. Some quantifiers can have two or more thresholds, e.g., *between 3 and 6* has two thresholds, 3 and 6.

Some quantifiers like *at least 5* have clear truth conditions with the threshold equals 5. Other quantifiers, like *many*, have various thresholds depending on the context (?, ?). Moreover, *many* and *few* are ambiguous between cardinal and proportional reading (?, ?). According to cardinal reading, the threshold is a fixed number e.g., "*Many* students passed the exam" means more than 40 students. Proportional reading of *many*, in turn, refers to *many* as more than some proportion, e.g., "*Many* of the students passed the exam" means more than 40% of the students. In this paper, we focus only on proportional readings of *few* and *many*.

Individual differences seem likely in context-dependent quantifiers such as *many* and *few*. ? (?) showed that different speakers have different meanings of these quantifiers. More surprisingly, ? (?) found individual differences in the quantifier *most* within the experimental paradigm downplaying the role of context. This finding questions the underlying assumption of many studies (?, ?, ?, ?) that participants have a dominant representation of *most*.

In the current paper, we adopt the truth-conditional semantics of quantifiers. Previous studies have shown (?, ?, ?) that truth-conditional semantics are suitable to model production of quantifiers. Building up on these findings, we propose a model of the quantifier verification task. We operationalized the truth-conditional representation of quantifiers as a threshold parameter around which the truth value of quantified sentences flips. Moreover, we extend the truth-conditional semantics framework by allowing for the individual differences in thresholds. We performed a cluster analysis to systematically investigate the subgroups of participants.

1.2 The order of quantifiers on the mental line

The meanings of the quantifiers considered here highly overlap. They constitute the sets of alternatives for each other (e.g., ?, ?, also cf. ?, ?). The first studies that looked into the order of quantifiers on a scale tried to link quantifiers with proportions for psychometric purposes (?, ?, ?). They found that participants were less consistent in the usage of some quantifiers than others. For example, low-magnitude quantifiers (e.g., *few*, *several*) were more context-dependent than high-magnitude quantifiers (e.g., *many*, *most*, ?, ?).

Recently, ? (?) have shown that quantifiers can be ordered on the mental number line. However, the distance between meaning representations does not have to be equal (see also ?, ?). For example, low-magnitude quantifiers (e.g., *few*, *almost none*) were more separated from each other and had sharper representations than high-magnitude quantifiers (*almost all*, *most*, *many*). Based on participants' semantic similarity judgements on the 7-point Likert-like scale, they also showed that some quantifiers are semantically more similar than others. For example, *many* is more similar to *most* than to *few*. Moreover, ? (?) showed that the change in the meaning representation of one quantifier (e.g., *many*) affects the threshold of the polar opposite quantifier (e.g., *few*). This effect is present in the reinforcement learning paradigm (?, ?) or via adaptation during exposure (?, ?).

The above studies did not account for the individual differences in thresholds. In contrast, we investigated the relationship between quantifier meanings taking into account the between-subjects variability in thresholds to shed more light on how quantifiers are represented on the mental number line on the individual level.

1.3 Other sources of between-subject variability

The truth-conditional semantics specify a threshold for each quantifier. In this view, the threshold is a point on a mental line around which the truth value of the quantifier flips. So far, we assumed that the position of the threshold on the mental line varies between quantifiers and participants. However, quantifiers can differ in how precise their meaning boundaries are. We will refer to this phenomenon as vagueness. The role of vagueness in natural language has been extensively debated in the linguistic and philosophical literature (see in ?, ?, ?, ?). In a nutshell, vagueness express the intuition that meaning boundaries are gradable. Rather than having a truth-conditional representation given as a rigid point on a mental line, participants can change their threshold slightly from trial-to-trial. Therefore, we included a separate parameter in our model to test the effect of vagueness independently of the threshold.

The borderline cases constitute a key characteristic of vagueness. Quantifiers like *more than half* or *fewer than half* have clear threshold, namely half. In contrast, the thresholds for *many* and *few* are not specific due to the borderline cases. For example, if we agree that the sentence "*Many* of the students failed the exam." is true when 20% of students failed, we will also probably agree that the sentence is true when 19% failed. Thus, the threshold for accepting a statement as true for *many* and *few* is fuzzy even given a fixed context (?, ?).

Some studies showed that *most* is also vague (?, ?, ?). ? (?) claimed that *most* and *more than half* are represented on different underlying scales. *More than half* has to be represented on the ratio scale, while *most* requires only the semiordered scale. The latter scale allows less precise comparisons, and, therefore, the meaning of *most* is more variable. Moreover, ? (?) showed that participants were less consistent about their threshold for *most* than for *more than half*.

Both threshold and vagueness can give rise to individual differences. Participants might disagree about the threshold for a given quantifier, as well as, they might have a different level of certainty about the exact position of the threshold on the mental line. In addition, while verifying quantified sentences, participants sometimes make errors. The response error in quantifier verification tasks depends on quantifier complexity (?, ?), working memory demands (?, ?), or polarity (?, ?, ?). For example, participants process negative quantifiers slower and with a higher error rate than when they process positive quantifiers (?, ?, ?, ?).

The individual differences in task performance hinder the interpretation of behavioral data. For example, previous studies (?, ?) argued that the same overall proportion of errors in the verification task for *most* and *more than half*

speaks in favor of the same truth conditions of these quantifiers. In contrast, another study (?, ?) showed that the accuracy for *most* is lower than for *more than half* when the proportion is slightly above 50%. ? (?) interpreted this asymmetry as a difference in quantifier pragmatics rather than truth conditions. Finally, ? (?) showed that the accuracy for *most* is lower than for *more than half* relative to their estimated thresholds. These studies show that the response error is a crucial measure of participants' performance. However, its interpretation is not unequivocal. We included the additional response error parameter in our model to account for differences in accuracy between negative and positive quantifiers and to disentangle the measure of error from the measures of threshold and vagueness.

Thus far we specified three parameters that could be sensitive to individual differences: threshold, vagueness, and response error. Moreover, we suggested that the behavioral measures in the linguistic task may reflect the interplay between parameters. For example, we can imagine that participants may have the same truth conditions for *most* and *more than half* and yet perform worse while verifying *most* because of other reasons. Moreover, participants may make more errors when verifying vague quantifiers. Response errors and vagueness, in turn, can lead to variability in thresholds. These interdependencies might lead to confounds when interpreting the experimental data. Therefore, it appears to be crucial to tease apart the effect of each parameter. However, as far as we know, the relationship between threshold, vagueness, and response error has not been systematically investigated.

1.4 Current study

To test the individual differences in quantifier representations and the relationship between the meanings of different quantifiers, we asked participants to judge the truth of a sentence involving a quantifier against the proportion given as a number between 1% and 99%. We chose proportional quantifiers from three groups: quantifiers with sharp meaning boundaries (*fewer than half* and *more than half*); vague and context-dependent quantifiers (*few* and *many*); and one quantifier that falls between these groups (*most*). We fit a computational model to the response data to estimate three parameters for every quantifier and participant. We predicted that each model parameter will capture a different aspect of participants behavior. We hypothesised that threshold parameter will capture a quantifier specific truth conditional representation. We predicted also a higher value of the vagueness parameter for vague quantifiers and that participants would make more mistakes while verifying the negative quantifiers.

To establish the subgroups of participants with different meanings, we performed a cluster analysis on the threshold parameter. We predicted that all participants would have the same threshold for *fewer than half* and *more than half* because these quantifiers already refer to the threshold, namely half. In contrast, we predicted that we would find between-clusters variability in thresholds for vague quantifiers like *most*, *many*, and *few*. We also hypothesized that only vague quantifiers would contribute to clustering on the threshold. In addi-

tion to clustering on threshold parameters, we performed two cluster analyses on vagueness and response error to see which quantifiers contributed to clustering (see Appendix **??** and **??**). We expected that *few*, *many*, and *most* would contribute to clustering on vagueness and negative quantifiers to clustering on response error.

To address our second research question, we explored how the meaning of one quantifier relates to other quantifiers. In contrast to previous studies (?, ?, ?, ?, ?, ?), we also looked into the order of quantifiers on a mental scale on the individual level within the clusters of participants.

