Quantifiers and verification strategies: connecting the dots

Natalia Talmina[1], Arnold Kochari[2], and Jakub Szymanik[2*]

[1] Johns Hopkins University, Baltimore, Maryland, U.S.A.
talmina@jhu.edu
[2] ILLC, Universiteit van Amsterdam, Amsterdam, the Netherlands
a.kochari@uva.nl
jakub.szymanik@gmail.com

Abstract

In this paper, we replicate the influential study of Hackl (2009), making more specific algorithmic-level predictions based on Hackl's findings. Hackl argued that two semantically equivalent quantifiers *more than half* and *most* are associated with different verification strategies. The results of our experiment diverge in several respects from the original study. We explain the results by focusing on two potential confounds in Hackl's 2009 experimental set-up: different roles that working memory can play in the verification of different quantifiers and individual differences suggesting the use of various cognitive strategies.

1 Introduction

In his influential paper, Hackl (2009) explores whether there is a cognitively significant difference in the specifications of truth conditions of quantifiers *most* and *more than half*, captured below.

(1) a. $[\![\text{most}]\!] = |A \cup B| > |A - B|$
b. $[\![\text{most}]\!] = |A \cup B| > \frac{1}{2}|A|$

(2) a. $[\![\text{more than half}]\!] = |A \cup B| > |A - B|$
b. $[\![\text{more than half}]\!] = |A \cup B| > \frac{1}{2}|A|$

Hackl argues on conceptual and linguistic grounds that (1a) is the preferred option over (1b) for *most*, while (2b) is a better way to express *more than half* compared to (2a). Subsequently, he also suggested that although the two denotations are truth-conditionally equivalent, the way in which they are specified appears to point to distinct verification procedures. *More than half* explicitly calls for dividing the total number of A's in half in the course of verification, while verifying *most* requires comparing the total number of A's that are B's (e.g. the number of dots that are blue) with the number of A's that are not B's (e.g. the number of dots that are not blue).

In order to understand whether there is a difference in verification profiles that are triggered by *most* and *more than half*, Hackl conducted an experiment where participants had to verify visual scenes (pictures containing rows of dots of different colors) against sentences like *Most of the dots are blue* or *More than half of the dots are blue*. He applied the Self-Paced Counting paradigm, which is similar in spirit to the widely used self-paced reading paradigm: instead of having access to the whole scene at once, participants have to press a button to proceed through the scene step-by-step while the time they spend on each screen is measured. Based

*The author have received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

on the time spent on each screen, one can make inferences about the processes that took place at that point.

The results showed there was no significant difference in overall reaction (i.e. screen inspection) times or accuracy, which Hackl takes to indicate that subjects treated *most* and *more than half* as equivalent expressions. However, there was a significant difference in reaction times *per screen*, when excluding the final screen where the decision was made: verifying *more than half* took subjects consistently longer than verifying *most* (during the inspection of screens which did not yet reveal the correct answer). Hackl observes that this difference makes sense if *most* favors a kind of lead counting strategy — "keeping track of whether the target color leads overall and by how much" (p. 89). Contrary to verifying *more than half*, the lead counting algorithm does not call for dividing the total number of dots in half. The design of the experiment made the task easier for the strategy assumed to be adapted for *most*: in each screen, it was easy to evaluate whether there were more dots in the target color than in the other color, whereas division by half was not as easy in this set-up.

Additional evidence for a direct relationship between meaning and verification comes from other experimental paradigms: Pietroski et al. (2009) and Lidz et al. (2011) conducted experiments in which participants had to verify visual scenes that sometimes had advantages (in speed or accuracy) for possible strategies associated with the meaning of *most*: a strategy Pietroski et al. called *OneToOnePlus*, which requires pairing objects A that have property P with objects A that don't have property P, and a *selection strategy* that requires estimating the cardinalities of sets of all objects that are being compared. They have found that participants did not make use of these strategies when the setup made them advantageous, but rather used the subtraction strategy throughout the experiment. They interpret this result as supporting the Interface Transparency Thesis, which states that "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." (Lidz et al., 2011, p. 233).

