Quantifiers and verification strategies: connecting the dots (literally)

MSc Thesis *(Afstudeerscriptie)*

written by

Natalia Talmina (previously Natalia Philippova)
(born May 7th, 1990 in Moscow, Russia)

under the supervision of **Jakub Szymanik** and **Arnold Kochari**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
June 19th, 2017	Dr. Floris Roelofsen (chair)
	Prof. Dr. Robert van Rooij
	Dr. Jakub Dotlačil
	Dr. Jakub Szymanik
	Arnold Kochari, M.A.

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

The meaning of natural language expressions is usually identified with the conditions under which this expression is true. An alternative view – the procedural approach to meaning – identifies the meaning of an expression with an algorithm (or algorithms) for judging whether the expression is true or false. However, the relationship between meaning and verification is a complex one: as Hunter et al. (2016) argue, identifying a verification procedure with the truth conditions of an expression is an oversimplification. Instead, several authors have suggested that meanings come with *verification weight* that makes certain verification strategies more preferable by default, even when the context of a task would make a different strategy more accurate or efficient. An experimental study by Hackl (2009) illustrates this point by providing evidence that quantifiers *most* and *more than half*, albeit truth-conditionally equivalent, trigger distinct default verification profiles.

The problem with this type of evidence, however, is that a number of confounding factors can interact with the choice of a verification strategy: differences in subjects' cognitive resources, the type of linguistic input, and the kind of task at hand, just to name a few. In this thesis, we will present the results of two experimental studies that partially address this problem by controlling for individual executive resources and making explicit predictions about the strategies underlying the verification of *most* and *more than half*. We will argue that while there are differences in how subjects verify *most* and *more than half*, they do not result in completely distinct patterns. Finally, we will propose a different approach to the relationship between the meaning and verification of quantifiers. We will propose that instead of corresponding to one default verification strategy, quantifiers are associated with a collection of strategies, some of which overlap for different quantifiers. The choice of a strategy among those is ultimately defined by multiple factors, such as context, the task at hand, personal preferences and resources, and the type of input.

Acknowledgements

I hope my supervisors, Jakub Szymanik and Arnold Kochari, know what a privilege it has been to work under their guidance. Whenever I came into Jakub's office with a heap of swirling semi-ideas (which was, admittedly, more than half of the time), he would somehow makes sense of them, and I would leave with respectable predictions and hypotheses. Arnold was incredibly generous with his time and expertise; his optimism and enthusiasm made me believe that even the most frustrating challenges were not insurmountable.

I would like to thank the members of my thesis committee, Robert van Rooij, Jakub Dotlačil, and Floris Roelofsen, for their interest in my work and for the inspiring discussion during the defense. I am additionally thankful to Floris for being my academic mentor throughout the past two years and always taking the time to give me thoughtful advice when I needed it most. At the ILLC, thanks also go to Fenneke Kortenbach and Tanja Kassenar for their help with any administrative issue that ever came up.

To my community in Amsterdam: my life here has been full of laughter and joy thanks to you. Alison, your wisdom and eloquence have been a continuous source of strength. Laura, thank you for your incredible warmth and all the cake. Thank you, Lisa, for all of our coffee-break conversations and Italian dinners. Zoi, thanks for being a great travel companion. Levin, Lucy, Jonathan, Johannes, Melina, Jakob, Bonan, Heidi – you are part of some of my happiest memories in Amsterdam.

My friends in Russia deserve a special thanks for always being there for me. Lena, none of this would have been possible if you hadn't become my math tutor all these years ago. Daria, your friendship has been a constant through distance and time – thank you.

Finally, to my family: thank you for always trusting my judgment and giving me the courage to venture into the difficult and the unknown. I am endlessly grateful for your unconditional support, patience, and love.

Contents

1 Introduction

Connections between language and other cognitive functions are ubiquitous, but intricate: the more we find out about them, the more questions we face.

Take bilingualism as one of the more prominent examples. A growing body of research supports the claim that bilingualism has lifelong benefits: it has a positive effect on the development of executive control in children (see Bialystok 2009 for overview), but also offers protection against cognitive decline that sometimes occurs at older age (Bialystok et al., 2007).

Another well-known example is the effect of color terms on color perception. Unlike English, Russian has two main color terms for *blue*: *siniy* (dark blue) and *goluboy* (light blue). A study by Winawer et al. (2007) revealed that Russian speakers were faster to discriminate two colors that belonged to different linguistic categories in Russian (i.e. if one color was considered *siniy* and the other *goluboy*) than colors that were from the same linguistic category (both *siniy* or both *goluboy*). English speakers did not show this advantage.

There are connections that appear to run even deeper, such as the apparent co-developmental relationship between syntax and Theory of Mind. Hale and Tager-Flusberg (2003) conducted a training study in which children were trained on comprehension of sentential complements (such as *Mary thought that John was working in the garden*). Their performance on Theory of Mind tasks was significantly better than before training – and better than the control group, who were trained on relative clauses instead.

Then there is the relationship between language and number cognition. The very fact that we can perform complex mathematical operations like multiplication and division is due to the fact that we have number words, which help us discriminate between quantities that only have the slightest difference. Decimals, one thousands, square roots and milliseconds are some of the things we can understand because we have words for them. But the effect of language on our mathematical skills does not end there. A growing body of research suggests that Specific Language Impairment impacts the development of mathematical skills in children. Donlan et al. (2007) found that children diagnosed with SLI had difficulties with counting and calculation, but they were able to grasp the logical principles underlying simple arithmetic. Newton et al. (2010) found that SLI performance on the reduced array selection task was above language controls.

There is a lot of uncertainty in each of these lines of inquiry. Some of this uncertainty concerns itself with what feels like really big questions: if language can affect how we perceive the colors of the world, what else about it can be different for a German speaker and an Urdu speaker? How do we go about finding out without repeating some of the more unfortunate mistakes of Benjamin Lee Whorf?[1]

Other uncertainties are relatively small, but ever the more interesting. Consider the sentences below:

(1) a. Most Russians enjoy watching sports.
    b. ??More than half of Russians enjoy watching sports.

(2) a. More than half of Russians were assigned female at birth.
    b. ??Most Russians were assigned female at birth.

Intuitively, *more than half* does not fit into the sentence in (1b) – if it's not infelicitous, it at least sounds odd. Similarly, the sentence in (2b) reads like a false, or at least a misleading statement – but the same sentence with *more than half* instead of *most* is fine. Yet, the logical form of these two quantifiers is indistinguishable – and so should be their meanings.

Admittedly, postulating a profound difference between *most* and *more than half* might be a bit audacious based on these two examples alone. Questions about examples above can be dismissed and redirected into the realm of semantics/pragmatics interface. However, experimental results from a study by Hackl (2009) suggest that speakers use distinct verification strategies to judge whether sentences that involve these quantifiers are true or false. What we still don't know is what exactly that means.

The relationship between meaning and verification is a complex one. If knowing the meaning of an expression is knowing the conditions that make this expression true, does this entail knowing how to verify whether these conditions are met? We will get into the technical details in the following section, but consider for a second that a statement like *Most A's are B* is true if there are more A's that are B than there are A's that are not B. We can capture this set-theoretically as $|A \cap B| > |A - B|$. Then, we can suggest that when speakers verify whether *Most A's are B* is true they count the A's that are B, count the A's that are not be, and compare the two quantities.

[1] See Pullum (1991) for discussion.

The idea we have just outlined is that there are *default verification strategies* associated with a meaning of an expression. According to the proponents of this idea (Lidz et al., 2011; Pietroski et al., 2009), the meaning of an expression, in addition to providing truth conditions, has some *verification weight* that makes a certain verification procedure a more compelling choice even when circumstances make a different strategy more precise.

However, there are multiple reasons why reasoners may prefer to consistently use a particular strategy in a verification task. The nature of the task itself might make this strategy convenient, or they might develop a cognitive bias by adhering to the same strategy throughout the experiment. Finally, it is not exactly clear how default verification strategies intersect with individual differences in cognitive control, such as inhibition and working memory.