Finally, we tested the relationship between model parameters. We wanted to separate the between-participants variability in truth conditions (thresholds) from vagueness and response error by introducing three parameters into our model. To justify the inclusion of all model parameters we tested whether they are not highly correlated. This analysis was exploratory in nature.

Before estimating the parameters of the computational model from the data of our experiment, we explored the effects of the three parameters on potential data patterns. In particular, we wanted to separate vagueness and response error effects because they both lead to response variability. Response errors are a result of additional cognitive processes and should therefore occur after the participants compare the proportion given in the experimental trial to their internal threshold. As such, response errors are independent of proportion. In contrast, vagueness adds noise to the decision process. The noise is greater around the participants' threshold. As a result, the internal threshold shifts from trial-to-trial. As such, vagueness depends on the proportion.

Figure **??** presents how we conceptualized threshold, response error, and vagueness parameters. We chose the quantifier *more than half* for illustration. For the ideal responder, the proportion of 'true' responses below 50% is zero, and above 50% is one. The logistic curve has a sharp shape indicating a rapid shift from false to true responses at the threshold. When the responses are affected by vagueness, the perceived threshold varies from trial to trial, and the logistic curve increases gradually. The response error, in turn, does not change the shape of the response curve. Instead, it lowers the probability of the true response above the threshold and increases the probability of the true response below the threshold equally for all proportions. We also plotted the combined effect of response errors, vagueness, and threshold.

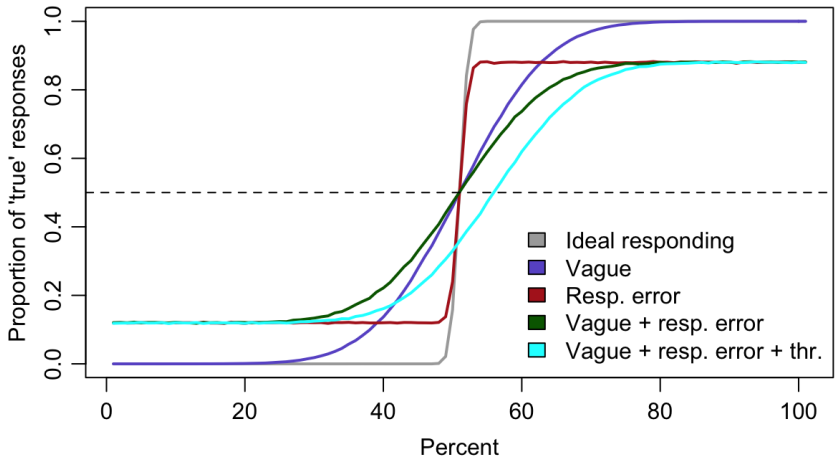


Figure 1: Predicted logistic curves under different threshold (thr.), response error (resp. error), and vagueness (vague) parameters. The dashed line indicates the 0.5 proportion of true responses. The percentage for which the logistic curve crosses the dashed line is the threshold.

2 Method

2.1 Data Availability

The data and analysis code is available at https://github.com/jstbcs/plingquant.

2.2 Participants

We recruited 90 participants via the online recruitment platform Amazon Mechanical Turk. We excluded 19 participants based on three exclusion criteria. Firstly, we excluded 11 participants who had 50% or more reaction times faster than 300 ms. Secondly, we excluded 7 participants who failed to obey the monotonicity of quantifiers. We defined the monotonicity criterion in the following way: for positive quantifiers (*many*, *most*, and *more than half*) we expected the probability of providing the true response to increase with increasing proportion. The opposite effect should hold for negative quantifiers. To apply monotonicity criterion, we fitted the generalized linear model to participants' response data with the proportion as a predictor and with by-subject random intercept and slope for proportion (*glmer* R function, ?, ?). We excluded participants, who had a negative slope for positive quantifiers or a positive slope for negative quantifiers. Finally, we excluded 1 participant, who took part in a similar

experiment. These exclusions meant that we included 71 participants (47 male, age $M = 35$, range: 22–59) in the final sample. Subjects gave informed consent prior to participation in the experiment. The study was approved by the Ethics Committee of the University of Amsterdam's Faculty of Humanities.

2.3 Experimental Design and Procedure

In our experiment, participants had to indicate whether the sentence with the quantifier: *most*, *many*, *few*, *fewer than half*, or *more than half* was true or false based on the sentence containing a proportion ranging from 1% to 99% (excluding 50%). We did not include the proportion 100%, because ? (?) showed that *most* has an upper bound on meaning and using it with 100% proportion is not accepted, although it is highly accepted with 99%. The upper bound of *most* could cause a divergence in the logistic function which we used in our model. We did not include 50%, because this proportion could be confusing for *more than half* and *fewer than half*.

While *most*, *more than half* and *fewer than half* have a proportional interpretation (?, ?), as explained above, *many* and *few* are ambiguous between cardinal and proportional reading (?, ?). For example, *many* could mean more than a certain number (cardinal reading) or more than a certain proportion (proportional reading, see Example **??**). We used explicit partitive 'of the' and present proportions as a percentage for all quantifiers to ensure the proportional reading and avoid confounds for ambiguous quantifiers. Moreover, by using the percentage format we enforced the precise comparison between proportion and the threshold. In this way, we minimized the differences between quantifiers in verification strategies. For example, in some experimental paradigms *most* is verified using approximate strategy (?, ?), while in others mixtures of strategies is used (?, ?).

The experiment started with a short training block to familiarize participants with the procedure. Next participants completed the 250 trials (50 per quantifier) in randomized order. At the end of the experiment, participants provided basic demographic information. Each trial of the experiment consisted of two sentences displayed on separate screens. The first sentence containing the quantifier was of the form "[*Most/Many/Few/More than half/Fewer than half*] of the gleerbs are fizzda." To read this sentence participants had to press the arrow down key and keep it pressed. When they advanced to the next screen, they read a sentence containing proportion e.g., "20% of the gleerbs are fizzda." Participants had to provide a response by pressing the right or left arrow keys corresponding to true or false judgment (counterbalanced between participants).

In our experiment, we used pseudowords generated from 50 English six-letters nouns and adjectives using *Wuggy* software (?, ?). We used pseudowords to avoid pragmatic effects associated with quantifiers. The original words were controlled for frequency (*Zipf* value 4.06, ?, ?). A native English speaker assessed the pseudowords in terms of how well they imitated English words.

2.4 Data pre-processing

We excluded trials with response times shorter than 300ms and longer than 2500ms (similar cut-offs to ?, ?). Altogether, we excluded 6% of trials. To be able to fit the same logit model to all quantifiers we flipped the true and false responses for *few* and *fewer than half*.

2.5 Computational Model

The logistic regression model is suitable for modeling the threshold variability (?, ?, ?). The model assumes that the probability that participants verify a statement as true or false depends on the proportion that was presented on a particular trial and the values of the logistic function parameters asymptote, midpoint and scale:

$$response \sim \frac{asymptote}{1 + exp(midpoint - proportion)/scale} \tag{1}$$

To accommodate individual differences and differences between quantifiers in the model, we used a three-parameter logistic regression model inspired by Item Response Theory (IRT). IRT determines the relationship between an individual's trait and the probability of providing a correct response for a given item (?, ?, ?). This relationship is expressed by the Item Response Function, which maps the IRT parameters (difficulty, discrimination, and guessing) onto the logistic function. The three-parameter model has a difficulty parameter, which determines the level of an individual trait necessary to provide a correct response (midpoint), a discrimination parameter that determines the steepness of the logistic curve (scale), and a guessing parameter that can adjust the logistic curve asymptotes.

In our model, the threshold corresponds to the difficulty parameter, vagueness to the discrimination parameter, and response error to the guessing parameter from the IRT model. We used a hierarchical Bayesian model to estimate the parameters for each participant-quantifier combination. To fit the model, we used the *rstan* package in R (?, ?) with 6 chains, 750 warm up iterations per chain and 2500 iterations per chain.

