Most, but not *more than half*, is proportion-dependent and sensitive to individual differences[1]

Sonia RAMOTOWSKA — *Institute for Logic, Language and Computation, University of Amsterdam*
Shane STEINERT-THRELKELD — *Department of Linguistics, University of Washington*
Leendert VAN MAANEN— *Department of Psychology, University of Amsterdam*
Jakub SZYMANIK — *Institute for Logic, Language and Camputation, University of Amsterdam*

Abstract. In this study we test individual differences in the meaning representations of two natural language quantifiers – *most* and *more than half* – in a novel, purely linguistic task. We operationalized differences in meaning representations as differences in individual thresholds which were estimated using logistic regression. We show that the representation of *most* varies across subjects and its verification depends on proportion. Moreover, the choice of the representation of *most* affects the verification process. These effects are not present for *more than half*. The study demonstrates the cognitive differences between *most* and *more than half* and individual variation in meaning representations.

Keywords: generalized quantifiers, *most*, *more than half*, individual thresholds, meaning representations, verification strategies.

1. Introduction

Imagine that there are three candidates in an upcoming election: candidates A, B and C. To win the election the candidate needs *most* of the votes. How would you check if the sentence "*Most* of the people voted for candidate A" is true? You can represent this sentence in many ways. You can, for example, think that the number of people who voted for candidate A is greater than half of all of the votes or alternatively that the number of people who voted for candidate A is greater than the number of people who cast for their votes on other candidates. In the literature (e.g. Lidz, Pietroski, Halberda, & Hunter, 2011; Pietroski, Lidz, Hunter, & Halberda, 2009; Tomaszewicz, 2013) there are several proposals how the meaning of *most* can be represented:

(1) Representations of *most*
  a. *most*(votes in election, votes on A) ⇔ |votes on A| > ½|votes in election|
  b. *most*(votes in election, votes on A) ⇔ |votes on A| > |votes on not-A|
  c. *most*(votes in election, votes on A) ⇔ OneToOnePlus(votes on A, votes on not-A)
  d. *most*(votes in election, votes on A) ⇔ |votes on A| > |votes on B| + |votes on C|
  e. *most*(votes in election, votes on A) ⇔ |votes on A| > |votes in election| - |votes on A|

[1] We would like to thank Milica Denic and Fabian Schlotterbeck for their comments on early draft of this paper. The study was founded by the European Research Council under the European Unions Seventh Framework Programme (FP/20072013)/ERC Grant Agreement n. STG 716230 CoSaQ.

For example, if you represent the meaning of *most* as in (1c), you have to pair each vote cast for candidate A both with votes for either candidate B or C. If you find at least one vote for candidate A left unpaired, then candidate A will win.

According to the Interface Transparency Thesis (Lidz et al, 2011: 233): "The verification procedures employed in understanding a declarative sentence are biased towards algorithms that directly compute the relations and operations expressed by the semantic representation of that sentence". Although some studies (Pietroski et al., 2009; Lidz et al., 2011) support the Interface Transparency Thesis, there is also evidence that people might prefer different verification strategies (Steinert-Threlkeld, Munneke, & Szymanik, 2015; Talmina, Kochari, & Szymanik, 2017; see for overview: Szymanik, 2016). In this paper, we will present findings demonstrating that there are individual differences in the representation of natural language quantifiers (expressions like: *most, more than half, fewer than half, many, few, some, all, at least*). We will show evidence for individual differences in meaning representations with a special focus on two natural language quantifiers: *most* and *more than half.*

1.1. *Most* and *more than half*

Most and *more than half* are examples of truth-conditionally equivalent quantifiers, that differ in many other aspects, e.g., they seem to trigger different verification strategies (Hackl, 2009) and have different pragmatic associations (Solt, 2016). Generalized Quantifier Theory (GQT, Mostowski, 1957; Barwise & Cooper, 1981; Peters & Westerståhl, 2008; Szymanik, 2016) is not able to distinguish between expressions that are logically equivalent, but generate different linguistic intuitions.

According to Hackl's (2009) linguistic analysis, *more than half* is a comparative expression, while *most* is the superlative form of *many* (i.e. MANY+EST). However, under this analysis *most* also satisfies proportional truth-conditions. In contrast, the opposite quantifier to *most* – *fewest* – has only a superlative reading. The lack of proportional reading for *fewest* cannot be explained on the grounds of GQT but falls out naturally from Hackl's analysis.
According to Hackl (2009) the linguistic differences between *most* and *more than half* are reflected in different basic logical representations of these quantifiers.

(2) Logical representations of *most* and *more than half*
  a. $most(A, B) \Leftrightarrow |A \cap B| > |A - B|$
  b. $more\ than\ half(A, B) \Leftrightarrow |A \cap B| > \frac{1}{2}|A|$

Although both logical forms satisfy the same truth-conditions and thus are indistinguishable from the perspective of GQT, they may trigger different cognitive verification strategies. The verification of *more than half* requires the comparison of cardinality of the target set ($|A \cap B|$) to half of the size of A ($\frac{1}{2}|A|$), while the verification of *most* requires comparison between the cardinality of the target set ($|A \cap B|$) and the cardinality of the complement set ($|A - B|$)[2].

[2] Pietroski et al (2009) and Lidz et al (2011) provided evidence that the verification strategy for *most* should be $most(A, B) \Leftrightarrow |A \cap B| > |A| - |A \cap B|$.

Hackl (2009) supported his linguistic analysis with experimental data. Using a novel paradigm – Self Paced Counting – he argued that *most* is verified using a vote-counting strategy. In this experiment, he did not find a difference in overall reaction times and accuracy between *most* and *more than half.* Hackl (2009) argued that this lack of differences is evidence that participants treated *most* and *more than half* as equivalent expressions. To summarize, Hackl (2009) argued that *most* and *more than half* are verified using different strategies, but that these two quantifiers are truth-conditionally equivalent and therefore they are both true above 50% proportion and false below.