While there are plenty theoretical arguments to support either claim, more empirical evidence is needed to provide more clarity about the relationship between meaning and verification. In this thesis, we are going to present results of two experimental studies whose goal was twofold: 1) to compare the verification profiles of *most* and *more than half* based on explicit predictions we formulated from prior research, and 2) to see whether the patterns we observed would persist across different settings and conditions. Based on the results of these experiments, we will argue against identifying the meaning of quantifiers with default verification procedures; instead, we will propose that quantifiers are associated with a collection of procedures, some of which may overlap for different quantifiers.

Thesis overview This thesis is structured as follows: in Chapter 2, we will lay the theoretical groundwork for our investigation. We will give a brief overview of Generalized Quantifier Theory, discuss the relationship between meaning and verification in more detail, and summarize the predictions of semantic automata theory about the involvement of working memory in the processing of proportional quantifiers. In Chapter 3, we will give a synopsis of several experimental studies that shed light on what processes underly the verification of *most* and *more than half*. We will discuss working memory and the Approximate Number System in more detail. In Chapter 4, we will provide the results of two experimental studies we have carried out. In the concluding Chapter 5, we will argue that quantifiers provide multiple verification strategies, and reasoners can switch between them.

2 Background

Before looking into the differences between quantifiers *most* and *more than half*, we will lay some groundwork by reiterating how two relevant semantic frameworks – Generalized Quantifier Theory and semantic automata theory – capture the meaning of these expressions. We will investigate why the indiscernibility of *most* and *more than half* in Generalized Quantifier Theory is problematic, and discuss the relationship between truth conditions of an expression and the verification strategies competent speakers employ to make sure that these conditions are met. We will also show that semantic automata theory sheds light on certain aspects of quantifier verification that might be crucial for inquiring into the connection between meaning and verification.

2.1 Generalized Quantifier Theory

Consider the sentence in (3a) below that contains the universal quantificational determiner *every*. This quantifier, along with other Aristotelean (*no*, *some*) and cardinal (*at least 3*, *more than 7*) quantifiers, can be easily expressed as a first-order relation between students and exams, such that for every student there is an exam that they dread, as captured in (3b).

(3) a. Every student dreads some exam.

b. $\forall x(\text{student}(x) \rightarrow \exists y(\text{exam}(y) \wedge \text{dreads}(x, y))$

However, not all types of quantifiers lend themselves so easily to such a convenient way of semantic notation. As was shown by Barwise and Cooper (1981), proportional quantifiers such as *most*, *more than half*, *two thirds*, etc. are not definable in first-order terms – there is no meaningful expression we could form out of variables, non-logical constants, and symbols $\exists, \forall, \neg, \vee, \wedge, \rightarrow$ to form an expression that would capture their meaning. Yet, these quantifiers form meaningful sentences fairly frequently in natural language[1], and some solution was necessary to solve this problem – some additional apparatus to enrich the expressive power of first-order logic.

The solution now commonly adopted in linguistics is Generalized Quantifier Theory (GQT), developed by Mostowski (1957) and Lindström (1966). We will give the definition of a generalized quantifier as formulated in Szymanik (2016):

Definition 1. A generalized quantifier Q of type $t = (n_1, \ldots, n_k)$ is a function assigning to every set M a k-ary relation Q_M between relations on M such that if $(R_1, \ldots, R_k) \in \mathsf{Q}_\mathsf{M}$, then R_i is an n_i-ary relation on M, for $i = 1, \ldots, k$. Additionally, Q is preserved by bijections, i.e., if $f : M \rightarrow M'$ is a bijection, then $(R_1, \ldots, R_k) \in \mathsf{Q}_\mathsf{M}$ if and only if $(fR_1, \ldots, fR_k) \in \mathsf{Q}'_\mathsf{M}$ for every relation $R_1, \ldots, R_k$ on M, where $fR = (f(x_1), \ldots, f(x_i))|(x_1, \ldots, x_i) \in R$, for $R \subseteq M^i$ (Szymanik, 2016).

According to this definition, a generalized quantifier is a function that maps a model $\mathbb{M}$ to a relation between relations on its universe M. Importantly, these relations are assumed to be semantic primitives (cf. Hackl 2009).

One consequence of this is GTQ's insensitivity to form: the internal composition of a quantifier does not affect their semantic behavior – i.e., distinct specifications of the same truth conditions are treated as equivalent. What this means is that as long as quantifiers express the same relation between sets, they are virtually indistinguishable in GQT. Take, for example, the expressions *at least 3* and *more than 2*: although there are systematic linguistic differences between the two (Geurts et al., 2010; Geurts and Nouwen, 2007; Hackl, 2000; Solt, 2016) – which possibly affect how speakers process sentences that involve these quantifiers – they share the same truth conditions, and are therefore treated as the same expression in GQT.

(4) a. $[\![\text{at least 3}]\!](\mathrm{A})(\mathrm{B}) = 1$ iff $|\mathrm{A} \cap \mathrm{B}| \geq 3$

b. $[\![\text{more than 2}]\!](\mathrm{A})(\mathrm{B}) = 1$ iff $|\mathrm{A} \cap \mathrm{B}| > 2$

As Hackl (2009) observes, the same problem arises when we consider multiple possible renditions of the truth conditions of some quantifier. As he shows for *no*, there are multiple equally good descriptions in GQT that are not discernible from each other:

(5) a. $[\![\text{no}]\!] = 1$ iff $\mathrm{A} \cap \mathrm{B} = \emptyset$

b. $[\![\text{no}]\!] = 1$ iff $|\mathrm{A} \cap \mathrm{B}| = 0$

c. $[\![\text{no}]\!] = 1$ iff $|\mathrm{A} \cap \mathrm{B}| < 1$

Similarly, there are multiple ways of expressing *most* and *more than half* in GQT. Moreover, as these two quantifiers are truth-conditionally equivalent, all descriptions that fit *most* also fit *more than half*.

[1] See Szymanik and Thorne (2017) for an estimation.

(6) a. $[\![\text{most}]\!] = 1$ iff $|A \cap B| > |A - B|$

b. $[\![\text{most}]\!] = 1$ iff $|A \cap B| > |A|/2$

(7) a. $[\![\text{more than half}]\!] = 1$ iff $|A \cap B| > |A - B|$

b. $[\![\text{more than half}]\!] = 1$ iff $|A \cap B| > |A|/2$

However, while these renditions of truth conditions are essentially equivalent, Hackl (2009) argues that conceptually, (7b) is a better way to capture the truth conditions of *more than half* – as it directly calls for dividing the total number of objects in half; and on the other hand, that (6a) is a more accurate description of *most*, which Hackl (2009) suggests treating as a superlative form of *many*.

2.2 Truth conditions and verification strategies

According to one of the most influential ideas in natural language semantics, knowing the meaning of any natural language expression – for instance, the sentence in (8) below – amounts to knowing the conditions under which the sentence is true.

(8) Most of the dots are blue.

These truth conditions can be seen as functions from contexts to truth values: if there are more blue dots than dots of other colors in a given picture or scene, the function corresponding to the truth conditions of (8) would map this context to true. Conversely, if a picture contains more non-blue dots than blue ones, that context would be mapped to false.

However, while this idea is fairly straightforward, it is not apparent from the truth conditions alone how exactly that function is executed (Pietroski et al., 2009; Steinert-Threlkeld et al., 2015); if, after looking at an image depicting yellow and blue dots, a competent speaker of English confirms that (8) is true, how did she verify that the relevant truth conditions have been met? We can think of several viable options: she could simply count all the dots and the blue dots, then subtract the latter from the former, or she could count the blue dots and the non-blue dots and compare those cardinalities, or try to estimate any of these number without appealing to any arithmetical operations. As Pietroski et al. (2009) point out, even asking a friend who is sitting nearby for a solution would count as a verification strategy. The choice of a strategy can depend on many factors such as the type of expression being verified (i.e. verifying a negated statement is not the same as verifying quantified expressions), the kind of stimuli (countable objects vs. mass), and the time for which the stimuli is presented, among others.