However, a lot of the questions remain unanswered. While the results of the studies can be interpreted to be indicative of a direct relationship between meaning and verification, the choice of a particular verification strategy could be guided by multiple other factors, such as cognitive load, the presentation of the stimuli (the types of objects in a visual scene, their position on the screen, time of presentation, etc.), a bias towards using the same initially assumed strategy throughout the whole experiment, etc. (see Kotek et al. 2015 for similar arguments).

One such potentially influential factor is working memory load. Previous research on the involvement of working memory in verification of quantified expressions has shown that proportional quantifiers like *more than half* and *most* require higher working memory load than other types of quantifiers (Szymanik and Zajenkowski 2010; Zajenkowski et al. 2011). Furthermore, Steinert-Threlkeld et al. (2015) present data suggesting that verification of *most* and *more than half* may involve working memory to a different extent. The set-up used by Hackl, a self-paced counting task, would be expected to put significant strain on working memory – subjects did not have access to the whole visual scene they had to verify. As Steinert-Threlkeld et al. (2015) point out, the mode of presentation (paired vs. random stimuli, dots vs. letters) impacts the degree of interaction between working memory and verification; for *most* and *more than half*, this impact is different.

Even if the results are taken at face value, it is not clear what this relationship is like without detailed algorithmic-level predictions of how different specifications of truth conditions are computed.

Motivated by these considerations, we will present the results of an experimental study

comparing the verification profiles of *most* and *more than half*, which replicates Hackl's original setup (specifically, his Experiment 1) with some modifications. The results of our experiment suggest that there are some consistent differences in how speakers verify *most* and *more than half* – the former tends to rely more on approximation, and the latter tends to trigger a more precise strategy. However, these differences do not seem to originate from the different specifications of truth condition associated with these quantifiers – the verification patterns we have observed were inconsistent with algorithmic-level predictions based on the specifications of truth conditions alone. Instead, the choice of a particular strategy in our task depended on various factors: both *most* and *more than half* were impacted by individual working memory capacity and the changes in experimental setup from Hackl (2009) that elicited more approximative procedures.

More importantly, our exploratory analysis revealed that three groups of participants can be distinguished based on their preferred strategy. This points to the fact that verification procedures are individualized and flexible – they depend on the type of the task and input, as well as on cognitive resources of subjects. All of these considerations lead us to suggest that instead of triggering a particular default procedure, each quantifier is associated with a collection of verification strategies. Among others, this idea has been previously proposed by Suppes (1982), who argued that the meaning of a sentence can be treated not just as one procedure, but as a collection of those (see also Szymanik 2016). We can expect, then, that some of these procedures overlap for *most* and *more than half*.

2 Current study

2.1 Predictions

Hackl (2009) argues that verifying *most* requires participants to keep track only of the color that is leading at any given moment, the so-called lead-counting strategy. Verifying *more than half*, on the other hand, does not rely on lead-counting: it instead requires keeping track of how many dots the subjects saw in both colors and then comparing the two quantities – using either precise calculations or approximation. This latter procedure is more demanding, as reasoners would have to store a bigger amount of information in their working memory, as well as performing manipulations with it.

Given these considerations, we expect that participants with higher working memory capacity will have shorter reaction times for *more than half* (compared to participants with lower working memory scores), because these participants will be better able to cope with the additional load. At the same time, we expect that subjects with higher memory scores will make fewer mistakes when verifying sentences overall, resulting in a smaller gap in accuracy between *most* and *more than half*[1].

Prediction 1. The higher working memory capacity, the smaller will be the RT effect (the difference in reaction times between most and more than half) and the smaller the accuracy effect (difference in accuracy).

Our second prediction tests Hackl's hypothesis that *most* requires a lead-counting strategy and that a self-paced counting paradigm facilitates such a strategy. On each trial, we coded one screen to have an advantage for the target color: after viewing the first few increments of the scene, which were ambiguous as to what color was leading, subjects saw a screen that

[1] Although overall difference in accuracy was not significant in Hackl's study, subjects were more accurate when verifying *more than half*.

contained a clear advantage for the target color. Assuming that this advantage is irrelevant for verifying *more than half*, we make the following prediction:

Prediction 2. Reaction times on the target screen (namely, screen 4) for *most* will be significantly lower than reaction times on the target screen for *more than half*.