In contrast to Hackl's (2009) findings, other studies (Talmina et al., 2017; Kotek, Sudo, & Hackl, 2015) showed that participants might not treat *most* and *more than half* as equivalent quantifiers. Firstly, in a replication study, Talmina et al. (2017) found that *more than half* is verified slower than *most*. This finding questions Hackl's (2009) argument that participants treated *most* and *more than half* as equivalent quantifiers. Moreover, Talmina et al. (2017) suggested that subjects might have used various verification strategies for both quantifiers. Talmina et al.'s (2017) findings suggest a more complex picture, showing that people might differ in their representation of quantifiers.

Secondly, Kotek et al. (2015) support the hypothesis that *most* and *more than half* have different meaning representations. They found a difference between *most* and *more than half* in terms of their sensitivity to proportion. While *more than half* was judged equally likely as false for proportions below 50% and true for proportions above 50%, *most* exhibited an asymmetry. It was judged true for proportions above 50% less often than *more than half.* Kotek et al. (2015) concluded that the asymmetry between *most* and *more than half* for proportions above 50% might be explained by pragmatic associations of these quantifiers (Solt, 2016).

In particular, Solt (2016) explained the differences between *most* and *more than half* in terms of their scale structure requirements. *More than half* requires precise comparison, which is only possible on a ratio scale. *Most* has lower scale requirements and can be verified on a semi-ordered scale. On a semi-ordered scale, one of the two proportions to be compared is greater than another, when it is greater by some value. The semi-ordered scale allows only for imprecise, approximate comparisons. As a consequence, *most* has a preferred interpretation of "significantly greater than *more than half*".

The differences in required scale structure for *most* and *more than half* are reflected in their pragmatics (Solt, 2016). Solt (2016) found, in corpus data, that *most* is used with higher proportions or in the context, in which the precise comparison is not possible. *More than half*, in turn, expresses proportions slightly above 50% and occurs in the context, in which the precise data are available. Although Solt (2016) found clear differences in usage of *most* and *more than half*, she used corpus data that does not provide evidence for differences in processing and verification of these quantifiers. Therefore, based on her findings, it is not possible to know whether the differences between *most* and *more than half* should be attributed to semantics or rather to the pragmatics of these quantifiers.

Solt's (2016) claim that *most* has a strong "significantly *more than half*" interpretation was supported by other studies (Ariel, 2003; Pezzelle, Bernardi, & Piazza, 2018). For example, Ariel (2003) found a similar pragmatic tendency to use *most* with the higher proportions than *more than half* in a questionnaire study. Moreover, she argued that *most* and *more than half* are semantically different. *Most* is an upper-bounded quantifier, while *more than half* has no upper-bound. In addition, Pezzelle et al. (2018) investigated the meaning boundaries of several quantifiers, among others *most*. They asked the subjects to select, from a restricted choice, a quantifier that best describe a given scene. They found that *most* was used for proportions between 40% and 100% with a peak around 70%. Its usage highly overlapped with *many*, however *most* was chosen more often. Unfortunately, Pezzelle et al. (2018) have not studied *more than half* so the direct comparison between these two quantifiers on the selection task is not available.

To summarize, the strong preferences to use *most* with higher proportions stands in conflict with the treatment of *most* as a quantifier with a 50% threshold. It also raises a question if *most* has only one possible representation – truth-conditionally equivalent to *more than half*. The existing evidence suggests that there are differences between *most* and *more than half*, which might result in the differences in thresholds in these quantifiers. While *more than half* has a clear threshold, the threshold for *most* might vary between 50% and higher proportions. According to the truth conditions, *most* should have the same threshold as *more than half*. However, experimental evidence (Kotek et al, 2015) and corpus data (Solt, 2016) suggest that *most* can also have a higher threshold. The fact that *most* has two possible interpretations raises the question of whether this quantifier is represented in the same way by all language users. Only a few studies (e.g. Yildirim, Degen, Tanenhaus, & Jaeger, 2016; Talmina et al, 2017) investigated individual differences in quantifiers. Because quantifiers like *most* are sensitive to different interpretations, it might be also possible that people differ in how they represent quantifiers.

Thus, the question arises: are the differences between *most* and *more than half* outlined above reflected in individual differences in thresholds? Before presenting our methods for answering this question, in the next section we review studies showing that individual differences in natural language are widespread.

1.2. Individual differences in natural language

Individual differences in natural language are exhibited in many phenomena related to variation in performance of cognitive functions such as working memory and executive function (Kidd, Donnelly, & Christiansen, 2018), environmental variables (Kidd et al., 2018) or efficiency in updating predictions (Reuter, Emberson, Romberg, & Lew-Williams, 2018). They are present in many domains of language processing: morphosyntactic processing (Tanner & Van Hell, 2014), language production (Barlow, 2013), representation of words in context (Halff, Ortony, & Anderson, 1976), understanding of grammar constriction (passive voice) and universal quantification (Street & Dabrowska, 2010), among others. Individual differences are also characteristic for language disorders like dyslexia (Heim et al., 2008) or dysgraphia (Döhla, Willmes, & Heim, 2018).

In contrast, only a few studies have investigated individual differences in meaning representations. Talmina et al. (2017) found that some people use precise verification strategies for quantifiers, while others use estimation-based strategies. Furthermore, Yildirim et al. (2016) showed individual differences in listeners' expectations about the speaker's interpretation of quantifiers. Speakers can also adjust their representation of the quantifier meaning to the listener (Yildirim et al., 2016) or learn a new representation of a quantifier (Heim et al., 2015). Heim et al. (2015) showed that change in representation of one quantifier adjusts other quantifier representations: for example, a change in the representation of *many* affects the representation of *few*.

In addition, studies investigating the scalar implicature *some-not all* (e.g. Bott, Bailey, & Grodner, 2012; Spychalska, Kontinen, & Werning, 2016) show that people can be grouped with regards to their preferences in interpretation of natural language quantifiers into so-called pragmatic or logical responders. The logical responders tend to interpret *some* according to its semantic, literal meaning: *some* As are B iff the number of As than are B is greater than zero. This interpretation includes also the possibility that *all* As are B. The pragmatic responders, in turn, judge sentence *some* As are B as false if in fact *all* As are B. This division is also reflected in differences in ERPs N400 and late positivity between two groups of responders (Spychalska et al., 2016).