The important part, however, is how these strategies relate to the formal specification of truth conditions in (6) and (7) – and this relationship is not straightforward. It might be tempting to assume that the truth conditions in (6) correspond to particular verification strategies: after all, they explicitly ask for subtracting the number of blue dots from the total number of dots or dividing the total number of dots in half and comparing the result with the number of blue dots – which are, we have argued, both possible strategies to verify whether the sentence *Most of the dots are blue* is true. Indeed, there is evidence that the relationship between truth conditions and verification strategies is constrained (see Dummett 1973; Horty 2007; Suppes 1980 for discussion and Lidz et al. 2011; Pietroski et al. 2009; Steinert-Threlkeld et al. 2015 for experimental results). To make the difference between the two notions clearer, we will follow Lidz et al. (2011); Odic (2014); Pietroski et al. (2009), among others, in making an analogy to Marr's levels of computation (Marr, 1982).

So far, we have discussed formal properties of natural language quantifiers. The primary goal of formal grammars is to provide us with *computational-level* descriptions of what we are computing: for example, dependencies and movement in the case of syntax, truth conditions in the case of semantics. These descriptions aim to be as abstract as possible and describe speakers' competence without specifying how exactly certain functions are computed. On the other hand, experimental studies concern themselves primarily with *algorithmic-level* questions about how certain functions are implemented in the brain: how we perceive information and memorize parts of it, process strings of sounds and translate them into sentences, perform calculations or spatial orientation tasks (Odic, 2014).

Still, these levels of description are not essentially disparate, nor are they meant to exist independently from each other – to the contrary, formal grammars provide strong foundations for algorithmic-level hypotheses that can be tested experimentally, and vice versa, experimental results can inform semantic theories. Even though it is implausible that meaning could be equated with verification, there is a possibility that "meanings are individuated at least as finely as truth-procedures" (Pietroski et al., 2009, p. 561). In other words, if abstract computational-level descriptions of the language system provide several plausible descriptions of grammar – in our case, equally plausible specifications of the truth conditions – these can inform algorithmic-level hypotheses about which functions are actually implemented by human cognition.

If a function X turns out to be preferred over function Y, it does not necessarily imply that the specification of truth conditions whose make-up follows more closely the calculations necessary to perform function Y is not a valid option. To give an example, if we find out that competent speakers prefer to verify (8) by dividing the total number of (blue and non-blue) dots in half, and then comparing that numerosity with the number of blue dots, it doesn't mean that alternative renditions of the truth conditions for *most* are incorrect, or that competent English speakers never use any other verification strategies – where they check, for example, whether there is a non-blue dot for every blue one. What this means, instead, is that we have grounds to speculate that there are canonical specifications of truth conditions that constrain the choice of a verification strategy: given other choices, speakers are more likely to pick the verification strategy that is closely related to the canonical ways of computing the relevant truth conditions.

The question about capturing the truth conditions of some expression, thus, is on a different level from the question about how the meaning of that expression is verified in a given context. However, these two questions can inform each other, and the relationship between truth conditions and verification procedures is likely constrained: canonical specifications of truth conditions can be used as default verification procedures.

2.3 Quantifier verification: a view from automata theory

Semantic automata theory identifies a quantifier of the class Q with a language L_Q describing all elements (models) of the class class Q corresponding to a quantifier and a machine M_Q that computes the truth-conditions of a sentence containing a quantifier of the class Q. In line with the procedural approach to meaning[2], quantifiers from Definition 1 can be alternatively seen as classes of models. Quantifiers that are defineable in first-order logic – for instance, Aristotelean (*all*, *some*, *no*) and cardinal (*at most 5*, *at least 4*) – are recognized by finite-state automata (FSA); proportional quantifiers, however, are not definable in first-order logic and require a pushdown automaton (PDA), which augments finite-state automata with a memory stack (van Benthem, 1986; Mostowski, 1998).

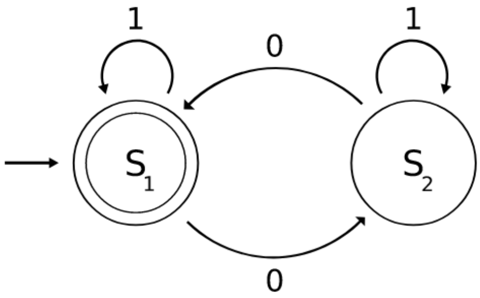


FIGURE 2.1: Example of a finite state automaton

This distinction is not merely theoretical. McMillan et al. (2005) tested the idea that pushdown automata can be viewed as verification procedures internalized by competent speakers. The authors hypothesized that the differences in complexity between these two types of quatifiers would be reflected in processing: verifying higher-order quantifiers associated with push-down automata would place higher demands on reasoners' working memory than verifying sentences with first-order definable quantifiers, which correspond to FSAs. This hypothesis was confirmed: the authors found that only higher-order quantifiers recruit the prefrontal cortex associated with the executive function, which includes working memory. As we will show in Chapter 3, further experimental studies have confirmed that processing proportional quantifiers requires an additional memory load.

[2] According to this approach, algorithms that competent speakers rely on to judge whether a natural language expression is true or not can be identified with meanings. See Szymanik (2016) for discussion.

2.4 Conclusion

In this section, we have outlined aspects of semantic theory of quantifiers that raise some questions about the meaning and verification of *most* and *more than half*.

We have presented Hackl's argument that Generalized Quantifier Theory is problematic when it comes down to discerning the meaning of quantifiers with equivalent truth conditions: in GQT, *most* and *more than half* are indistinguishable. We have noted that this is problematic for quantifiers that are systematically different in terms of semantics, pragmatics and verification.

We have also observed that there are several ways of specifying truth conditions for both of these quantifiers, and argued that this difference might be cognitively relevant for the way in which *most* and *more than half* are processed. We have noted that although truth conditions and verification strategies are not equivalent, it is plausible that there is a connection between the two notions.

The next cheaper builds up on the current discussion with a detailed and systematic overview of the verification profiles of *most* and *more than half*. We will give a summary of several experimental studies that 1) explicitly compare how reasoners verify sentences with *most* and *more than half* in the context of a self-paced counting task; 2) processing proportional quantifiers; 3) verification of *most* is consistent with a classic psychophysical model of the Approximate Number System.

3

Semantics and verification profiles of *most* and *more than half*

Having explored the relationship between meaning and verification in the previous section, we are now equipped to tackle the verification profiles of *most* and *more than half* in further detail. In this section, we will summarize the results of several studies that shed some light on the problem at hand. We will do so in three stages. First, we will reiterate the results of Martin Hackl's 2009 self-paced counting experiment, which explicitly investigated differences in default verification procedures underlying the meanings of these two quantifiers. Second, we will look into the experimental evidence of high working memory involvement in proportional quantifier verification. Finally, we will look at the verification profile of *most* in more detail and relate experimental results that show it exhibits certain properties of Approximate Number System. We will try to fit these three directions of research together to see what is still missing from our picture.

3.1 Differences in verification of *most* and *more than half*: Martin Hackl's experiment

In his 2009 paper, Hackl explores whether there is a cognitively significant difference in the specification of the truth conditions for *most* and *more than half* in (6) and (7), repeated below.

(9) a. $[\![\text{most}]\!] = 1 \text{ iff } |A \cap B| > |A - B|$

b. $[\![\text{most}]\!] = 1 \text{ iff } |A \cap B| > |A|/2$

(10) a. $[\![\text{more than half}]\!] = 1 \text{ iff } |A \cap B| > |A - B|$

b. $[\![\text{more than half}]\!] = 1 \text{ iff } |A \cap B| > |A|/2$

He argues on conceptual and linguistic grounds that (9a) is the preferred option over (9b) for *most*, while (10b) is a better way to express *more than half* than (10a) and 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, 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)[1].

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 in a step-by-step fashion (see Figure 3.1).