The model was specified in the following way. Let i indicate participants, $i = 1, ..., I$, j indicate the quantifier, $j = 1, ..., 5$, and k indicate the trial for each quantifier, $k = 1, ..., K_{ij}$. Then Y_{ij} is the i-th participant's response to the j-th quantifier in the k-th trial, and $Y_{ijk} = 1$ if participant indicated true, and $Y_{ijk} = 0$ if participant indicated false. Then, we may model Y_{ijk} as a Bernoulli, using the logit link function on the probabilities:

$$Y_{ijk} \sim Bernoulli(\pi_{ijk}) \tag{2}$$

where the probability space of π maps onto the μ.

$$\pi_{ijk} = \gamma_{ij} + (1 - 2\gamma_{ij}) logit^{-}1(\mu_{ijk}) \tag{3}$$

The additional parameter γ_{ij} determines the probability of making a response error on either side of the threshold, namely erroneously saying true, or erroneously saying false. Each participant-quantifier combination has its own response error parameter estimate. The parameter μ_{ijk} has a linear model explication:

$$\mu_{ijk} = \frac{c_{ijk} - \beta_{ij}}{\alpha_{ij}} \tag{4}$$

where c_{ijk} indicates the percentage centered at 50%, parameters β_{ij} indicate the threshold, and parameters α_{ij} correspond to the vagueness of the quantifier.

We defined prior probabilities on response error (γ), threshold (β), and vagueness (α) parameters:

$$\gamma_{ij} \sim Beta(2, 20) \tag{5a}$$

$$\beta_{ij} \sim Normal(\delta_j, \sigma_j^2) \tag{5b}$$

$$\alpha_{ij} \sim log - Normal(\nu_j, \sigma_{\alpha_j}^2) \tag{5c}$$

$$\nu_j \sim Normal(0, 5^2) \tag{5d}$$

$$\sigma_{\alpha_j}^2 \sim Invers - Gamma(2, 0.2) \tag{5e}$$

$$\sigma_j^2 \sim Invers - Gamma(2, 0.2) \tag{5f}$$

$$\delta_j \sim Normal(0, 5^2) \tag{5g}$$

The hierarchical nature of the distributions for α_{ij} and β_{ij} indicate that we estimated the effect of threshold and vagueness for each participant under the assumption that they had a common mean and variance. The vagueness and threshold priors were fairly uninformative to avoid the inclusion of incidental constraints. Vagueness (α_{ij}) came from a log-normal distribution to ensure only the positive estimates. Its mean (ν_j) had a normal distribution, and its variance ($\sigma_{\alpha_j}^2$) was drawn from Inverse-Gamma distribution, as this distribution is typically used to model variance. For the thresholds (β_{ij}) we used a normal distribution with a common, normally-distributed mean (δ_j) and the same variance distribution (σ_j^2) as for α_{ij}. The response error (γ_{ij}) came from a more informed distribution with most of its mass below an error rate of 20% for each true and false response[2].

2.6 Cluster analysis

We ran an exploratory cluster analysis for the threshold parameter[3] estimating the clusters using the k-means clustering method (*kmeans* function in R, ?, ?). We determined the optimum number of clusters by using the elbow plots and Silhouette width. We chose the k-mean clustering because it could be apply to

[2]To reduce the complexity of the model, we did not use hierarchical modeling for response errors.

[3]See cluster analysis for vagueness and response errors in Appendix.

relatively small data set and does not impose too much structure into the data (?, ?).

2.7 Linear Discriminant Analysis

To assess the contribution of the model estimates to the clustering, we performed a linear discriminant analysis (LDA). We used the stepwise procedure Wilks' lambda assessment (*greedy.wilks* function in R package *klaR*, ?, ?) to determine which variable contributed significantly to cluster formation. Next, we ran the LDA (*lda* function in R package *MASS*) to test how accurately the selected variables could predict the clusters. To validate the LDA, we ran a leave-one-out cross validation.

3 Results

3.1 Estimated parameters

The estimated model parameters are shown in Table **??**. Figure **??** shows the estimated item response curves for each participant-quantifier combination; the overall response curves for the quantifiers are represented by the bold, colored lines. We found greater individual variation in thresholds for *most*, *many* and *few*, compared to *more than half* and *fewer than half*. At the group level, quantifier thresholds were represented in the following order (Friedman test $\chi^2(4) = 134$, $p < 0.001$, moderate effect size $W = 0.47$): *few* had the lowest threshold, followed by *many*, then were *fewer than half* and *more than half*, and *most* had the highest threshold (pairwise comparison, Wilcoxon Signed Rank Test with Bonferroni correction).

Table 1: Mean (*SD*) parameters of individual participants for each quantifier, and additionally for threshold parameter the percent corresponding to mean thresholds.

	Threshold	Vagueness	Response error
Few	-.103 (.073), 39.7%	.016 (.001)	.062 (.042)
Fewer than half	-.006 (.027), 49.4%	.002 (.00004)	.074 (.047)
Many	-.061 (.094) 43.9%	.019 (.003)	.048 (.024)
More than half	.001 (.012) 50.1%	.001 (.00003)	.042 (.019)
Most	.029 (.056) 52.9%	.009 (.001)	.047 (.024)

The quantifiers *fewer than half* and *more than half* were the least vague as indicated by the steep response curves in Figure **??**. Moreover, *few* was more vague than *fewer than half* ($V = 2556$; $p < 0.001$), *many* was more vague than *more than half* ($V = 2556$; $p < 0.001$), *many* was more vague than *most* ($V = 2556$; $p < 0.001$), and *most* was more vague than *more than half* ($V = 2556$; $p < 0.001$), p - values based on Wilcoxon Signed Rank Test. We also found

that *fewer than half* had a greater response error than *more than half* ($V = 2323$; $p < 0.001$), and *few* had greater response error than *many* ($V = 1809$; $p = 0.002$), p - values based on Wilcoxon Signed Rank Test. As predicted, the vague quantifiers had a higher value of vagueness parameter and negative quantifiers had higher value of response error parameter.

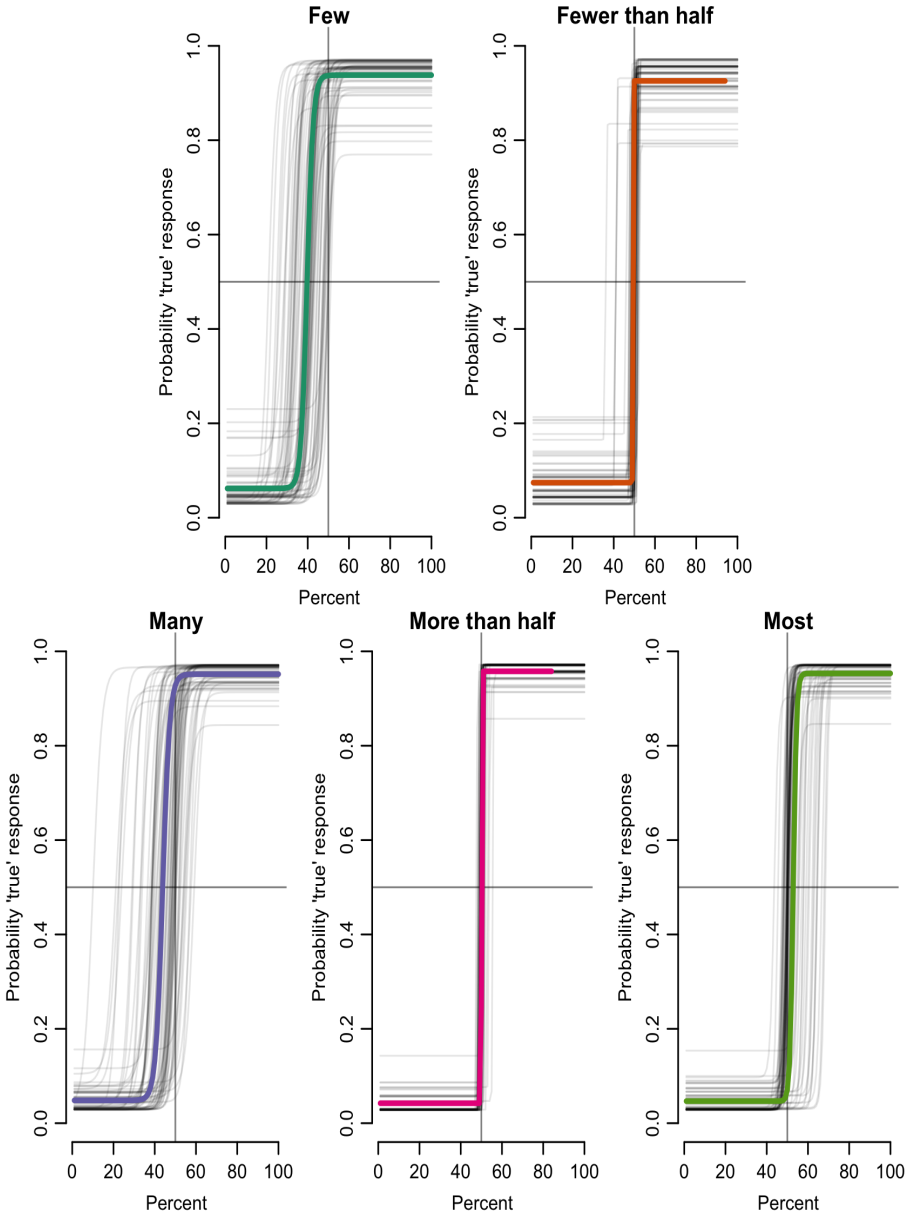


Figure 2: The logit curves estimated for each quantifier. The color lines indicate the mean curves.