1.3. Current study

The current study tests the effect of individual differences in representations of the quantifiers *most* and *more than half*. We operationalized the individual differences in quantifier representation as individual thresholds. We asked participants to verify a sentence with quantifiers based on proportion, given as a percentage. We used quantifiers that intuitively varied in sharpness of their meaning boundaries: *more than half*, *fewer than half, most, many* and *few*. We used proportions given as percentage in order to force a proportional reading for all quantifiers. We formulated the following predictions.

According to GQT, *most* and *more than half* are truth-conditionally equivalent and therefore, should have the same threshold: 50%. Moreover, there should be no difference in the interpretation of these quantifiers between participants. In contrast to GQT, previous studies (Solt, 2016; Ariel, 2003; Kotek et al., 2015) showed *most* has also the "significantly greater than *more than half*" interpretation and it is dispreferred with proportions around 50%. These findings give a prediction that the threshold for *most* should be higher than the threshold for *more than half*. Finally, the number of studies (Yildirim et al., 2016; Talmina et al., 2017) showed that quantifiers, like other natural language expressions, are sensitive to individual differences in representation. We hypothesized that participants might vary in terms of which reading of *most* they prefer. Therefore, we predicted that:

(H1) Participants will have different representations for *most* and *more than half*.

Following Hackl (2009) we assumed that the choice of the verification strategy depends on the cognitive representation of the quantifier. Moreover, according to Solt (2016), *most* is verified using an imprecise, estimation-based strategy. Pietroski et al. (2009) and Lidz et al.

(2011) showed that the usage of the estimation-based strategy results in proportion-dependent performance. However, they did not contrast the proportion-dependent performance of *most* with *more than half* and so we cannot conclude that the effect they found was a consequence of linguistic properties of *most* rather than the task design. We directly compared the effect of proportion on speed of verification (reaction times) between *most* and *more than half.* Following Pietroski et al (2009) we hypothesized that when the proportion is close to 50% the verification of *most* should be more difficult. We predicted that:

(H2a) The verification of *most*, but not *more than half,* is proportion-dependent.

We also aimed to see if we can capture the effect of variation in representations between participants in their reaction times. We assumed that if participants have different representations of quantifiers, they also use different verification strategies. Therefore, we predicted that:

(H2b) Differences in representation will be reflected in differences in verification speed.

2. Methods

2.1. Participants

We collected data from 90 subjects. After exclusion criteria were applied, the final sample consisted of 47 male (age: *M* = 35, *SD* = 11, range: 22-59) and 24 female participants (*M* = 34.5, *SD* = 10, range: 22-59). 6 female and 18 male participants graduated high school, 6 female and 15 male subjects finished high school education and started college, 12 female and 14 male participants graduated college or obtained higher degree. Each participant received 4 US$ for participation. The study was a part of the project that received European Research Council and University of Amsterdam, Faculty of Humanities Ethics Committee ethical approvals.

2.2. Design

Participants were presented two sentences. The first sentence was of the form “Q of the As are B”, where Q was one of the quantifiers: *most*, *more than half, many, fewer than half, few* and As and Bs were pseudowords generated with the Wuggy software (Keuleers & Brysbaert, 2010) from English 6 letters adjectives and nouns. An English native speaker assessed pseudowords; we excluded them if they were too close to real English words or did not sound like plausible English words. 50 pseudo-adjectives and 50 pseudo-nouns were chosen and randomly paired. We checked frequency (Zipf value) of the original adjectives and nouns in SUBTLEX-US database (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The Zipf value of final lists were both 4.06. Each quantifier occurred with each pair of pseudowords only once and in a random order.

The second sentence presented to participants was of the form “*p*% of the As are B”, where As and Bs were the same pseudowords as in the first sentence and *p%* was a randomly generated proportion form 1% to 99%, excluding 50%. In the case of *most*, *more than half* and *fewer than half,* the proportions above and below 50% were counterbalanced within participants. Because *most* does not have a clear upper boundary (Ariel, 2003) we did not include the proportion 100%.

2.3 Procedure

Our experiment was conducted on Amazon Mechanical Turk. Participants had to decide if the first sentence is true based on information from the second sentence. They were presented 250 pairs of sentences, 50 per each quantifier. Firstly, they had to press the arrow down button and keep it pressed as long as they wanted to see the first sentence on the screen. Secondly, they had to press arrow down button again to read the second sentence with proportion. Finally, they had to choose the arrow left or arrow right buttons for true or false response. The response buttons were balanced between-subjects.

Before the proper experiment started, participants practiced the procedure for 8 trials in a training block. In the training block we used the quantifiers *some*, *all*, and *none* in the first sentence. At the end of the experiment participants were asked to provide basic demographic information (e.g., gender, age, education background).

2.4. Preprocessing reaction times (RT) data

Before we estimated individual thresholds we excluded reaction times shorter than 300 ms and longer than mean+2SD for each quantifier and true/false responses separately.

2.5. Logistic regression model

In order to estimate participants' individual thresholds we applied logistic regression using R *nls* self-starting function (Bates & Chambers, 1992):

$$(3)\ \mathrm{P(T)} \sim \frac{1}{1+e^{(p_0-p)/s}},$$

with starting values: $p_0 = 50$, $s = 4$ for *most*, *more than half* and *many*, and $p_0 = 41$, $s = -5$ for *few* and *fewer than half.*

P(T) indicates the probability that a participant provided a "true" response, and *p* the percentage introduced on every trial. The estimated parameters were p_0 – participant individual threshold – and *s* – the steepness of logistic regression curve.

The individual threshold could not be estimated using the *nls* function if a participant's "true" and "false" responses did not overlap. In those cases, we computed thresholds as the average of the highest proportion for which a participant responded "false" and lowest proportion for which he or she responded "true" (vice versa for *few* and *fewer than half*).

3. Results

3.1. Excluded participants

We excluded 11 participants, who had 50% or more responses below 300 ms. Additionally, we ran the *glmer* function in the R package *lemrTest* (Kuznetsova, Brockhoff, & Christensen, 2017) separately for each quantifier, with random slope for each participant. The random slopes indicate whether the probability of response "true" increases or decreases with increasing proportion. We assumed that the random slope for quantifier *most*, *more than half* and *many* should be positive and for *fewer than half* and *few* negative. We excluded 6 participants, who did not meet this criterion. Finally, we excluded two participants from further analysis because their estimated threshold was higher than 100% or lower than 0%.