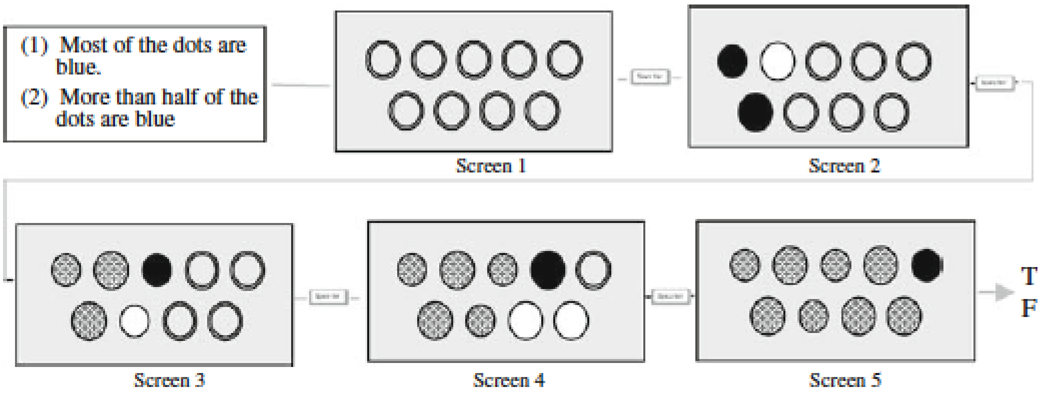

(1) Most of the dots are blue.
(2) More than half of the dots are blue
Screen 1
Screen 2
Screen 3
Screen 4
Screen 5
T
F

FIGURE 3.1: Sequence of events in the Self-Paced Counting paradigm

This setup allows to measure how much time participants spend processing information at each step. At the beginning of a trial subjects heard target sentences played over the speakers and saw two rows of dots displayed on the screen. At first, participants did not know the color of the dots, as only their outline was visible. As subjects pressed the space bar, increments of 2 to 3 dots were uncovered, and the previously seen dots were masked again. Subjects had to verify whether the sentence they had heard over the speakers was true or false. They could answer at any point during the trial by pressing the appropriate key on their keyboards, but the design of the experiment made it impossible to determine the truth or falsity of the sentence within the first four screens.

[1] Note that this is not the only way to specify truth conditions for *most*.

The overall scores (accuracy and reaction times) for *most* and *more than half* turned out not to be significantly different, which Hackl takes as evidence that participants treat these expressions as equivalent in the context of a self-paced counting task. The author found, however, that there was a significant screen-by-screen difference: verifying *more than half* took subjects consistently longer than *most*. Hackl observes that this difference makes sense if *most* favors a kind of lead counting strategy – checking whether there is a non-blue dot for every blue dot. The design of the experiment made the task easier for such a strategy: in each screen, it was easy to evaluate whether there were more dots in the target color than in the other color. Moreover, the self-paced setup made it easy to keep track of the difference.

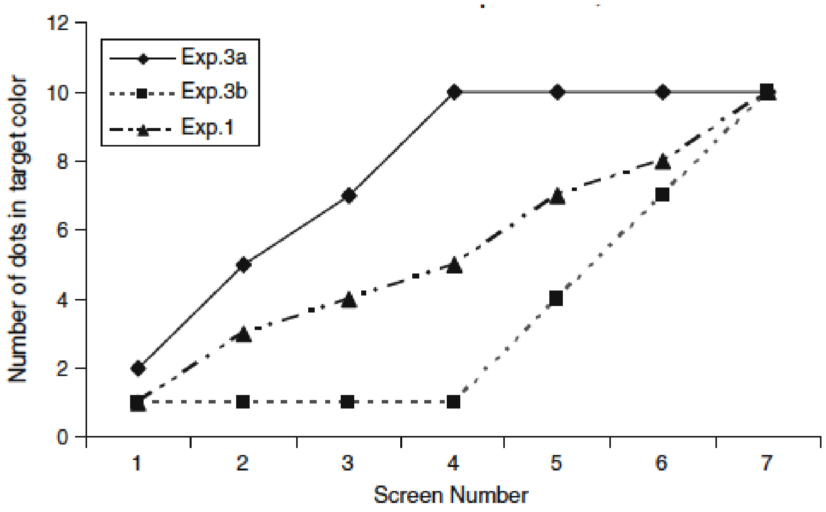


FIGURE 3.2: Item schema in Hackl (2009) showing how stimuli were distributed between screens in the "early" and "late" conditions

To follow up on this finding and provide further support for the hypothesis that *most* and *more than half* trigger distinct verification procedures, Hackl conducted another experiment in which he manipulated the arrays of dots in such a way that the verification procedure triggered by *most* was either facilitated or impeded, while the verification procedure triggered by *more than half* remained unaffected. In particular, the distribution of dots in the target color was manipulated: in condition (a), the "early" condition, nearly all dots in the target color were at the beginning of the trial, while in condition (b), the "late" condition, nearly all of them were at the end.

It was revealed that both quantifiers were affected by the distributional asymmetries of the target items; however, the verification strategy triggered by *most* was more sensitive to them. *Most* required more time in the "late" condition while the difference between the two conditions was not significant for *more than half*.

3.2 Verification tasks and working memory

As we have seen in the previous chapter, the distinctive feature of proportional quantifiers like *most* an *more than half* is that processing them requires some involvement of working memory. Working memory (WM) is a basic cognitive mechanism that can temporarily store pieces of information, as well as manipulate them for processing tasks (Baddeley, 2003; Miyake and Shah, 1999). A line of research has focused specifically on the degree of working memory load in quantifier verification tasks.

Szymanik and Zajenkowski (2010) examined the effects of working memory load in three groups of quantifiers: proportional, parity, and numerical. Based on the predictions of the automata-based quantifier verification model, they hypothesized that asking subjects to hold arbitrary information in short-term memory would disproportionally affect the difficulty of verifying these types of adjectives. In particular, the authors hypothesized that the difficulty would decrease in the following order: proportional quantifiers, numerical quantifiers of high rank, parity quantifiers, numerical quantifiers of low rank. The experiment consisted of two elements: the sentence verification task and the memory task. At the beginning of each trial, participants were asked to memorize a 4- or 6-digit string of numbers. After that, subjects had to judge the truth-value of sentences such as *More than half of the cars are red* or *An even number of cars are blue* against visual scenes presented on the screen. After completing the sentence verification task, they were asked to recall the string they had memorized.

The authors' hypothesis was confirmed in the 4-digit condition, and crucially, proportional quantifiers proved to be the most difficult, with the highest reactions times and the poorest accuracy. However, the differences between the considered types of quantifiers were not significant in the 6-digit condition. The authors observed a decrease of accuracy in numeric recall with simultaneous increase in performance on quantifier verification task, which they considered to be trade-off between processing and storage.

Zajenkowski et al. (2011) compared the processing of several groups of quantifiers in patients diagnosed with schizophrenia and a healthy control group. Participants had to verify sentences that contained natural language quantifiers, and patients with schizophrenia were consistently slower than controls on all types of quantifiers. However, the difference in accuracy was only significant for proportional quantifiers. Zajenkowski et al. (2011) suggested that the longer RTs allowed patients to verify Aristotelean, parity, and numerical quantifiers almost as accurately as the control group. However, slower processing did not result in the same match with the controls' accuracy. The authors suggested this is also due to the high engagement of working memory. As there is evidence that patients diagnosed with schizophrenia often have impaired executive function – especially control or the supervision of cognitive processes – it is possible that simultaneously processing and keeping stored information was too demanding for the patient group.

Steinert-Threlkeld et al. (2015) explored the impact of different presentations of a visual scene and working memory load on proportional quantifier sentence verification. Subjects were presented with two types of stimuli: objects were either scattered (i.e. spread randomly across the screen) or paired (objects appeared in pairs where one of them was the target – for instance, a yellow dot and a blue dot). Moreover, there were two types of objects: yellow and blue dots, and characters 'E' and 'F'. The participants were asked three types of questions: 1) *Are more than half of the dots yellow?*, 2) *Are more than half of the letters 'E'?*, and 3) *Are most of the dots yellow?*. The authors were interested whether distinct verification strategies can be used to complete the same task (i.e. verifying the truth values of sentences against a visual scene). The authors expected that paired stimuli would trigger a different verification procedure than scattered stimuli. To test whether this is indeed the case, Steinert-Threlkeld et al. included a digit recall task to manipulate working memory load, expecting that if subject consistently used only one strategy, the effect of this additional resource restriction would be consistent across all conditions.