2.2 Participants

Thirty-five (8 female, 24 male, 1 genderfluid) subjects were initially recruited for the study via Prolific.ac, all native speakers of English. Participants were between 18 and 35 years old and were located in the United States. They viewed the experiment in their web browsers, and the average completion time was 14 minutes. Subjects received £2.50 as compensation.

2.3 Materials

The experiment consisted of two sections. In the first section, the digit span task (Schroeder et al., 2012), subjects had to memorize sequences of digits and reproduce them in reverse order. We administered this task as a measure of the working memory capacity of each participant. In the second section, the quantity judgment task, participants had to compare statements such as *Most of the dots are blue* and *More than half of the dots are blue* against visual stimuli, as in Hackl (2009). They were required to press a button for whether the sentence matches the visual stimulus or not. The experiment consisted of 24 target items: 12 sentences with the quantifier *most* and 12 with the quantifier *more than half*. In each group of sentences, 6 of the statements were true (i.e., when the subjects saw the statement *Most of the dots are blue*, it was followed by a scene that matched that description) and 6 were false. The experiment also included thirty-six fillers – sentences with non-proportional quantifiers such as *At most six of dots are yellow*, *Some dots are blue*, *Few dots are green*, etc.

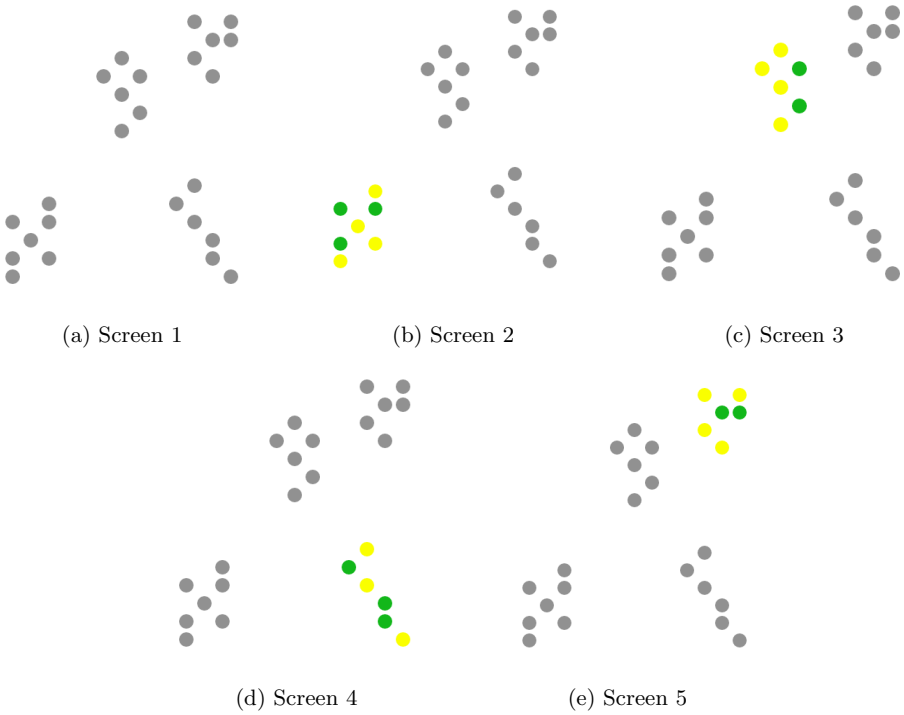
The visual stimuli consisted of pictures of dots scattered across the screen. On the trials with the two quantifiers of interest, it was never clear whether the statement was true or false until screen 5, the last screen. In our analyses below, we focus only on the first 4 screens. On some of the filter trials, it was clear whether the statement was true before reaching the last screen, on the second or third screen. Subjects were informed that they could press one of the answer keys at any point during the trial (see Talmina 2017 for further details).

The main difference between the current setup and Hackl (2009) was the position of the dots – in our experiment they were scattered (see Figure 1), while in Hackl (2009) they appeared in two rows.

2.4 Procedure

For the digit span task, participants saw sequences of digits. Each digit was displayed for 1000 ms. At the end of the sequence participants were given as much time as they needed to fill in the digits in the reversed order. The length of sequences gradually increased and the task ended at the point when participants made three errors in a row. No feedback was given to participants about their performance.