3.2. Individual thresholds

We estimated individual thresholds for each quantifier. Figure 1 presents individual thresholds distributions among quantifiers and summarizes descriptive statistics of thresholds. The mean accuracy for all quantifiers above and below threshold was high: *many* 95%, *most* 96%, *more than half* 97% both above and below threshold, and *few* and *fewer than half* 94% above threshold and 90% below threshold. The mean reaction times above thresholds were: *many* 991.88 ms (sd = 384.51), *most* 1025.06 ms (sd = 502.67), *more than half* 925.28 ms (sd = 342.48), *few* 1081.76 (sd = 421.89) and *fewer than half* 1068.96 (sd = 374.52). The mean reaction times below thresholds were: *many* 1097.24 (sd = 421.21), *most* 1035.28 (sd = 434.23), *more than half* 942.25 (sd = 306.84), *few* 1181.70 (sd = 425.60), *fewer than half* 1172.03 (sd = 475.01).

We tested if there are differences in mean individual thresholds between quantifiers. We found a significant main effect of threshold ($F_{4,345} = 9.21$, $p < 0.001$). After applying Bonferroni correction on the significance level, we found that the mean threshold for *few* was lower than the threshold for the other quantifiers; the mean threshold for *many* was lower than for *most*, *more than half* and almost significantly lower ($p = 0.056$) than for *fewer than half*; and the threshold for *fewer than half* was lower than the threshold for *most*. Importantly, the mean threshold for *most* was higher than the threshold for *more than half.*

A Kolmogorov-Smirnov test revealed that the distribution of thresholds is different for *most* and *more than half* ($D = 0.30$; $p < 0.01$).

Taken together the results show that *most* has a higher threshold than *more than half*, but also that participants differ in their representation of *most*. While in the case of *more than half* almost all participants had a threshold of 50%, in the case of *most* some participants had a threshold between 50-70%.

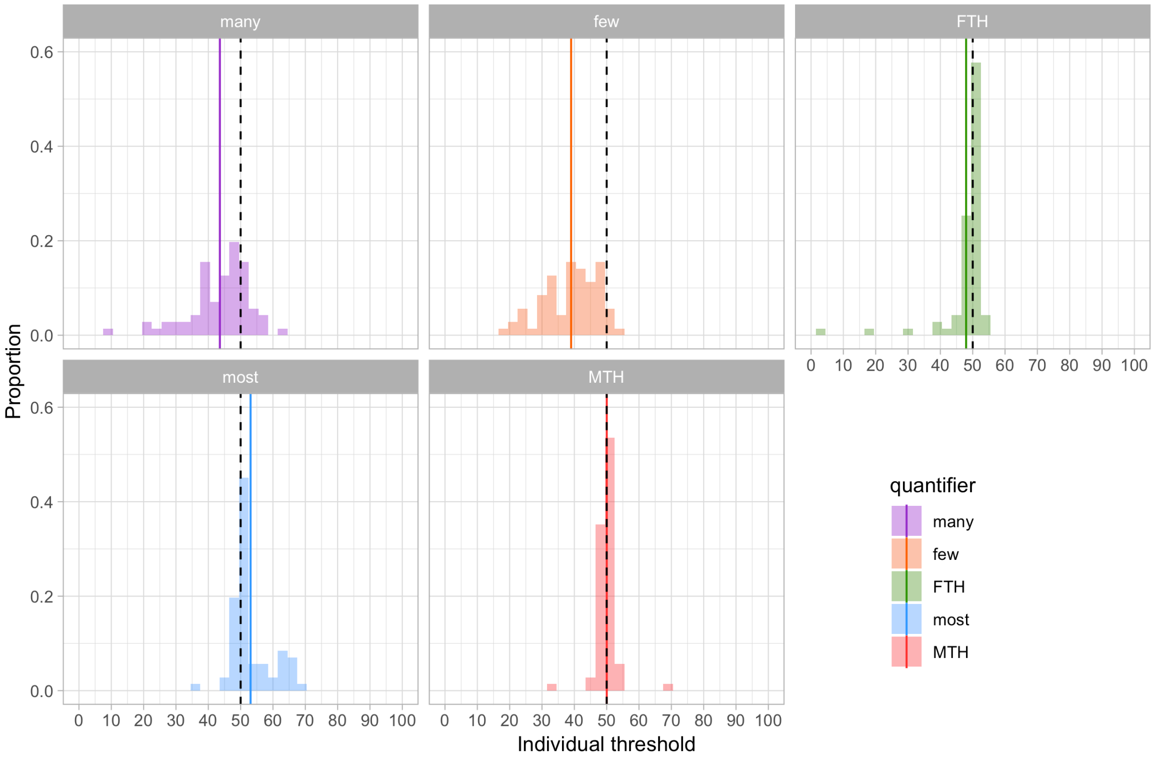

Figure 1: Histograms present individual thresholds distribution in each quantifier. The dashed lines indicate 50%, the solid lines indicate mean individual threshold. The mean threshold for *many* is 44% (sd = 10), *few* 39% (sd = 8), *fewer than half* (FTH) 48% (sd = 7), *most* 53% (sd = 6), *more than half* (MTH) 50% (sd = 4).

3.3. Proportion effect on reaction times

To understand the effect of proportion on reaction times, we re-coded all responses relative to individuals' thresholds. We coded "true" responses that are above the individual threshold and "false" responses that are below the threshold as correct responses. We ran a mixed effect regression model (R package *lmerTest*; Kuznetsova et al., 2017) with reaction times as dependent variable and quantifiers (*most*, *more than half*), proportion (*z*-scored) and responses (true/false) and their interactions as predictors. Firstly, we tested the random effects structure. Following Barr, Levy, Scheepers, & Tily (2013), we tried to keep the random structure maximal until the model converged. We used the best path forward algorithm and included random slopes that significantly improved the model (tested by anova function in R; see Appendix A). If two random slopes were significant, we included the one that had lower *p*-value. To this model we included by-subject random intercept and by-subject random slope for proportion. We used *more than half* as baseline.