The authors found that in the case of *more than half*, both the accuracy and reaction times of the sentence verification task were affected by the type of stimulus. Moreover, the interaction of stimulus type and working memory load had significant effects on accuracy and RTs in the digit recall task: the difference in RTs and accuracy between low and high working memory load conditions was greater for scenes with scattered objects than for paired objects. Interestingly, the authors did not find the same interaction effects for *most*. Accuracy and RTs in the digit recall task showed effects of working memory condition – however, there were no significant interaction effects of stimulus type and WM. Steinert-Threlkeld et al. concluded that working memory demands for *most* are not affected by the presentation of stimulus in the same way as *more than half*, and that the two quantifiers have distinct verification profiles.

Crucially, even though this conclusion seems to mirror Hackl's results, his interpretation of the difference between *most* and *more than half* does not fit Steinert-Threlkeld et al.'s data. If *most* indeed favors a verification procedure based on lead-counting, there would be a significant difference in working memory demand between paired and random stimuli, which Steinert-Threlkeld et al. did not observe.

3.3 More about *most*: Approximate Number System

Pietroski et al. (2009) compared two distinct notational variants of specifying the truth condition for *most*, in (11) below, to see if there is a psychological significance, i.e. whether one of the algorithms is used as the default verification strategy. The procedure in (11a) requires comparing two cardinalities: the number of blue dots and the number of non-blue dots; the alternative representation in (11b) calls for verifying whether some but not all of the blue dots can be paired off with non-blue dots.

(11) a. $\text{GreaterThan}[\#\{x : \text{Dot}(x)\&\text{Blue}(x)\}, \#\{x : \text{Dot}(x)\&\neg\text{Blue}(x)\}]$

b. $\text{OneToOnePlus}[\{x : \text{Dot}(x)\&\text{Blue}(x)\}, \{x : \text{Dot}(x)\&\neg\text{Blue}(x)\}]$

To test whether the relationship between the truth conditions in (11) and the default verification profile triggered by the quantifier *most* is constrained, Pietroski et al. conducted an experiment, in which participants had to verify visual scenes against statements containing *most*. They presented participants with scenes that favored using OneToOnePlus (pictures in which the dots were paired) strategy and scenes which make using this strategy difficult (pictures in which the dots were scattered randomly). They predicted that if this variation does not affect participants' accuracy, then *most* is probably not understood in terms of correspondence.

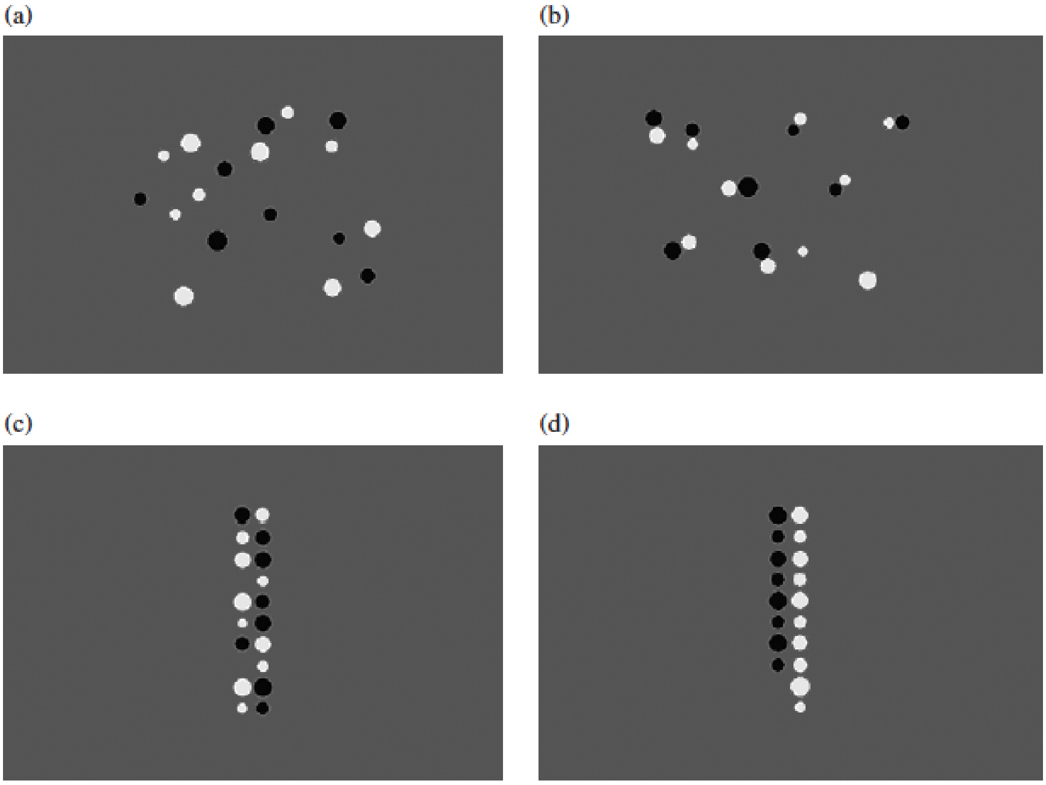


FIGURE 3.3: The conditions in Pietroski et al.'s experiment: (a) Scattered Random, (b) Scattered Pairs, (c) Column Pairs Mixed, (d) Column Pairs Sorted

During each trial, participants saw a screen with dots of two colors (yellow and blue) for 200ms and were asked to answer the question *Are most of the dots blue?*. There were four conditions: Scattered Random, Scattered Pairs, Column Pairs Mixed, and Column Pairs Sorted (see Figure 3.3). The authors expected that if *most* is indeed understood in terms of a one-to-one correspondence, subjects would perform better on those trials where the dots were paired. However, it turned out only performance on the Column Pairs Sorted condition was significantly higher than on other trials; there were no significant differences between Scattered Random, Scattered Pairs, and Column Pairs Mixed. The authors conclude that these results suggest that the meaning of *most* is not specified in terms of one-to-one correspondence.

Similarly, Lidz et al. (2011) compared the two specifications of *most* below. The specification in (12a) corresponds to what Lidz et al. call a *selection* verification procedure that enumerates (or estimates the number of) the blue dots, then the nonblue ones, and compares the two numerosities. The specification (12b) triggers a verification procedure that estimates the overall number of dots, then the number of blue dots, subtracts the latter from the former, and verifies that the result is smaller than the number of blue dots.

(12) a. $> (|\text{DOT} \cap \text{BLUE}|, |\text{DOT} - \text{BLUE}|)$

b. $> (|\text{DOT} \cap \text{BLUE}|, |\text{DOT}| - |\text{DOT} \cap \text{BLUE}|)$

They hypothesized that if subjects used the selection procedure, their performance should be higher on trials where there are two colors on the screen. The motivation behind this is that the selection procedure yields more accurate results, and would therefore be the optimal procedure to use when it's available. However, on screens with multiple colors, identifying all non-blue dots, assessing their cardinality and subtracting this number would have been too difficult to do in 150ms; therefore authors expected that participants would switch to a less accurate subtraction procedure. However, they found that there was no significant difference in accuracy between trials with just two colors and trials with multiple colors. The authors concluded that subtraction procedure was used throughout the experiment. Moreover, they claimed that this result supports 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).

More importantly, Pietroski et al. and Lidz et al. found strong evidence that participants used a cognitive resource called *the Approximate Number System* to solve the tasks. The Approximate Number System (ANS) is an evolutionary ancient system of representing numerosity that humans share with other animals (Dehaene, 1997). It is often contrasted with arbitrary systems of exact number representation – or, simply put, the way counting is taught in schools: by representing number exactly. This method involves arbitrary symbolic, discrete representations of number – number words like *seven*, *decimal*, *one tenth* – to perform very precise numeric operations (Spaepen et al., 2011).