3.2 Individual differences in thresholds

In the next step, we explored the associations between threshold parameter across quantifiers to reveal potentially systematic patterns (see Figure **??**). The

Pearson correlations between thresholds were negligible or weak. The lack of significant correlations between thresholds of *fewer than half* and *more than half* and other quantifiers is not surprising given a little between-participants variability in truth conditions of *fewer than half* and *more than half*. The lack of correlations between thresholds of vague quantifiers requires a comment. A significant positive correlation between thresholds would indicate individual differences in preference toward placing the thresholds high or low on a mental scale for all quantifiers (e.g., the higher threshold for *many*, the higher threshold for *few*). A significant negative correlation would indicate individual differences in distance between thresholds (e.g., a high threshold for *many* and a low threshold for *few* indicate greater distance, while a low threshold for *many* and a high threshold for *few* indicate smaller distance). Our analysis did not reveal systematic patterns. Therefore, it gives additional reason for the cluster analysis, because the lack of correlation might be caused by different position of thresholds in the subgroups.

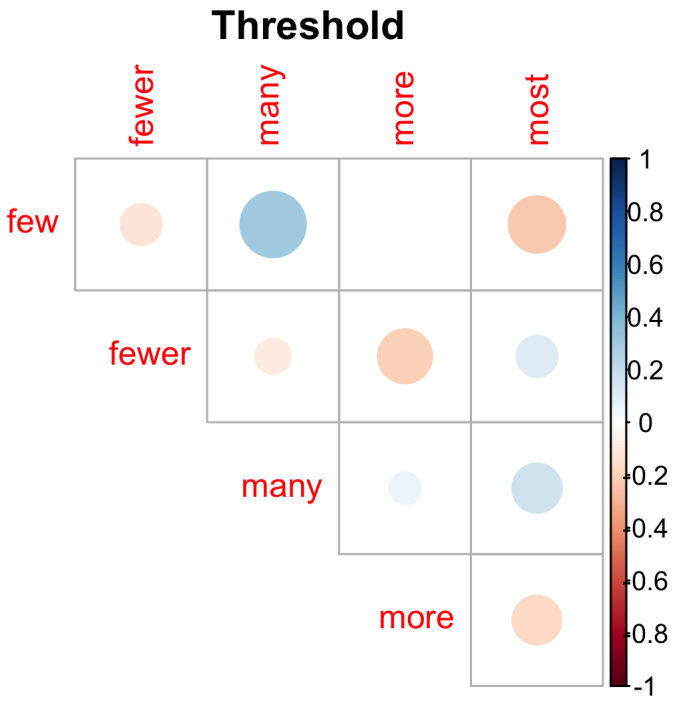


Figure 3: The Pearson correlations between thresholds (significance level *** 0.001, ** 0.01,* 0.05). The p - values were adjusted using the Bonferroni correction.

3.2.1 Cluster analysis results

Following the correlation analysis we performed the k-mean clustering on the threshold parameter from all quantifiers. This analysis addressed our first research question concerning the subgroups of participants. We hypothesized that participants will cluster into groups with aligned meanings.

The methods to determine the optimum number of clusters for threshold gave ambiguous results. The elbow plot indicated 3 or 4 clusters, while the Silhouette method preferred 5 clusters. We chose the simplest solution, comprising 3 clusters, because the additional clusters consisted of only 4 participants, making interpretation difficult. The three clusters were indistinguishable for the quantifiers *fewer than half* and *more than half*, but differed substantially in thresholds for the quantifiers *few*, *many*, and *most*.

The first cluster ($N = 13$) consisted of participants with a higher mean threshold for *most* and a lower threshold for *few*, the second cluster ($N = 34$) included participants who had thresholds for all quantifiers close to 50%, and the last cluster ($N = 24$) consisted of participants who had similar a mean threshold for *few* and *many* (see Table **??**).

Table 2: Mean (*SD*) threshold parameter in each cluster and percentage corresponding to mean thresholds, 3-cluster solution.

Quantifier	Cluster 1 ($N = 13$)	Cluster 2 ($N = 34$)	Cluster 3 ($N = 24$)
Few	-.15 (.05)	-.05 (.04)	-.15 (.07)
	35%	45%	35%
Fewer than half	.001 (.01)	-.012 (.03)	-.002 (.02)
	50.1%	48.8%	49.8%
Many	.014 (.06)	-.022 (.04)	-.16 (.09)
	51.4%	47.8%	34%
More than half	-.00006 (.006)	.002 (.01)	.0007 (.01)
	49.99%	50.2%	50.07%
Most	.10 (.05)	.009 (.03)	.02 (.05)
	60%	50.9%	52%

3.2.2 Linear Discriminant Analysis results

For thresholds, as expected, we found that only vague quantifiers contributed to the clustering: *many* ($\lambda = 0.42$, $p < 0.001$), *few* ($\lambda = 0.24$, $p < 0.001$), and *most* ($\lambda = 0.16$, $p < 0.001$). Figure **??** shows the combined effect of the three quantifiers on the clustering. The LDA accuracy in classification into Clusters 1 to 3 based on thresholds for *many*, *few* and *most* was 97%, and the leave-one-out cross validation accuracy was 94%.

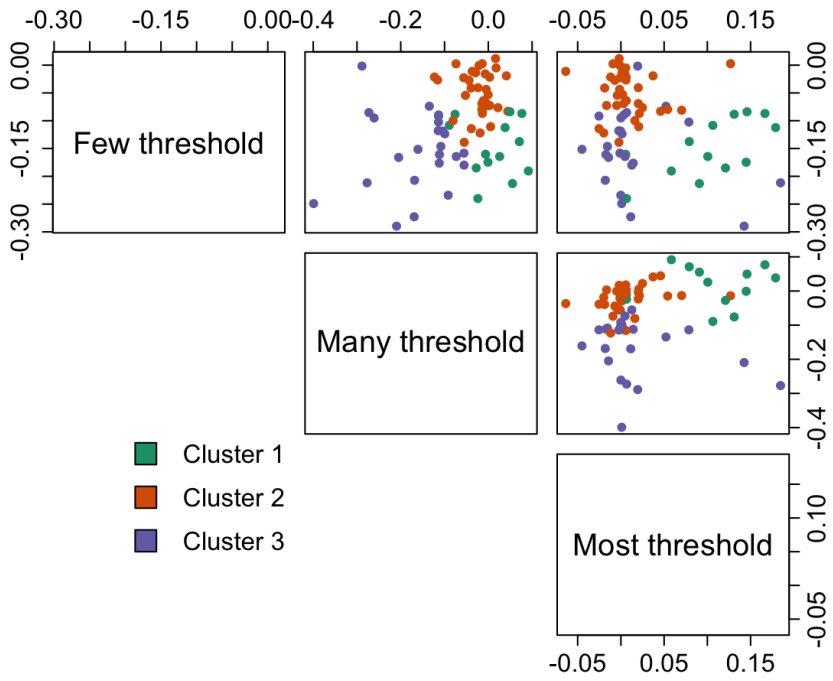
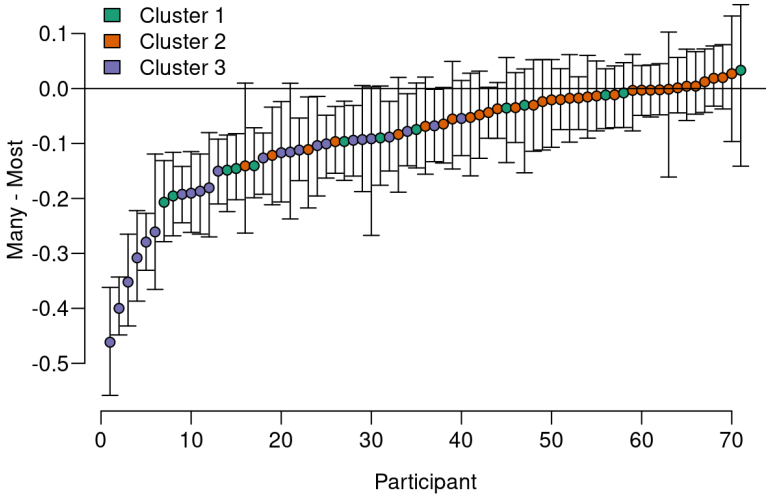