Secondly, we tested the significance of the fixed effects. Table 1 and Figure 2 summarize the effects. Here we focus on the most important one. We found no significant effect of proportion (β = -22.16; t = -0.96; p = 0.34), but a significant quantifier-proportion interaction (β = -133.18; t = -3.99; p < 0.001), meaning that the proportion had greater effect on RTs in case of *most* than *more than half.* Additionally, we found a significant main effect of quantifier (β = 216.82; t = 6.43; p < 0.001).

This finding shows that, in contrast to *more than half*, the verification of *most* is dependent on proportion, meaning that the verification is slower when the proportion is close to 50%.

Table 1: The summary of regression models comparing the effect of proportion between *most* and *more than half*.

Effect	estimates	*t* value	*p* vale
Intercept	947.15	26.28	< 0.001
Prop	-22.16	-0.96	0.34
Quant	216.82	6.43	< 0.001
Resp	48.86	1.52	0.13
Prop:quant	-133.18	-3.99	< 0.001
Prop:resp	82.96	2.59	< 0.01
Quant:resp	-42.41	-0.94	0.35
Prop:quant:resp	237.40	5.29	< 0.001

Notes: Prop. – main effect of proportion; Quant. – main effect of threshold; Resp. – main effect of response; Prop:quant – proportion threshold interaction; Prop:Resp – proportion response interaction; Quant:resp – threshold response interaction; Prop:quant:resp – three way interaction.

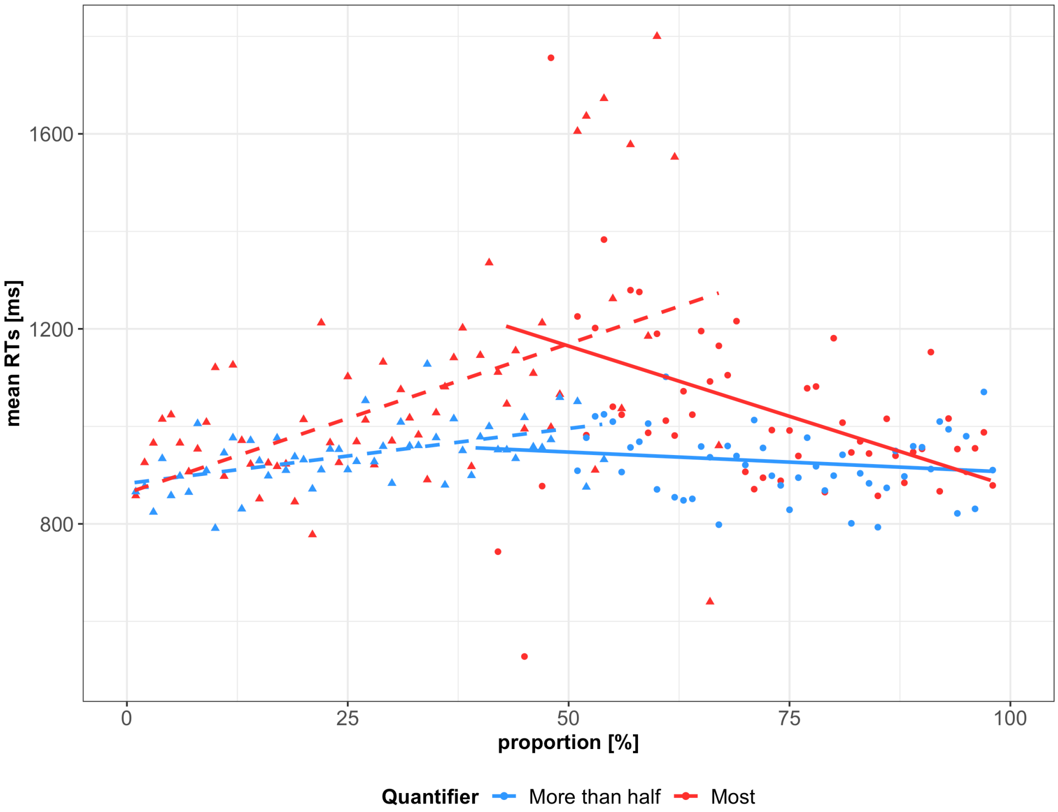

Figure 2: The figure presents mean reaction times for each proportion and each quantifier. The triangles correspond to responses below threshold and circles to responses above threshold. The dashed lines illustrate the model predictions for responses below threshold and solid lines for responses above threshold. The red lines for *most* are steeper indicating the proportion effect for this quantifier. For clarity of the figure we constrained the y-axis to 500-1800 ms.

3.3. Individual threshold as predictor of reaction times

In the next step we tested if the individual thresholds predict the speed of the verification process. In the regression model we used reaction times only for correct responses. We *z*-scored the proportion and threshold variables. We tested each quantifier separately using linear mixed effect regression model in R package *lmerTest* (Kuznetsova et al., 2017). We used reaction times as dependent variable and proportion, individual threshold, response type (true/false) and their interactions as predictors. We set true responses as the baseline level. We used the same procedure to include random slopes as in model comparing proportion effect for *most* and *more than half.*

To test the fixed-effect structure, we used the same procedure as above: we started with the maximal model and excluded those effects that did not improve model by using anova function in R (see Appendix A). We summarize all regression models' coefficients in Table 2 and included detailed description of results for *fewer than half, many and few* in Appendix B.

Table 2: The summary of regression models estimates with significance level (. < 0.1; * < .5; ** < .01; *** < .001) for all quantifiers.

Effect	*More than half*	*Fewer than half*	*Most*	*Many*	*Few*
Intercept	941.55***	1136.85***	1238.77***	1110.22***	1541.56***
Prop.	-18.26	-39.45.	-216.17***	-158.04***	316.89***
Thr.	-13.99	15.51	216.98***	-8.13	-237.56***
Resp.	53.20*	-33.30	-67.63	208.57***	-368.56***
Prop:Thr			-111.36**	-9.46	-114.40**
Prop:Resp	78.46**		381.99***	352.29***	-462.63***
Thr:Resp	24.47*		-170.40***	-217.42***	197.08***
Prop:Thr:Resp			132.66**	-105.31**	152.98***

Notes: Prop. – main effect of proportion; Thr. – main effect of threshold; Resp. – main effect of response; Prop:Thr – proportion threshold interaction; Prop:Resp – proportion response interaction; Thr:Resp – threshold response interaction; Prop:Thr:Resp – three way interaction.