ANS does not need to be learned explicitly: it is present in infants and nonverbal adults (Gordon, 2004; Izard et al., 2009), and it allows us to make numerical discriminations and perform certain numeric operations, such as estimating the cardinality of a set. As it is apparent from its name, the ANS does not generate exact representations of numerosity – instead, number is represented as a continuous Gaussian activation of several numerical values on a mental number line (Dehaene, 1997; Halberda et al., 2012; Odic, 2014; Piazza et al., 2004). This means that when we are looking at a scene that contains a certain number of dots – say, seven – we will not be able to verify that there are exactly seven dots just by using ANS. However, we will be able to estimate that this number is somewhere around seven – maybe six or eight. Generally, the more representations of these activations overlap, the more difficult it is to discriminate between them (Feigenson et al., 2004; Halberda and Feigenson,

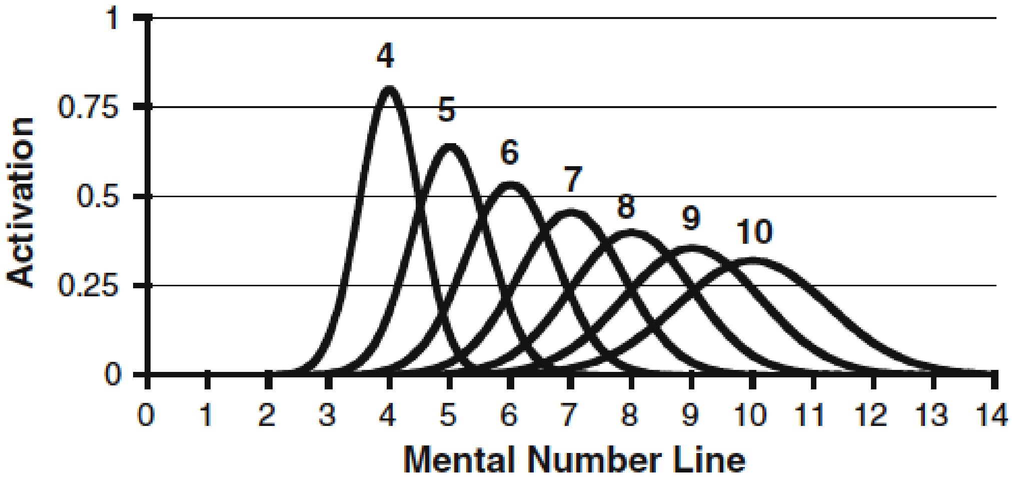


FIGURE 3.4: Representation of numerosities on the mental number line

2008; Odic, 2014). ANS is characterized by its compliance with Weber's law, which states that discriminability is determined by the ratio of the two values being compared. For example, if we are comparing two sets of dots and have to pick the largest one, it is as easy to do when there are 6 blue and 12 yellow dots as when there are 60 blue and 120 yellow dots. When the difference between two values remains the same, but the numerosities increase (e.g. from 6 and 12 to 60 and 66), it becomes more difficult to compare them (the size effect). When of the two values being compared one remains the same but the other one increases (from 6 and 12 to 6 and 20), the comparison becomes easier (the distance effect).

For all their differences, ANS and the system of precise number representation, which requires rule-learning, are not completely independent from each other. Children in numerate cultures learn how to map representations of the ANS onto discrete number words by the time they are 5 years old (Le Corre and Carey, 2007), and there is evidence that later in life, whenever an adult sees a precise representation of numerosity – such as an Arabic numeral – or performs mental calculations, the ANS gets activated (Dehaene, 1997).

Even in tasks that require precise judgments, some affects of the ANS are present. For instance, Zajenkowski et al. (2014) compared the difficulty of proportional quantifier processing under different semantic conditions. In particular, they were interested how difficult it would be for subjects to verify proportional quantifiers against a scene depending on the number of objects present on the screen. They found that the numerical distance between two cardinalities that must be compared was significant for accuracy and reactions times: the bigger was this distance, the better the performance. Moreover, this result was significant no matter the total number of objects in a scene.

As Pietroski et al. (2009) and Lidz et al. (2011) point out, the properties of the ANS we have discussed make it incompatible with using a OneToOnePlus strategy:

ANS does not generate representations of unit or exact differentiations between cardinalities – “the ANS will not deliver a representation of something *as* exactly one” (Pietroski et al., 2009, p. 567). Since the system does not represent unity or minimal differences between discrete cardinalities (Leslie et al., 2008), they conclude speakers cannot rely on ANS to implement a OneToOnePlus algorithm.

3.4 How it all comes together

The papers we have reviewed in this chapter all take slightly different approaches to understanding the relationship between specification procedures and verification profiles of natural quantifiers in natural language. Hackl (2009) was interested in comparing two quantifiers whose truth conditions are identical in Generalized Quantifier Theory, but, he suspected, were verified differently by speakers. Zajenkowski et al. (2011), Zajenkowski et al. (2014) and Szymanik and Zajenkowski (2010) focused on the role of working memory load in verification of different types of quantifiers. Lidz et al. (2011) and Pietroski et al. (2009) compared multiple ways of specifying the truth conditions for *most* to see if there is psychological significance.

Despite the differences in experimental paradigms, these lines of research are mutually informative and share the same questions at their core. As we are interested in comparing quantifiers *most* and *more than half*, we will now look in more detail at how Hackl’s results could be informed by the studies we related in this chapter.

First of all, we will note that, based on his experimental findings, Hackl argues that *most* and *more than half* trigger distinct verification strategies. However, from an earlier discussion, we have seen that we cannot equate distinct renditions of truth conditions with verification strategies. Although Hackl observed certain differences in the verification of *most* and *more than half*, it’s not clear whether they are constrained by the differences in the specifications of truth conditions between the two quantifiers, or are an artifact of the experimental setup.

For instance, Hackl reports significant screen-by-screen differences in reaction times between *most* and *more than half*, but interestingly, they do not sum up to significant overall difference. It is not clear why this happens, especially since there were no explicit hypotheses about what RTs and accuracy rates are expected from each screen.

As we have seen, one of the things that impacts the processing of proportional quantifiers like *most* and *more than half* is working memory – and the task in Hackl’s experiment requires significant working memory load: participants had to remember the strings of dots they had previously seen to solve the task. However, working memory was not controlled for, and we don’t know exactly how it impacts performance.

Finally, the results of Lidz et al. and Pietroski et al., as well as Steinert-Threlkeld et al.’s, argue against a pairing strategy for *most*. Lidz et al. and Pietroski et al. in particular argue that *most* requires a verification strategy that is based on ANS and is therefore incompatible with OneToOnePlus. However, it is also important to point out that the experimental setting used by Lidz et al., Pietroski et al. and Steinert-Threlkeld et al. is very different from Hackl's. So while it is convincing that the pairing strategy was not preferred by speakers in their experiments, it could be triggered by a different experimental paradigm, such as Self-Paced Counting.

3.5 Conclusion

In this chapter, we have overviewed a series of experimental studies that explored the processes underlying quantifier verification. We have presented the experimental results from Hackl (2009), that show that *most* and *more than half* are indeed processed differently, despite truth-conditional equivalence.

We have summarized the results of studies by Zajenkowski et al. (2011), Szymanik and Zajenkowski (2010) and Steinert-Threlkeld et al. (2015) that tested the involvement of working memory in verification tasks and provided evidence that the classification of quantifiers in automata theory is cognitively plausible, as proportional quantifiers like *most* and *more than half* require higher working memory involvement than first-order quantifiers. Finally, we have given an overview of the verification profile of *most* from Lidz et al. (2011) and Pietroski et al. (2009).

However, as we pointed out, verification of *most* and *more than half* – as well as the question about default verification profiles – are not completely settled issues; a lot remains to be investigated. We have pointed out that several aspects of Hackl's study make the interpretation of their experimental results problematic. Moreover, given the discussion about the relationship between default renditions of truth conditions and default verification strategies, can we extrapolate the differences discovered by Hackl to verification in general? Or are they a result of a particular setup?

In the following chapter we will present the results of an experimental study which aimed to answer some of these questions.

4 Experiments

In this section, we present the results of two experimental studies that explore differences in verification procedures triggered by *most* and *more than half*, and try to answer the question about whether these difference arise from distinct default verification profiles. We will show that, contrary to the results of Pietroski et al. (2009), Lidz et al. (2011) and Hunter et al. (2016), our data do not support the claim that there are default verification strategies for these quantifiers. Although we will observe some distinctive features of *most* and *more than half*, we will provide evidence that subjects in our studies varied in their choices of verification procedures, suggesting that there is a collection of strategies associated with each quantifier.