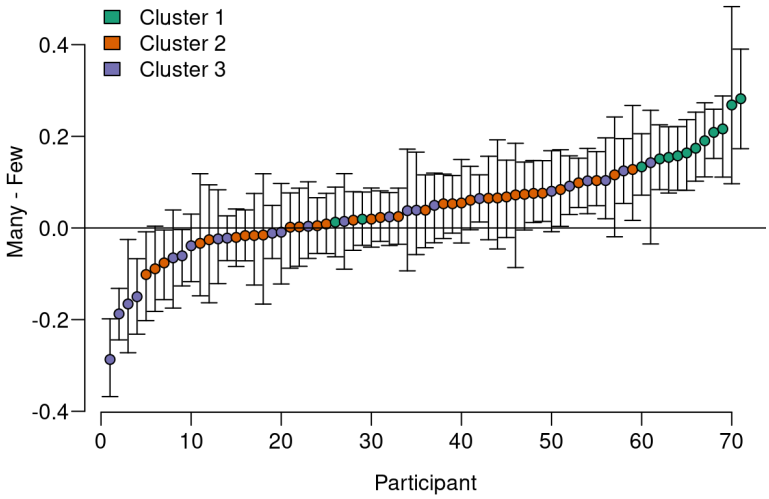
Figure 4: Three clusters for threshold based on *few*, *many*, and *most* parameters. The parameters' values of thresholds for three quantifiers (*few*, *many*, and *most*) that contributed to clustering are plotted against each other. Colors are used to indicate the cluster membership: Cluster 1 ($N = 13$) is indicated in green, Cluster 2 ($N = 34$) in orange, and Cluster 3 ($N = 24$) in purple.

3.3 Mental line of quantifiers

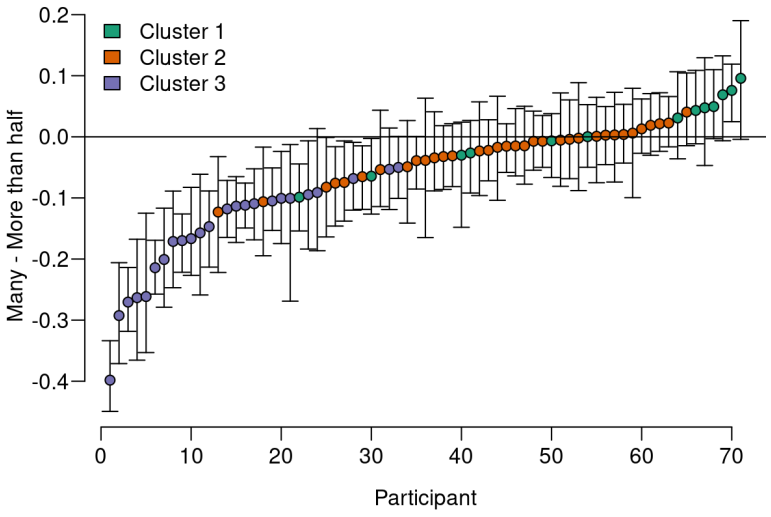
Our second research question concerned the order of quantifiers on a mental line in the subgroups of participants. We investigated whether all participants would have the same order of vague quantifiers on a mental line. Figure **??** shows that all participants had a lower or equal thresholds for *many* than for *most*. However, the distance between thresholds was higher in Cluster 3 than in other clusters. Figure **??** shows that the vast majority had a higher threshold for *many* than for *few*. The greatest distance between thresholds was in Cluster 1, while the smallest was in Cluster 3. Figures **??** and **??** show that all participants in Cluster 3 had a lower threshold for *many* than for *more than half* and *fewer than half*. In sum, the results indicate a rather stable order of quantifiers among three clusters, however, the distance between thresholds varies substantially among clusters.


Cluster 1
Cluster 2
Cluster 3
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
Many - Most
0
10
20
30
40
50
60
70
Participant

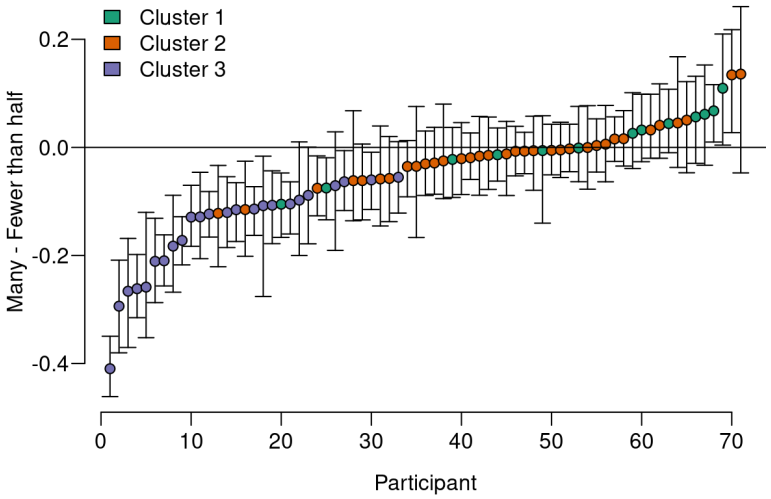
(a)


Cluster 1
Cluster 2
Cluster 3
0.4
0.2
0.0
-0.2
-0.4
Many - Few
0
10
20
30
40
50
60
70
Participant

(b)



(c)



(d)

Figure 5: Differences between individual thresholds ordered from smallest to largest. **a.** The difference between the threshold for *many* and *most*. **b.** The difference between the threshold for *many* and *few*. **c.** The difference between the threshold for *many* and *more than half*. **d.** The difference between the threshold for *many* and *fewer than half*. Colors are used to indicate cluster membership: Cluster 1 is indicated in green ($N = 13$), Cluster 2 in orange ($N = 34$), and Cluster 3 in purple ($N = 24$). The error bars indicate the 95% credible intervals.

3.4 The interrelationship between vagueness, threshold, and response error

Finally, we addressed our third research question about the interrelationship between vagueness, threshold, and response error. We correlated the model parameters for each quantifier (Figure **??**). Although this analysis was exploratory in nature, it gave additional rationale for our modeling choices. Significant high correlations between all parameters of our model would be problematic. It would mean that the parameters of our model do not capture unique source of variability in the data. Thus a more parsimonious model would be desired. Therefore, we wanted to test whether there were any systematic patterns of correlations between parameters across quantifiers. We found a significant negative correlation between threshold and vagueness for *few* (r = -0.33) and *many* (r = -0.31). We also found correlations between threshold and response error for *fewer than half* (r = -0.32), and response error and vagueness for *many* (r = 0.53) and *most* (r = 0.52). In general, the correlations did not reveal systematic patterns across quantifiers. The lack of systematic correlations between vagueness and response error parameters gives additional support to the choice to model these parameters as two separate mechanisms.