<u>*More than half*</u> We predicted that the individual thresholds should not influence the reaction times during verification of *more than half.* We included by-subject random intercept. We used model comparison to determine the best model. The best model did not include three-way interaction between proportion, response and threshold ($\chi^2(1) = 0.39$; p = 0.53) and interaction between proportion and threshold ($\chi^2(1) = 2.23$; p = 0.14). As predicted, we did not find a main effect of threshold for *more than half* (β = -13.99; t = -0.60; p = 0.55).

<u>*Most*</u> We hypothesize that the individual threshold should predict the reaction times during verification of *most*. We found that the best random structure of *most* includes by-subject random intercept and by-subject random slope for proportion. We found a main effect of threshold (β = 216.98; t = 4.20; p < 0.001) and a significant interaction between threshold and response type (β = -170.40; t = -3.61; p < 0.001), indicating that the threshold effect was smaller for false responses. Finally, we also found a significant interaction between threshold and proportion (β = -111.36; t = -3.03; p < 0.01) and a three-way interaction between

proportion, response and threshold (β = 132.66; t = 3.17; p < 0.01), meaning that the threshold affected the proportion effect, but only for true responses.

All together, these findings show that the individual thresholds affect the speed of the verification process in vague quantifiers e.g. *most*, but not in quantifiers that have a clear threshold like *more than half.* It is worth to mention that the effect of threshold for *most* was asymmetric and present only for responses above threshold.

4. General discussion

The main goal of this paper was to investigate variability of meaning representations between subjects and assessing whether *most* and *more than half* are truly equivalent. We tested differences in meaning representations by estimating individual thresholds for quantifiers. We also tested for differences in the verification process of *most* and *more than half* by looking into participants' reaction times.

According to Hackl's (2009) linguistic analysis, *most* and *more than half* are verified using different strategies. Following Hackl's (2009) findings, Solt (2016) postulated that *most* is verified using approximate strategy. As a consequence, *most* should have a "significantly greater than *more than half*" interpretation. Solt (2016) found supporting evidence for her theory in corpus data.

Following Solt's (2016) findings we considered *most* as a vague quantifier, which can have a literal interpretation, equivalent to *more than half*, and a "significantly greater than *more than half"* interpretation. To test the first hypothesis, we estimated individual thresholds using logistic regression for *most* and *more than half* and three other quantifiers: *few, fewer than half* and *many*. We found that the threshold for *more than half* is 50%, while in the case of *most,* there is higher variation in thresholds. Moreover, the mean threshold for *most* was higher than mean threshold for *more than half.* This finding clearly suggest that *most* is more sensitive to individual interpretation.

In contrast to our finding, Pietroski, Lidz, Hunter, Odic, & Halberda (2011) conducted additional analyses on Pietroski et al.'s (2009) data to support their claim that subjects had a 50%-threshold for *most* in their experiment. They investigated the deviation of accuracy from the Approximate Number System model predictions and concluded that the deviation did not increase when the ratio approached 1. The disparity between our and Pietroski et al. (2011) findings might be explained in different ways. Firstly, Pietroski et al.'s (2011) analysis is indirect and specifies only the deviation from the model predictions. In our analysis we estimated individual thresholds directly form participants responses. Therefore, our analysis does not require any additional assumptions about the correctness of the model. Secondly, Pietroski et al. (2009, 2011) ran a visual stimuli task design in a way that forced ANS performance. We, instead, gave our participants a purely linguistic task with unlimited time to provide responses. Therefore, our task is able to detect subtle differences in natural language quantifiers' representation, while Pietroski et al. (2009, 2011)'s task confounds the linguistic effects with the influence of visual and number cognition.

The disparity between our and Pietroski et al.'s (2011) results shows the advantage of using a novel purely linguistic task. Verification processes of quantifiers are often studied using visual stimuli (e.g. Pietroski et al., 2009; Bott, Augurzky, Sternefeld, & Ulrich, 2017; Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; Zajenkowski & Szymanik, 2013; Szymanik, 2016). For example, Pietroski et al. (2009) and Lidz et al. (2011) used a number cognition model – the Approximate Number System model (Dehaene, 1997) – to test the verification of *most*. We decided to use a purely linguistic paradigm, because the verification process of quantifiers against visual scene can be affected by many non-linguistic factors. For example, if the verification of quantifiers is based on ANS, then factors like type of task (Gilmore, Attridge, & Inglis, 2011; see for review: Dietrich, Huber, & Nuerk, 2015), duration of display (Cheyette & Piantadosi, 2019; Inglis & Gilmore, 2013), and set size (Dietrich, Nuerk, Klein, Moeller, & Huber, 2019) will affect the verification process regardless of quantifier representation. Therefore, we think that the verification of quantifiers should be studied also in purely linguistic tasks to test to what extent the effects found in picture tasks can be attributed exclusively to semantic processing.

It is worth stressing that although we found differences in the interpretation of *most* and *more than half*, they are not completely in line with the Solt (2016) and Ariel (2003) findings. Solt (2016) and Ariel (2003) found that *most* is preferred for proportions above ~65%-70%. We found that some participants had thresholds above 60% for *most*, but the majority of participants had a threshold lower than 60%. This might mean that Solt (2016) and Ariel (2003) captured some additional pragmatic effects on *most*, that pushed the threshold of this quantifier higher. In contrast, our task was very abstract (e.g., we used pseudowords) which mitigates the influence of a pragmatic interpretation on *most*. Moreover, Ariel (2003) tested Hebrew *rov* for *most*, while we tested English *most*. We cannot exclude the possibility that the differences in findings might be explained by differences in languages.

Secondly, we found that the verification of *most* is proportion-dependent in terms of reaction times. The verification of *most* takes longer when the given proportion is close to 50%. No such effect was found for *more than half*. These findings extend the previous studies. Pietroski et al. (2009) showed that the verification of *most* is dependent on proportion in terms of accuracy by using an ANS model. However, they (Pietroski et al., 2009) did not contrast *most* with *more than half* to show that these quantifiers differ in verification process.