4.1 Experiment 1

4.1.1 Motivation

As we have discussed in the previous sections, the results of Hackl's study leave open several questions about verificational differences between *most* and *more than half*.

First of all, there is the question of working memory load effect in the processing and verification of quantifiers. We have presented experimental evidence from several studies that explored the extent of WM load in proportional quantifier processing, which all point to the fact that processing *most* and *more than half* requires additional executive resources. We have also argued that the design of a self-paced counting also places demands on WM load, which might affect the interpretation of observed differences between *most* and *more than half*.

We have seen that the mode of stimuli presentation can make a difference on how subjects verify quantified statements. For instance, in Pietroski et al. (2009), the Column Pairs Sorted condition (which is also visually similar to Hackl's rows of dots) elicited a different verification strategy compared to other conditions in the experiment. While the differences that Hackl reports might be due to the fact that *most* and *more than half* have distinct verification profiles, there is also a possibility they are constrained by the way in which the stimuli were presented.

Finally, while Hackl observed several differences in the verification of *most* and *more than half*, they are difficult to pin point: we know that the two quantifiers are processed differently, we just don't know *how* they differ. We will attempt to answer this question by making explicit predictions about the lead counting strategy that Hackl postulates for *most*.

4.1.2 Predictions

As we have seen in Chapter 3, studies such as Szymanik and Zajenkowski (2010) and Steinert-Threlkeld et al. (2015) shed light on the involvement of working memory in the processing of proportional quantifiers: proportional quantifiers tend to require more working memory capacity compared with other types of quantifiers. We have also noted that the experimental design in Hackl (2009) requires additional memory load, as subjects have to keep track of the images they have seen. Putting these two factors together makes the interpretation of reported results in Hackl (2009) somewhat complicated: while the differences in RTs between *more than half* and *most* might be due to different demands these quantifiers place on working memory capacity, the extent of this effect cannot be readily determined. One problem is that, to our knowledge, working memory demands have not been assessed for each of these quantifiers separately. Although it's true that *most* and *more than half* place more demands on working memories than less complex quantifiers like *only three* or *some*, it is not clear whether they place *equal* demands.

The other problem is that the experimental design in Hackl (2009) prompted subjects to store and process large chunks of information in their working memory. For example, if subjects were to verify whether *More than half of the dots are blue* was true, they would have to remember that there were 2 blue dots and 1 yellow dot in the first screen, 1 blue dot and 2 yellow dots in the second screen, etc., and later retrieve from memory how many dots in each color they had seen. However, individuals vary in their working memory capacity, and these differences in turn relate to differences in linguistic processing (Bornkessel et al., 2004; Daneman and Carpenter, 1983).

These two aspects combined might affect the results of the experiment, as it is not clear to what extent the differences between the quantifiers are due to their properties (i.e., different demands they place on working memory), or other factors. Moreover, the experimental setup requires subjects to verify a big number of items in a row (60 items in Hackl's study), which could lead them to develop cognitive strategies that are more efficient in the context of the given task, such as trading accuracy for time. This, again, is another reason why it's not completely clear whether verification strategies are triggered by particular quantifiers or the task itself. In the current study, we will attempt to reach some clarity by controlling for individual working memory capacity, expecting a correlation with reaction times and accuracy.

Hackl (2009) argues that verifying *most* requires participant to keep track only of the color that is leading at any given moment. Verifying *more than half*, on the other hand, does not rely on lead-counting, meaning it would ostensibly require 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 higher working memory capacity will result in shorter reaction times for *more than half*. At the same time, we expect that subjects with higher memory scores will make fewer mistakes when verifying *most*. As in Hackl's experiment *more than half* had both higher accuracy and mean reaction time compared to *most*, we make the following prediction:

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).

So far, we have followed Hackl's argument that the rendition of truth conditions in (10b) would be a more accurate way to capture the meaning of *more than half*, as the algorithm specifically mentions dividing the bigger set in half; conversely, (9a) is more desirable for *most*. We have also seen some evidence in favor of distinct verification profiles that these quantifiers trigger; but is the choice of a verification strategy constrained by the default specification of truth condition for each quantifier?

Suppose a speaker is asked to verify whether the sentence *More than half of the dots are blue* is true, and they are presented with a visual scene in which dots are scattered across a picture – or placed neatly in rows, for that matter. As we have discussed earlier, we cannot guarantee, even if we know that all competent speakers interpret *most* as in (9a) what strategy our speaker would choose[1]: the questions about what abstract rules are captured by a speaker's linguistic competence and what functions implement this competence in their brain are at different cognitive levels. For whatever reason, the speaker might decide to settle the issue by rolling a dice, and there is no way we could predict their choice of dice-rolling as a strategy based on the default meaning of *most*.

Suppose, however, that as per suggestion of Hackl (2009) the speaker is biased to use an algorithm that is associated with the specification in (10b) – that is, that the choice of a verification strategy is constrained by the form of the expression in (10b). Then we would expect our speaker to solve the task by following these steps for *more than half*:

1. Calculate or approximate the total number of dots.

[1] Although there are some cognitive factors that could make the choice of one strategy more likely than others.

2. Divide that number in half.

3. Calculate or approximate the number of blue dots.

4. Compare the cardinalities of (2) and (3).

For *most*, the verification procedure might go something like this[2]:

1. Calculate or approximate the number of blue dots in the current screen.

2. Calculate or approximate the number of nonblue dots in the current screen.

3. Verify that the number in (1) is leading. If (2) is leading instead, switch the leading color.[3]

Note, however, that it would make a difference, when verifying *more than half*, whether the total number of dots is odd or even. If this number is even, the second verification step should be relatively easy: for instance, if there are 12 dots on the screen, it becomes immediately obvious that 12 divided by 2 is 6, and so there should be 7 blue dots for the sentence *More than half of the dots are blue* to be true. If there are 13 dots on the screen, however, performing the verification steps we sketched above becomes more difficult. Crucially, the verification procedure for *most* should not be affected by this variation: the oddness or evenness of the total number of dots is irrelevant for enumerating nonblue and blue dots. We can make the following prediction:

Prediction 2. Solving a verification task for the quantifier *more than half* would result in a longer reaction time and lower accuracy when the total number of dots is odd.

Finally, Hackl (2009) observed that there was a screen-by-screen difference in reaction times between *most* and *more than half*. He argued that this is indicative of *most* using a lead-counting strategy (i.e. a similar strategy to OneToOnePlus) which requires speakers to consistently check what the leading color is. Hackl's rationale behind this explanation is that the lead-counting strategy is particularly well-suited for the Self-Paced Counting paradigm. Still, we have to verify that using distinct strategies leads to different screen-by-screen reaction times.

Consider the following scenario. As before, subjects are asked to verify sentences like *Most of the dots are blue* and *More than half of the dots are blue*. On screen 2, two dots are blue and one is yellow. On screen 3, two dots are yellow and one is blue. Then on screen 4, again, two dots are yellow and one is blue. If subjects are using a lead-counting strategy, it would be easy for them to react on screen 4 (let's call it the *target* screen), as it is very clear that the target color is leading. However, screen 4

[2] As we have seen, this is not the only possible verification procedure for *most*: in fact, Lidz et al. (2011) argue against a selection procedure, which requires speakers to select all non-blue dots. However, for the sake of the current argument, it does not matter which procedure for *most* we compare *more than half with*, as none of the procedures for *most* require dividing the total number of dots in half.

[3] The change in the leader would probably lead to an increase in reaction time.

does not facilitate the verification strategy triggered by *more than half* – if reasoners rely on a more precise strategy and keeping track of how many dots in each color they have seen, paying attention to which color is leading in which screen seems excessive. Then, we would expect that the RTs for *most* would be lower than for *more than half* on screen 4 (the target screen).

Prediction 3. Reaction times on the target screen for *most* will be significantly lower than reaction times on the leading screen for *more than half*.

4.1.3 Participants

Thirty five (8 female, 24 male, 1 genderfluid) subjects were initially recruited for the study via Prolific.ac, all native speakers of English. They viewed the experiment in their web browsers, and the average completion time was 14 minutes. Subjects received £2.50 as compensation. Subsequently, we removed the subjects who spent less than 10 seconds reading the instructions, resulting in a pool of 33 subjects. Participants who had failed to answer correctly on at least 70% of catch trials (trials with unambiguous answers) had been removed from the study at an earlier stage and did not receive compensation.