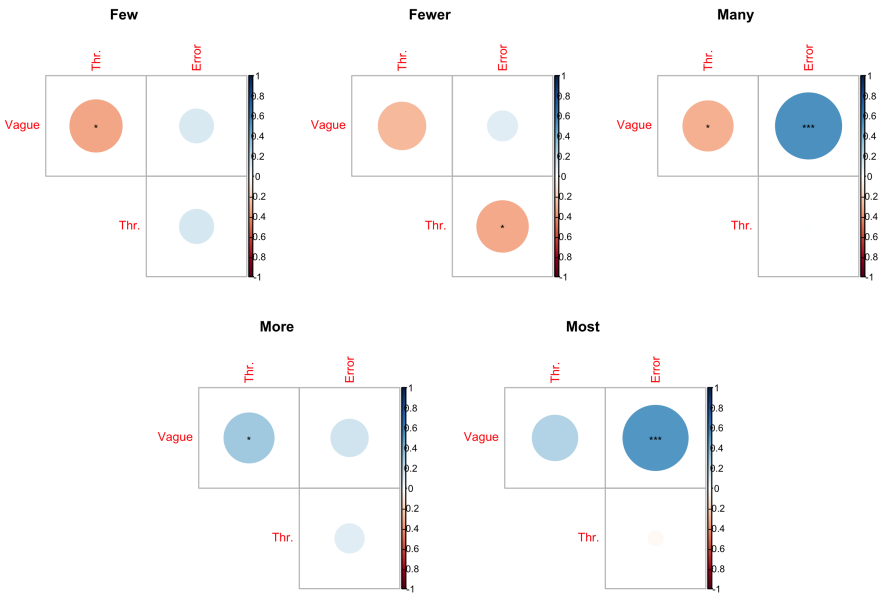


Figure 6: Correlations of parameters for each quantifier (significance level *** 0.001, ** 0.01, * 0.05). The p - values were adjusted using the Bonferroni correction.

4 Discussion

Previous studies showed that quantifiers are organized on a mental scale (?, ?, ?) and that participants use their internal threshold to verify proportional quantifiers (?, ?). However, little has been known about individual differences in the organization of quantifiers on the mental line. The main goal of this study was to identify subgroups of participants with substantially different meanings of quantifiers. We then investigated whether these subgroups organize quantifiers differently on the mental line.

First, we examined the correlations between quantifiers for the threshold parameter. This analysis revealed a lack of significant correlations and further motivated the analysis of the subgroups. We ran a cluster analysis on the threshold parameter of each quantifier. We identified three groups of participants with different constellations of thresholds across quantifiers indicating shifts in quantifier meaning. As initially predicted, quantifiers with sharp meaning boundaries, like *fewer than half* and *more than half*, did not contribute to clustering, and they had similar thresholds in all groups. In contrast, thresholds for *many*, *few*, and *most* varied considerably between clusters. In all groups, *most* had the highest threshold. However, the mean threshold varied between clusters. In the first cluster, the mean threshold was 60%, and in the second and third clusters, the mean thresholds were just slightly above 50%, at 51% and 52%, respectively. For *few*, participants in the first and third clusters had mean thresholds of 35%, and in the second cluster, the mean threshold was 45%. The mean threshold for *many* was the most diverse between groups. It ranged from 51% in the first cluster to 48% in the second cluster to and 34% in the third cluster. Together, these results show that individual differences in truth conditions of quantifiers and their position on the mental line are tangible.

Second, we investigated the order of quantifiers on a mental line. According to the correlation of meaning hypothesis (?, ?), participants may assign different thresholds for the same quantifier, however, they should agree on the order of quantifiers on a mental line. Our results support this hypothesis with the exception that *many* has a more flexible meaning than the other quantifiers.

The third goal of this paper was to look into the relationship between threshold, vagueness, and response error. As predicted, we found that quantifiers with fuzzy meaning boundaries had a higher vagueness value and that negative quantifiers had a higher response error value. We investigated the correlations across parameters for all quantifiers. The lack of systematic patterns in the data supports our modeling choices and justifies usage of the three-parameter model. We will discuss the implications of these findings in more detail in the following subsections.

4.1 The order of quantifiers on the mental line

Because we failed to find correlations at the group level between thresholds of different quantifiers, we zoomed into the mental line of the subgroups of participants. We observed that the clusters mainly differed in the range of the

mental line. Participants in the first cluster had the most stretched mental line, ranging between 35% and 60%, with a clear order of thresholds, where *few* was the lowest and *most* was the highest. In contrast, the second group had the most shrunk mental line, ranging between 45% and 51%. The mental scale of the last group stretched between 34% and 52%. Moreover, we found that the position of *many* on a mental line was the most flexible and varied substantially between subgroups.

? (?) found that although participants assigned different numerical equivalence to quantifiers, they were consistent about the order of the quantifiers. Moreover, according to the correlation of meanings principle (?, ?) the between-participants variability in meanings should be constrained by the communication need. Our findings indicate that the order of quantifiers across participants was rather stable, with the exception of *many*. We further discuss the possible explanation of this finding and the relationship between vague quantifier pairs: *few* and *many* (the polar opposites), and *many* and *most*.

4.1.1 *Many* vs. *few*

We found an asymmetry between *many* and *few* with regard to their positioning on the mental scale. The position of *many* on the mental scale was more flexible than the position of *few*. In the second and third clusters, the mean threshold for *many* was lower than for *more than half* and *fewer than half*, but in the first cluster, it was higher (see Table **??** and Figure **??**).

The flexibility of *many* on the mental scale cannot be explained by its context-dependency. First, in our experiment, we used an artificial context by introducing pseudowords. There was no reason for participants to have different expectations about the context. Second, based on the literature (?, ?), we may predict the opposite pattern of results. The low-magnitude quantifiers, such as *few*, are more context-dependent than high-magnitude quantifiers (?, ?). Moreover, they can change their threshold depending on the reference set (?, ?) and they are more separated from each other on the mental scale than high-magnitude quantifiers (?, ?).

We attribute the asymmetries in our study to competition between quantifiers. While *few* was less than 50% and *most* was more than 50% for all clusters, *many* had to compete with both quantifiers for a place on the mental line. As a result of this competition, *many* had a greater variation in threshold and, at least for some participants, was more vague. We observed two tendencies concerning the threshold of *many* (see Figure **??**). The first tendency was to either keep the threshold for *few* and *many* close together (Clusters 2 and 3) or far apart (Cluster 1). The second tendency was to either keep the threshold for *many* close to *most* (Cluster 2, and to some extent 1) or far from most (Cluster 3, see Figure **??**). Despite these tendencies, almost all participants had a higher threshold for *many* than for *few*, and all participants had a higher threshold for *most* than for *many*. Altogether, this finding shows that the position of *many* on the mental line is more flexible than the position of *few* and it explains the membership of the clusters. Nonetheless, in all clusters participants treated *few*

as less than *many*, and *many* as less than *most*.

4.1.2 *Many* vs. *most*

Previous studies and linguistic analysis (?, ?, ?) stressed similarities between *most* and *many*. First, ? (?) analyzed *most* as a superlative of *many* (*many*+est). This analysis predicts that *most* has to be more than *many*. Our data support this prediction. We showed that not only the mean threshold for *many* was lower than for *most* in all clusters, but also all participants had a higher threshold for *most* than *many* regardless of the cluster's membership (see Figure **??**). While all participants treated *most* as the superlative of *many*, the distance between thresholds of these quantifiers was different depending on the cluster. The greatest distance was in the third subgroup.

Second, ? (?) showed substantial overlap in the production of *most* and *many*. Both quantifiers cover comparable proportions on the mental scale. In contrast, our results show individual differences in the distance on the mental line between *most* and *many*. For example, in the third cluster, the mean threshold for *many* was considerably lower than the mean threshold for *most*, while in the second cluster, both thresholds were close to 50%.

Lastly, ? (?) found that *many* is used less frequently than *most*. We think that the quantifier's vagueness could be one of the sources of the difference in frequency. The high perceived vagueness of *many* lowers its usefulness. The more vague the quantifier, the less information it conveys. However, people try to be as informative as possible (?, ?) and therefore avoid the usage of uninformative quantifiers with very flexible meanings. This explanation generates a new prediction to test in future work: participants who perceive *many* as vaguer should also use it less often in a production experiment.

4.2 Relationship between model parameters within quantifiers

With regard to the relationship between the three model parameters, we did not find a systematic pattern of correlations among quantifiers (see Figure **??**). The only significant correlation between threshold and response error was for *fewer than half*. This correlation was, however, strongly affected by the outlier participants with a low threshold for *fewer than half* (see Figure **??** in Appendix **??**). Overall, this lack of correlation shows that the variation in thresholds reflects variation in the meaning representations and it is not an artefact of task performance.