There are at least two possible explanations of the proportion effect for *most*. Firstly, it might be a consequence of a difference in verification strategy. *More than half* is verified using a precise strategy, comparing the given proportion to 50%. In the case of *most,* participants had to compute the proportion of As that are not B given the proportion of As that are B. They computed the number of As that are not B approximately, which results in greater proportion-dependent performance. Although we used a purely linguistic task, it is possible that participants engage the Approximate Number System into the verification process. Previous studies (e.g. Moyer & Landauer, 1967; Hinrichs, Yurko & Hu, Psychology, & 1981) show that ANS effects, e.g. distance effect, can be found even in a symbolic number comparison task.

According to the second possible explanation, the proportion effect of *most* is a result of the pragmatic strengthening. On the one hand, participants represented *most* as *more than half*; on the other hand, they had a strong pragmatic preference towards using *most* for higher proportions. Before they made a decision, they had to choose between these representations. Future studies need to shed light on disentangling these two competing explanations.

In addition to the proportion effect, we tested if the differences in thresholds in vague quantifiers (*most, many, few*) will affect the verification process. We found an effect of threshold on reaction times in vague quantifiers, but not in quantifiers with sharp meaning boundaries (*more than half* and *fewer than half*). The lack of threshold effect for *many* was one deviation from this result.

The results presented in this paper clearly suggest that *more than half* and *fewer than half* have unequivocal thresholds. In contrast, *most, many* and *few* have varied thresholds. The literature about *many* and *few* (e.g. Partee, 1988) consistently claims that these two quantifiers are highly context dependent and that they can have various interpretations. Our findings suggest that *most* exhibits similar effects. Further experimental studies are needed to explain how these meanings change and are selected in the context. It is possible that the specific context will trigger pragmatic reasoning about *most* and push the thresholds even higher.

Our study fits with an increasing number of findings on individual differences in natural language. Although many studies (e.g. Newstead & Coventry, 2000) investigated how the interpretation of vague quantifiers depends on contextual features like set size (Newstead, Pollard, & Riezebos, 1987; Newstead & Coventry, 2000), size of the stimuli and its position with relation object that creates context (Newstead & Coventry, 2000) or the number of non-target objects (Coventry, Cangelosi, Newstead, & Bugmann, 2010), little attention has been paid to individual differences in meaning representations. We aimed to bridge this gap by finding individual differences in natural language on example of quantifiers.

Our study also has several limitations. Firstly, the task was very abstract. On the one hand, this can be considered as an advantage, because abstract tasks limit pragmatic reasoning and allow us to test semantic differences between quantifiers. On the other hand, it makes quantifiers like *many* and *few* hard to interpret. Secondly, we tested a wide range of proportions, which means that we had only a limited number of trials per proportion and participant. We tried to compensate for this problem by including a large number of participants into our study (Rouder & Haaf, 2018) and excluding the outliers' responses. Thirdly, in our study all five quantifiers were a within-subject variable. It is therefore possible that estimated thresholds are affected by interaction between quantifiers. For example, some participants might have used the same 50% threshold for *most* and *more than half* to simplify the task (assuming that it is easier to perform the task, when participants have to remember only one threshold instead two). It would be worth testing if the same, or even stronger results, can be observed in a between-subject design. Finally, the logistic regression method, which we used to estimate the thresholds, was not always successful. In future work, we hope to overcome this difficulty by applying more complex methods to estimate the underlying properties of the verification process, such as evidence accumulation modeling (Anders et al, 2015; Ratcliff & McKoon, 2018).

This study contributed to the discussion about differences between *most* and *more than half* by showing that *most* exhibits more sensitivity to individual differences and is proportion-dependent. In this way, we showed that truth-conditionally equivalent expressions differ in meaning and that *most* is a vague quantifier with various meaning representations. We showed differences between *most* and *more than half* in a novel, purely linguistic task. By using this task, we avoided confounds between semantic meaning of the expression and other cognitive systems and we were able to directly compare *most* with *more than half.* Finally, we presented a new method to investigate individual differences in meaning representations.

References

Anders, R., Riès, S., van Maanen, L., Alario, F.-X. (2015). Evidence accumulation as a model for lexical selection. *Cognitive Psychology, 82*, 57-73.

Ariel, M. (2003). Does most mean 'more than half'? *Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Phonetic Sources of Phonological Patterns: Synchronic and Diachronic Explanations*, 17–30.

Barlow, M. (2013). Individual differences and usage-based grammar. *International Journal of Corpus Linguistics*, *18*(4), 443–478.

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.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219.

Bates, D. M., & Chambers, J. M. (1992). "Nonlinear Models". In S. J. M. Chambers & T. J. Hastie (Eds.), *Statistical Models*. Pacific Grove: Wadsworth & Brooks/Cole.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123–142.

Bott, O., Augurzky, P., Sternefeld, W., & Ulrich, R. (2017). Incremental generation of answers during the comprehension of questions with quantifiers. *Cognition*, *166*, 328–343.

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences*, *116*(36), 17729–17734.

Coventry, K. R., Cangelosi, A., Newstead, S. E., & Bugmann, D. (2010). Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, *2*(2), 221–241.

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

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

Dietrich, J. F., Huber, S., & Nuerk, H. C. (2015). Methodological aspects to be considered when measuring the approximate number system (ANS) - a research review. *Frontiers in Psychology*, *6*, 1–14.

Dietrich, J. F., Nuerk, H.-C., Klein, E., Moeller, K., & Huber, S. (2019). Set size influences the relationship between ANS acuity and math performance: a result of different strategies? *Psychological Research*, *83*(3), 590–612.

Döhla, D., Willmes, K., & Heim, S. (2018). Cognitive profiles of developmental dysgraphia. *Frontiers in Psychology*, *9*.

Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system.

Quarterly Journal of Experimental Psychology, *64*(11), 2099–2109.

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

Halff, H. M., Ortony, A., & Anderson, R. C. (1976). A context-sensitive representation of word meanings. *Memory & Cognition*, *4*(4), 378–383.

Heim, S., McMillan, C. T., Clark, R., Golob, S., Min, N. E., Olm, C., … Grossman, M. (2015). If so many are "few", how few are "many"? *Frontiers in Psychology*, *6*.