For the sentence judgment task, at the beginning of each trial, subjects saw the sentence that they had to verify in 24pt font on their screen. The time of presentation was not limited, and the subjects had to press the spacebar to proceed to the image. In the first screen (see Figure 1 for an example of a sequence of screens), only the outlines of the dots were visible. The subjects had to press the spacebar to move through the screens, gradually exposing the colors of dots.



(a) Screen 1 (b) Screen 2 (c) Screen 3

(d) Screen 4 (e) Screen 5

Figure 1: Example of a sequence of events in a trial.

When subjects uncovered a new screen, the dots they had previously seen were covered again. They also were not allowed to go back between screens. Participants were informed that they could respond during any part of the trial by pressing the "Y" key (for "yes") on their keyboard if they thought the sentence matched the visual stimuli or the "N" key (for "no") if they thought the sentence was not a correct description of the visual stimuli.

We recorded the information about the time it took the subjects to press a key on every screen, which key was pressed, and whether their response on every trial was correct (but no feedback was given to participants about these).

2.5 Results

Subjects made slightly more mistakes with *most* (in 17.9% of all *most* items) than with *more than half* (in 13.8% of all *more than half* items), but the difference was not significant (Wilcoxon rank-sum test $W = 81576; p = 0.1204$). When analyzing reaction time data, we only looked at the correctly responded trials. Overall reaction times were significantly affected by quantifier: participants took longer verifying *more than half* than *most* when looking at total reaction times ($W = 2043000; p < 0.001$). The latter finding differs from Hackl's, who interpreted the lack of significant overall difference to mean that subjects treated the two quantifiers as equivalent.

Hackl (2009) also reported a difference in reaction times for the first 4 screens (when collapsing across two quantifiers) where participants cannot yet make a decision about whether

the visual scene matches the sentence. In his study, the reaction times gradually increased from screen 1 to screen 4. In our own study, there was also a main effect of screen (Kruskal-Wallis test $H(3) = 140.39; p < 0.001$), but we do not observe such a gradual increase per screen. Instead, in our study participants took significantly longer on screen 1 in comparison to other screens. Focused comparisons of the mean ranks between screens support that: the significant differences were found mostly between screen 1 and other screens.

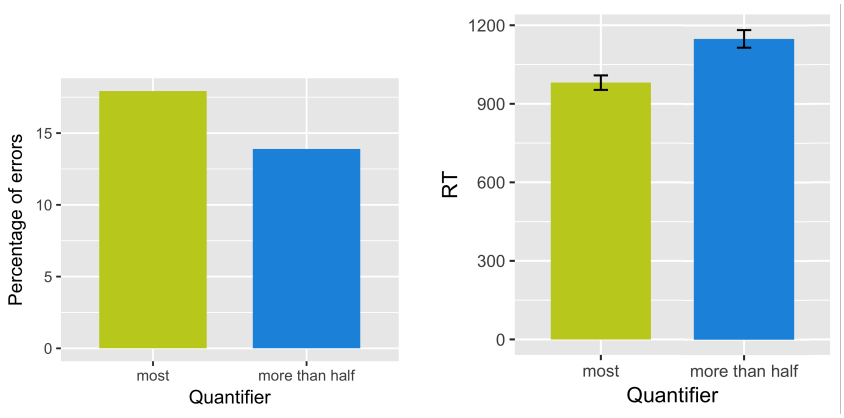


Figure 2: Percentage of errors per quantifier (left plot) and overall reaction times (right plot). The error bars for the reaction times indicate the standard error value.

We further investigated whether there was a screenwise difference in the RT effect – i.e., whether the discrepancies in reaction times between *most* and *more than half* were significantly larger on some screens than on others. To do so, we collected mean reaction times for every subject per every screen, and calculated the difference between *most* and *more than half*. However, no significant differences were found. Thus, the difference between two quantifiers was of similar size for all screens. This goes against Hackl's suggestion that *most* favors lead counting and the prediction that we subsequently made based on this that on the target screen, screen 4, the difference between *most* and*more than half* should be larger than on other screens.