We found small to moderate positive correlations between response error and vagueness parameters. As one could expect, the vaguer the quantifier, the more difficult it is to perform the task. However, the correlation was only significant for *many* and *most*. In addition, the lack of systematic correlations between vagueness and response error shows that they correspond to two different processes that should be modeled by separate parameters. Vagueness thus may correlate with response error for some quantifiers, but it can not be

equated with threshold-independent erroneous responding (cf. ?, ?). However, the correlation between vagueness and response error was more consistent (at least in direction) than between the other parameters. Relatedly, the overall magnitude of the vagueness parameters was quite small. One reason for both the correlations and the low magnitude might be an issue in identifiability in the model. In the future, it might be helpful to change the parameterization of the model to reduce overlap between the two parameters, for example by modeling asymmetries in vagueness around the threshold.

Finally, we found significant correlations between vagueness and threshold parameters for *many* and *few*, but, importantly, not for *most*. This finding challenges the explanation proposed by ? (?), according to which participants verify *most* using the approximate strategy (?, ?). Consequently, the verification of *most* is noisy around 50%. To reduce the noise, participants prefer thresholds significantly greater than 50%. This theory predicts that participants with higher thresholds for *most* will perceive it as a vaguer quantifier than participants with lower thresholds. In our model, we captured the noisiness of verification in the vagueness parameter. The lack of significant correlation between vagueness and threshold for *most* does not support Solt's explanation. Instead, it suggests that some participants assigned different truth conditions to *most* and *more than half*.

Overall, the results of the correlation analysis show that each parameter of the model captures a unique aspect of participant performance in the verification task. We interpret this finding as an empirical validation of the assumptions of our model (cf. Figure **??**).

4.3 Relationship of model parameters across quantifiers

We did not find significant correlations across quantifiers for the threshold parameter (see Figure 3), indicating that this parameter was quantifier-specific. In an additional analysis (see Appendix **??**), we tested whether the vagueness and response error parameters are quantifier-specific. Vagueness parameters were not correlated across quantifiers, indicating that similarly to threshold, it is quantifier-specific. In contrast, the response error parameter was significantly correlated across almost all quantifiers. This is in line with our considerations about the model – response error reflects cognitive abilities independent of the quantifiers, while vagueness contains information about quantifier meaning.

The correlations for the response error parameters were stronger between negative than positive quantifiers because of the greater variation in response error in negative quantifiers. Due to this variation, only negative quantifiers contributed to the clustering on response error. Response error, thus, reflects a combination of general task performance ability and specific difficulty in verification for negative quantifiers (?, ?, ?, ?). We noted that the cluster with a higher rate of response error was small ($N = 7$), probably because the task was generally easy. It would be worth testing whether the response error parameter contributes more to clustering in a more challenging task, for example, with visual displays instead of sentences.

4.4 Sources of individual differences

Our starting point for the investigation of individual differences in meaning representations of natural language quantifiers was the observation that language users can have different truth conditions for logical words (e.g., ?, ?). For example, previous studies (e.g., ?, ?) showed that two groups of speakers have different interpretations of the quantifier *some*. In this spirit, we demonstrated that this phenomenon is not limited to just one quantifier. We showed that there are three subgroups of participants with different thresholds for *many*, *few*, and *most*.

The question remains, however, how individual differences in thresholds emerge. We consider here a few possible explanations. First, we argue that individual differences are not due to the various verification strategies used by participants (cf., ?, ?). We think that this explanation is unlikely because the task design limited possible strategy choices. Participants verified the sentence with a quantifier by comparing their threshold to the proportion given as a number. Although the Approximate Number System (?, ?) could have interfered with the precise number system, it is rather unlikely that participants were unable to precisely compare proportions. In our task, there was no time pressure on the decision and the proportions were displayed on the screen for an unlimited period of time. We feel confident in rejecting the explanations based on the variability in verification strategies as a source of observed individual differences in thresholds in our study.

The individual differences in thresholds are also unlikely to be a result of the different cognitive abilities of our participants (e.g., ?, ?, ?, ?). We did not measure the working memory or executive function performance of participants, but our task was relatively easy and did not require much working memory or other cognitive function resources. Moreover, we included a response error parameter in our model, which accounted for variability in task performance (e.g., attention lapses or mistakes). We found that the majority of participants belonged to a low response error cluster (see Appendix **??**), indicating that they performed the task at a similar level of accuracy. Altogether, we conclude that the differences in thresholds between groups are due to different representations of the truth conditions of quantifiers.

Another possible explanation of individual differences in thresholds arises from studies on informativeness and pragmatic abilities. For example, ? (?) showed using the electroencephalography marker of semantic processing that participants with better pragmatic abilities are more sensitive to underinformative sentences compared to less skilled participants. According to this line of argumentation, participants in our study had the same meaning representations of quantifiers. The observed individual differences in thresholds thus are driven by a pragmatic mechanism.

While we can not rule out the pragmatic explanation of our result, we would like to point out some challenges to this interpretation. First, previous studies investigating the role of pragmatic abilities in the interpretation of underinformative sentences are inconclusive (cf., ?, ?, ?). Some studies (e.g., ?, ?) did

not find a correlation between pragmatic skills and sensitivity to underinformativeness. Secondly, it is difficult to provide a single pragmatic mechanism that would explain individual differences in thresholds of all quantifiers. For example, it has been argued that the higher threshold for *most* compared to *more than half* is a result of pragmatic strengthening (?, ?). Recent studies (?, ?) tested this prediction directly and challenged the pragmatic strengthening hypothesis. Moreover, the mechanism of pragmatic strengthening can not explain the individual differences in thresholds of *many* and *few*. The variability in thresholds for *many* and *few* is often explained by different expectations related to the contextual factors (e.g., ?, ?). However, in our study, the contextual information was fairly abstract and unlikely to elicit specific expectations. More research is needed to fully understand the complexity of the semantics-pragmatics interface and its effect on the structure of the quantifier mental line.

Finally, the individual differences may have emerged due to participants being sensitive to different pressures that shape the space of meanings. ? (?) put forward the idea that the space of natural concepts is optimally designed, meaning that it satisfies design criteria and constraints (see also ?, ?, for discussion). For example, the space of concepts should be informative but also parsimonious and learnable. Analogue criteria could be applied to the quantifier mental line. We suggest that the subgroups may differ in which criteria shape their space of meanings primarily, and as a result, they could apply different thresholds. Further studies would be needed to test this hypothesis. Moreover, future research should investigate how participants belonging to different clusters can communicate with each other. The adjustment of the threshold between participants called the semantic adaptation (e.g., ?, ?) is one of the possible mechanisms. Further studies should investigate the role of vagueness in semantic adaptation. The fuzziness around the threshold may facilitate the "meeting of minds" (?, ?) by leaving space for shifting the thresholds.

4.5 Conclusions

In the current study, we separated individual differences in meaning representations, such as vagueness and threshold, from general cognitive abilities reflected in a response error parameter using a Bayesian modeling approach. Based on the model's threshold parameters, we identified three clusters of participants assigning different meanings to vague quantifiers such as *most*, *many*, and *few*. We showed that these quantifiers have different positions on the mental scale in subgroups of participants. Our findings are consistent with the claim that logical words can have various semantic representations for different speakers. We believe that our approach could be helpful for studying individual differences in the representation of not only quantifiers but also other functional or content words. In sum, in this paper, we showed that our computational model can bring together formal semantic and psycholinguistic approaches to study meaning representations.