Heim, S., Tschierse, J., Amunts, K., Wilms, M., Vossel, S., Willmes, K., … Huber, W. (2008). Cognitive subtypes of dyslexia. *Acta Neurobiologiae Experimentalis*, *68*(1), 73–82.

Hinrichs, J. V., Yurko, D. S., & Hu, J. -m. (1981). Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 890–901.

Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: How are approximate number system representations formed? *Cognition*, *129*(1), 63–69.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13).

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

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae*, *44*(1), 12–36.

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

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, *18*(3), 178–182.

Newstead, Stephen E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, *12*(2), 243–259.

Partee, B. (1988). Many Quantifiers. *Proceedings of ESCOL 5*, 383–402.

Peters, S., & Westerståhl, D. (2008). Quantifiers in Language and Logic. In *Quantifiers in Language and Logic*. Oxford: Oxford University Press.

Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers, *Cognition, 181*, 117–126.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of "Most": Semantics, numerosity and psychology. *Mind and Language*, *24*(5), 554–585.

Pietroski, P., Lidz, J., Hunter, T., Odic, D., & Halberda, J. (2011). Seeing what you mean, mostly. *Syntax and Semantics*, *37*, 181–217.

Ratcliff, R., & McKoon, G. (2018). Modeling Numerosity Representation With an Integrated Diffusion Model. *Psychological Review*, *125*, 183–217.

Reuter, T., Emberson, L., Romberg, A., & Lew-Williams, C. (2018). Individual differences in nonverbal prediction and vocabulary size in infancy. *Cognition*, *176*, 215–219.

Rouder, J. N., & Haaf, J. M. (2018). Power, Dominance, and Constraint: A Note on the Appeal of Different Design Traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26.

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

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.

Steinert-Threlkeld, S., Munneke, G.-J., & Szymanik, J. (2015). Alternative Representations in Formal Semantics: A case study of quantifiers. *Proceedings of the 20th Amsterdam Colloquium*, 368–378.

Street, J. A., & Dabrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua*, *120*(8), 2080–2094.

Szymanik, J. (2016). *Quantifiers and Cognition: Logical and Computational Perspectives*. Cham: Springer.

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

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*(1), 289–301.

Tomaszewicz, B. (2013). Linguistic and Visual Cognition: Verifying Proportional and Superlative Most in Bulgarian and Polish. *Journal of Logic, Language and Information*, *22*(3), 335–356

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*, *67*(6), 1176–1190.

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.

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.

Appendix A

A.1. Proportion effect on reaction times – random structure

Tests for by-subject random effects: model with only intercept vs. model with random slope for proportion (χ^2 (2) = 16.12; p = 0.0003), model with only intercept vs. model with random slope for response (χ^2(2) = 13.04; p = 0.001), model with random slope for quantifier had singular fit, model with random slope for proportion vs. model with random slope for proportion and response (χ^2 (3) = 3.29; p = 0.35).

A.2. Individual threshold as predictor of reaction times – random structure

More than half: by-subject random slopes for proportion and response gave singular fit; Most: model with only by-subject intercept vs. model with by-subject random slope for proportion ($\chi^2(2) = 12.71$; p = 0.002), model with only by-subject intercept vs. model with by-subject random slope for response ($\chi^2(2) = 11.40$; p = 0.003); model with both random slopes gave singular fit; Many: model with only by-subject intercept vs. model with by-subject random slope for proportion ($\chi^2(2) = 6.43$; p = 0.04), model with only by-subject intercept vs. model with by-subject random slope for response ($\chi^2(2) = 3.25$; p = 0.2); Few: model with only by-subject intercept vs. model with by-subject random slope for proportion ($\chi^2(2) = 7.81$; p = 0.02), model with only by-subject intercept vs. model with by-subject random slope for response ($\chi^2(2) = 8.86$; p = 0.01); model with two by-subject slopes did not improve fit ($\chi^2(3) = 4.74$; p = 0.19); Fewer than half: model with only by-subject intercept vs. model with by-subject random slope for proportion ($\chi^2(2) = 14.80$; p = 0.0006), model with only by-subject intercept vs. model with by-subject random slope for response ($\chi^2(2) = 18.79$; p < 0.0001); model with two by-subject slopes improved fit ($\chi^2(3) = 7.89$; p = 0.04).

Appendix B

Fewer than half We predicted that the individual thresholds should not influence the reaction times during verification of *fewer than half* as in case of *more than half.* We included by subject random intercept and by-subject random slope for percent and response type. By using model comparison, we excluded three-way interaction ($\chi^2(1) = 0.22$; p = 0.64), threshold-response interaction ($\chi^2(1) = 0.15$; p = 0.7), threshold-proportions interaction ($\chi^2(1) = 1.62$; p = 0.20) and proportion-response interaction ($\chi^2(1) = 1.98$; p = 0.16). The final model for *fewer than half* included only three main effects. The effect of threshold was not significant ($\beta = 15.51$; $t = 0.62$; $p = 0.54$).

Many and few Finally we also predicted that the verification time of *many* and *few* should be threshold-dependent. We included by-subject random intercept for both quantifiers and by-subject random slope for proportion for *many* and by-subject random slope for response type for *few*. For *many* we did not find a significant main effect of threshold ($\beta = -8.13$; $t = -0.27$; $p = 0.78$) but did find a significant threshold-response type interaction ($\beta = -217.42$; $t = -4.53$; $p < 0.001$), meaning that the effect of threshold was greater for false responses. We also did not find a significant threshold-proportion interaction ($\beta = -9.46$; $t = -0.68$; $p = 0.50$), but did find a significant three-way interaction between proportion, response and threshold ($\beta = -105.31$; $t = -2.95$; $p < 0.01$), meaning that for responses false there was a threshold-proportion interaction.

For *few,* we found a main effect of threshold ($\beta = -237.12$; $t = -3.86$; $p < 0.001$), a significant interaction between threshold and response type ($\beta = 197.08$; $t = 3.55$; $p < 0.001$), a significant interaction between threshold and proportion ($\beta = -114.40$; $t = -2.84$; $p < 0.01$) and a significant three-way interaction between proportion, response and threshold ($\beta = 152.98$; $t = 3.59$; $p < 0.001$), meaning that the effect of threshold was stronger for true responses than for false responses and that it influenced the proportion effect stronger for true responses, than false responses.