To investigate the relationship between working memory capacity and accuracy and RT effects, we assigned a memory score to every subject based on the length of digit sequences they could remember in the reverse digit span task. We found a negative correlation between memory score and accuracy effect (Pearson's $r = -0.22$[2]), suggesting[3] that a higher memory score is related to lower difference in accuracy between *more than half* and *most*. Similarly, there was some negative correlation between memory score and RT effect (Pearson's $r = -0.2$), which also points to a connection between working memory capacity and lower processing times for more than half, in line with our predictions.

We have noted that the reaction times in the first screen were considerably longer than on other screens (and than RTs in the first screen in Hackl 2009). This might be explained by subjects attempting to estimate the total number of dots (or count them) before proceeding with the task. However, as we can see from Figure 4, not all participants behaved in this way: while some spent over 3000 milliseconds looking at the first screen, others were fast and

[2]We do not provide p-values as they are not considered to be diagnostic for correlation analyses. We only examine the correlation coefficient itself.

[3]See Talmina (2017) for a discussion of why we consider this result to be meaningful.

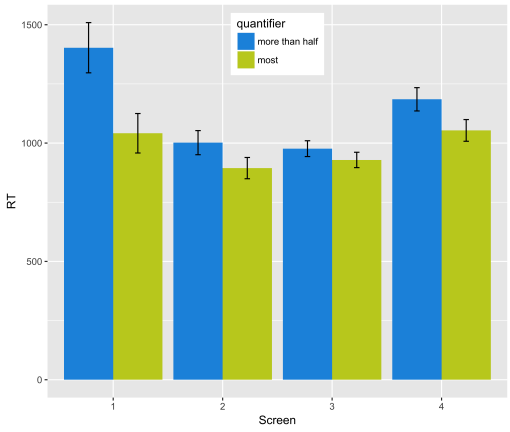


Figure 3: Mean reaction times per screen. The error bars indicate the standard error value.

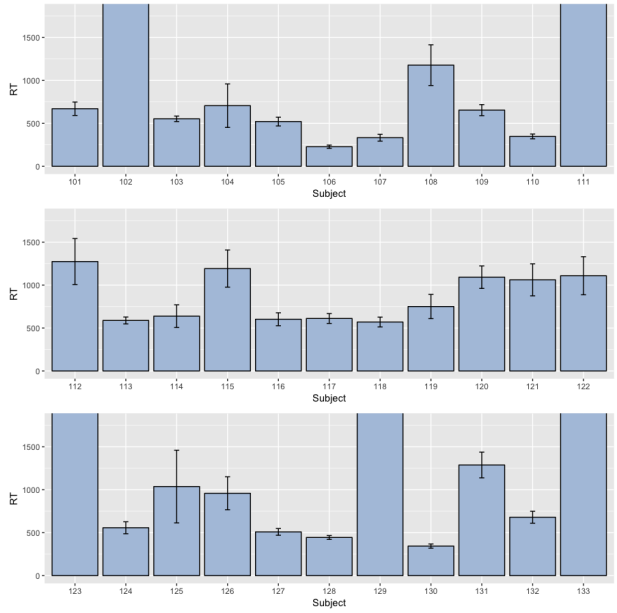


Figure 4: Mean RTs in the first screen for each participant.

took around 500 milliseconds to proceed to the next screen. Another group was in the middle: subjects who spent around 1000 milliseconds in the first screen on average.

As we've argued before, estimating the total number of dots requires additional executive resources, and is most likely justified when participants need to know precisely how many blue dots there need to be for a statement like *More than half of the dots are blue* to be true. In other words, people who look at the first screen for longer are probably going to use a more precise strategy. To explore this intuition, we divided our subjects into three groups based on average

time spent on the first screen: the "counting" group (on average, > 2000 ms spent on the first screen) who we suspect used a precise strategy throughout the whole experiment, the "mixed" group ($1000 - 2000$ ms) who we believe used a mixture of precise and approximative techniques, and the "fast" group (< 1000 ms) who were likelier to use an approximative strategy.