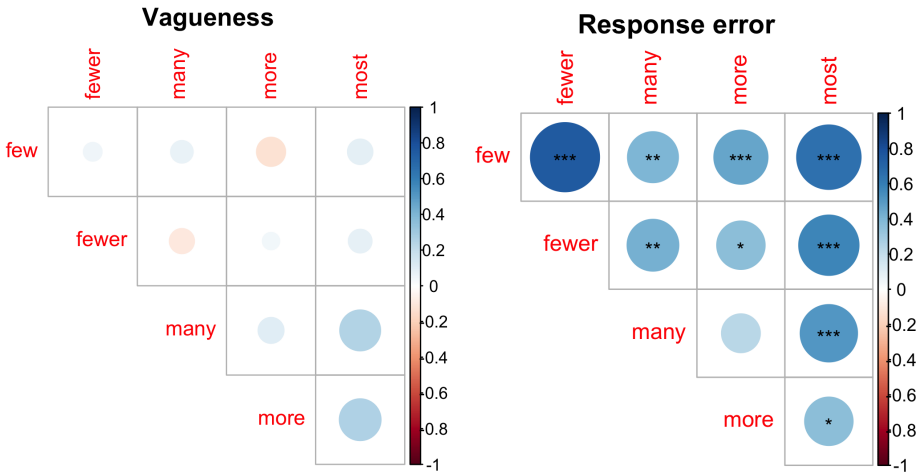
5 Acknowledgements

This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n.STG 716230 CoSaQ.

A Appendix

A.1 Correlations of vagueness and response error parameters across quantifiers

Figure **??** shows the correlations for vagueness, and Figure **??** for response errors. The correlations for vagueness were weak, suggesting that this parameter is quantifier-specific and not domain-general. In contrast, the correlations for response error varied, ranging from a strong correlation between *few* and *fewer than half* ($r = 0.75$), to the weakest correlation between *more than half* and *many* ($r = 0.24$, see Figure **??**). The strongest correlation was significantly higher than the weakest, Steiger's test $z = 4.72$, $p < 0.001$. This suggests that response error reflects general cognitive ability.



(a) (b)

Figure 7: **??** correlations between vagueness; **??** correlations between response error (significance level *** 0.001, ** 0.01,* 0.05). The p - values were adjusted using the Bonferroni correction.

A.2 Vagueness

A.2.1 Cluster analysis

We expected polar opposite quantifiers *few* and *many* to make comparable contributions to clustering on vagueness. What we observed instead was the asymmetry in *many* and *few*. The elbow plot and Silhouette method agreed that the two-cluster solution was optimal, identifying one cluster ($N = 24$) with high vagueness for *many*, and a second cluster ($N = 47$) with lower vagueness for *many* (Table **??**).

Table 3: Mean (*SD*) vagueness parameter in each cluster, 2-cluster solution.

Quantifier	Cluster 1 ($N = 24$)	Cluster 2 ($N = 47$)
Few	.016 (.001)	.016 (.001)
Fewer than half	.002 (.00004)	.002 (.00004)
Many	.023 (.002)	.017 (.001)
More than half	.001 (.00004)	.001 (.00002)
Most	.009 (.001)	.009 (.001)

A.2.2 Linear Discriminant Analysis

For the vagueness parameter, we expected vague quantifiers to contribute to the clustering. We found that only *many* contributed significantly to the clustering ($\lambda = 0.29$, $p < 0.001$). The LDA achieved 94% accuracy in classification of participants into clusters based on vagueness parameters for *many*, and the leave-one-out cross validation accuracy was 94%.

A.3 Response error

A.3.1 Cluster analysis

The elbow plot suggested that either two or three clusters should be optimal, but the Silhouette method indicated the 2-cluster solution. Assuming two clusters, we found a cluster of participants with few response errors ($N = 64$) and a cluster with more response errors ($N = 7$) across quantifiers, see Table **??**. This means that the majority of participants had a low response error rate. The difference in response error between clusters was most prominent for negative quantifiers.

Table 4: Mean (SD) response error parameter in each cluster, 2-cluster solution.

Quantifier	Cluster 1 ($N = 7$)	Cluster 2 ($N = 64$)
Few	.17 (.05)	.05 (.02)
Fewer than half	.19 (.03)	.06 (.03)
Many	.08 (.04)	.05 (.02)
More than half	.06 (.02)	.04 (.02)
Most	.09 (.03)	.04 (.02)

A.3.2 Linear Discriminant Analysis

We expected the response error parameter for negative quantifiers to contribute more to clustering. In line with this hypothesis, the Wilks test showed a significant contribution of response error parameters for *few* ($\lambda = 0.32$, $p < 0.001$) and *fewer than half* ($\lambda = 0.25$, $p < 0.001$), but not for *many*, *most* and *more than half*. Figure **??** shows the combined effect of the two quantifiers on clustering. Participants who made more errors while verifying *few* also made more errors for *fewer than half*. We used the LDA to predict the cluster membership for each participant based on response error parameters for *few* and *fewer than half*. The LDA achieved 99% accuracy, and the leave-one-out cross validation accuracy was 99%.

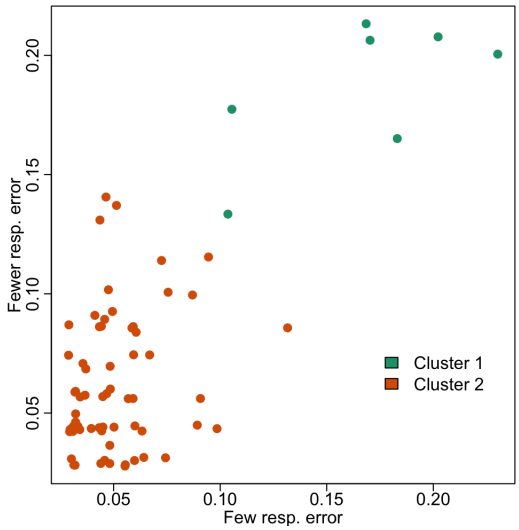
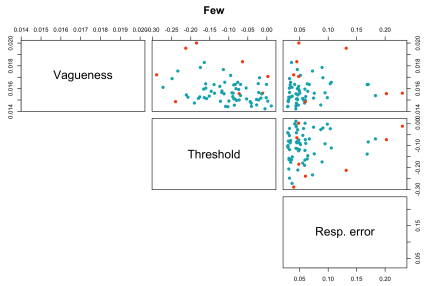


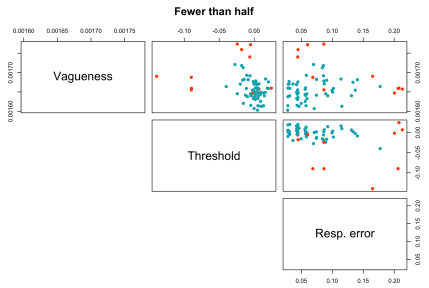
Figure 8: Two clusters for response error based on *few* and *fewer than half* parameters. The response error values of parameters for *fewer than half* are plotted against the response error values for *few*. Colors are used to indicate the cluster membership: Cluster 1 ($N = 7$) with high response error is in indicated in green, and Cluster 2 ($N = 64$) with low response error in orange.

A.4 Influential observations

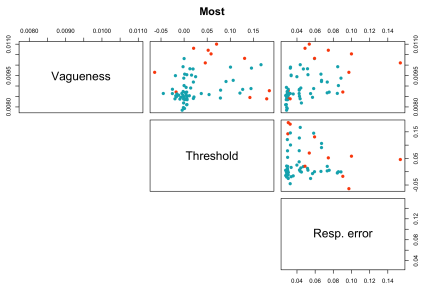
Figure 11 illustrates how relationships between model parameters for each quantifier are affected by influential observations. We computed the Cook's distance using the *ols plot cooksd bar* R function in the package *olsrr* (?, ?).



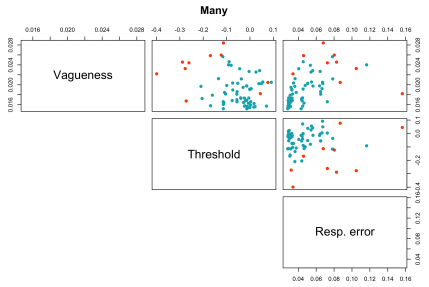
(a)



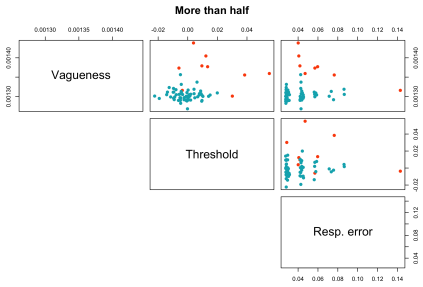
(b)



(d)



(c)



(e)

Figure 9: The scatter plots illustrate the relationships between model parameters (abbreviation Resp. error - response error) for each quantifier. The influential observations according to Cook's distance are indicated in red.