In the "counters" group (5 subjects), we found that there was a significant difference in accuracy between *most* and *more than half* ($W = 2100; p = 0.0148$), but no difference in reaction time ($W = 20757; p = 0.4057$). In the "mixed" group (9 subjects), we found no effect of quantifier on overall accuracy ($W = 4608; p = 1$), but *most* and *more than half* differed significantly in reaction times ($W = 81854; p = 0.012$). In the "fast group" (19 subjects), there was no effect of quantifier on accuracy ($W = 26676; p = 0.4694$), but it was a significant factor for reaction times ($W = 298730; p = 0.00042$).

This analysis was done in a post-hoc manner, so the existence of distinct strategies and their characteristics should be examined in future studies intentionally looking at this aspect.

3 Discussion and conclusion

The results of our experiment diverge in several respects from Hackl (2009). Hackl found no significant overall differences in RTs and accuracy between *most* and *more than half*, which he suggested served as evidence that subjects "treat the two expressions as essentially equivalent" when they are faced with a self-counting task. In the present study, however, we have found that mean difference in overall RTs between the two quantifiers was significant: participants took less time to verify *most* than *more than half*, suggesting that they did not treat in fact the two expressions as equivalent.

We also found a different pattern of screen-by-screen RT change: while Hackl observed a linear decrease in RTs as subjects proceeded through the scene, in our experiment RTs were highest in the first screen, then dropped in screen 2 and remained stable on screen 3, increasing again in screen 4. We built upon this finding to show that participants in our study used distinct verification strategies: while some started moving through the scene right away and relied on more approximative strategies, other spent more time on screen 1, where only the outlines of dots were visible, to count how many there were (see Figure 4).

Further, we investigated Hackl's hypothesis that the verification of *most* requires a lead-counting strategy. We have argued that if such a strategy is employed, subjects would take advantage of additional cues about the leading color that we have supplied in the target screen. As we have found no difference between the presence or absence of the RT effect (the difference in mean RTs between *most* and *more than half*) on the target screen compared to other screens, it is difficult to explain why the presentation mode would help subjects implement a lead-counting strategy, but they would ignore other cues.

The tendency for negative correlation between a subject's memory score and the accuracy effect (the difference between how many errors a subject made when verifying *more than half* and the number of errors she made when verifying *most*) suggests that the higher a subject's working memory capacity was, the fewer mistakes they made when verifying *most*. Higher working memory capacity possibly allowed participants to use a more precise strategy when verifying *most*. Similarly, it allowed reasoners to process *more than half* faster: the algorithm for this quantifier requires storing two numbers in memory, which was easier to do for participants with higher cognitive resources.

In summary, only Prediction 1 (the effect of working memory capacity on accuracy and RT effects) was supported by the data. Prediction 2 was not supported by the data, as we found no significant differences in RTs between *most* and *more than half* on the target screen.

However, we learned other things from our results. Compared to Hackl's results, we do find a difference in overall reaction times between the two quantifiers, which suggests a possible difference in their meanings as well. Moreover, our results highlight potential role of individual differences in competing these tasks (or, perhaps, in verification strategy that is used) which should be investigated in future studies. Hence, we believe that Hackl's experimental conclusions may be premature. There is not enough evidence to claim that *more than half* and *most* are semantically equivalent and associated with different verification strategies. The situation may be much more sophisticated. Each of these quantifiers may be associated with a collection of potential verification strategies which are used in different proportions by different subjects. The choice of the verification strategies depends not only on the quantifier but also on the context, e.g., experimental setup. Crucially, the verification strategies subjects choose seem to be sensitive to general cognitive constraints, like working memory.

References

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

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

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

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

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., and Marshall, P. S. (2012). Reliable digit span: A systematic review and cross-validation study. *Assessment*, 19(1):21–30.

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

Suppes, P. (1982). Variable-free semantics with remarks on procedural extensions. *Language, Mind and Brain*, pages 21–34.

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

Szymanik, J. and Zajenkowski, M. (2010). Quantifiers and working memory. In *Logic, Language and Meaning*, pages 456–464. Springer.

Talmina, N. (2017). Quantifiers and verification strategies: connecting the dots (literally). Master's thesis, ILLC, Universiteit van Amsterdam.

Zajenkowski, M., Styła, R., and Szymanik, J. (2011). A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders*, 44(6):595–600.