4.1.4 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. 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).

The first task consisted of 14 sequences of digits, ranging from 3 to 9 digits. Each number of digits appeared twice: there were two 3-digit sequences, two 4-digit sequences, etc. The sequences were created using a random number generator, but were the same for all participants. Subjects also saw two practice trials, one consisting of 2 digits and the other of 3 digits.

The second section consisted of 24 target items: 12 sentences with the quantifier *most* and 12 with the quantifier *more than half*. In each group, 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 visual stimulus that matched that description) and 6 were false. The visual stimuli consisted of pictures of dots scattered across the screen. All dots had a radius of 20 pixels and were situated within a grid with 10 rows and 10 columns. Grid spacing was set to 50 pixels. Dots were scattered in chunks of two or three dots with their location generated randomly. The number of dots that appeared on the screen varied between 10, 11, and 12 dots; there were 8 target items in each category (2 false *most*, 2 true *most*, 2 false *more than half* and 2 true *more than half*). On target trials, it was never clear whether the statement was true or false until screen 5, the last screen.

The color of the dots varied across trials; altogether, four different colors were used (yellow, blue, red, and green). Each trial only featured dots of two different colors: for example, if the sentence the subject had to verify was *Most of the dots are yellow*, the image would feature dots in yellow and one other color (for instance, blue). The difference between the true and false conditions was always kept to one or two dots (depending on whether the total number of dots was odd or even). In the true condition, if a trial had 12 dots in total, there would be 7 dots in the target color and 5 in the other color, and vice versa in the false condition. If a trial had 11 dots, there would be 6 dots in the target color and 5 dots in the other color in the true condition, and vice versa in the false condition. Screen 4 was consistently the "target" screen on all trials: on this screen, the target color always had an advantage. For instance, if blue was the target color, and there were 2 blue dots and 1 yellow dot in the second screen, followed by 1 blue and 2 yellow dots in the third screen, then the fourth target screen would contain 2 blue and 1 yellow dots.

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. There were 18 true and 18 false fillers, and, as with the target items, the total number of dots varied between 10, 11 and 12 dots, with 12 dots in each category. Of the 36 fillers, 13 were "catch" trials – unambiguous sentences, for which it was easy to judge whether they were true or false. Some examples of the catch items are in (13). Subjects also received three practice items similar to the fillers to familiarize themselves with the task.

(13) a. More than three dots are yellow.
b. Only six dots are red.
c. Only four dots are blue.

All stimuli were created using the JsPsych library for JavaScript (de Leeuw, 2015)[4], and the circles were drawn using the Snap.svg library.

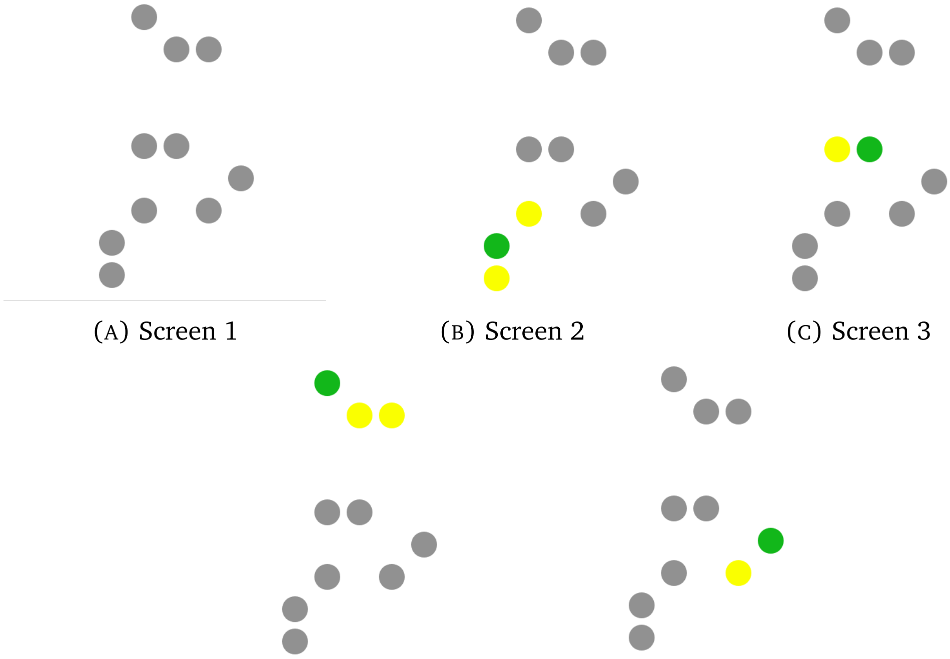
4.1.5 Procedure

Digit span task

Subjects views stimuli and solved the tasks in their web browsers, answering with keyboard keys or entering responses into text fields where necessary. In the digit span task, sequences of digits appeared on their screens, with each digit appearing on a separate screen for 1000 milliseconds. The break between digits was set to 200 milliseconds. Every sequence was preceded by the "+" sign presented for 250 milliseconds to draw participants' attention to the upcoming sequence. After seeing the sequences, subjects were asked to enter it into a text field in reverse order. If they

[4] See de Leeuw and Motz (2016) for discussion on the reliability of response time measurements collected using JavaScript relative to standard laboratory software.

made three mistakes in a row, the task stopped and they proceeded to the quantity judgment task.



(A) Screen 1

(B) Screen 2

(C) Screen 3

(D) Screen 4

(E) Screen 5

FIGURE 4.1: Sequence of events in a trial

Quantity 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 wasn't limited, and the subjects had to press the space bar to proceed to the image. In the first screen (see Figure 4.1), only the outlines of the dots were visible. The subjects had to press the space bar to move through the screens.

When subjects uncovered a new screen, the dots they had previously seen were covered again. They also weren't 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 statement was true or the "N" key (for "no") if they thought the statement was false. On some filler trials, they could indeed answer before reaching the final screen, as the answer became clear after the second or third screen.

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.

4.1.6 Results

As in Hackl (2009), we only analyzed data from subjects who had at least an 80% correct response rate; however, no subjects were excluded according to this criterion. When analyzing reaction times, we only used data from correctly answered trials. We also followed Hackl in excluding the fifth screen – the screen that varied between the true and false conditions – from our analysis.

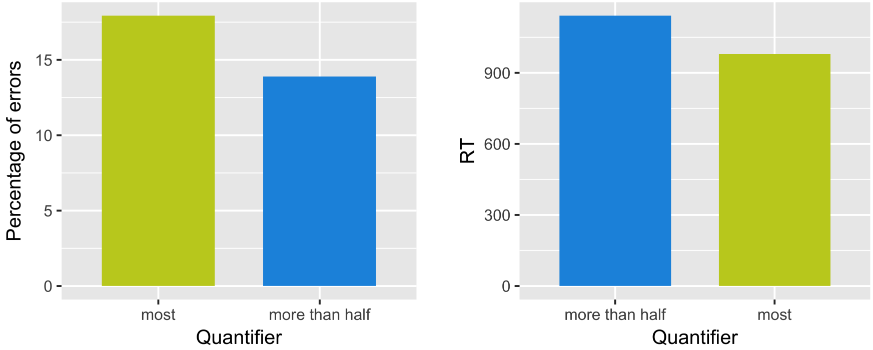


FIGURE 4.2: Overall RTs and percentage of errors per quantifier

Following Hackl (2009), we first analyzed the data in terms of overall accuracy and reaction times. As can be seen from Figure 4.2, 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) – however, this difference was not significant ($t(781.66) = 1.5548, p = 0.12$). The difference in overall reaction times, on the other hand, was significant ($t(2568.6) = -3.7445, p < 0.01$): as shown in Figure 4.2, subjects took longer to verify *more than half* than *most*.

A 2 (Quantifier) × 4 (Screen) repeated-measures ANOVA yielded a significant main effect of Quantifier ($F(1, 32) = 16.14, p < 0.01$). No other significant effects were found, suggesting that on all screens, quantifier *more than half* took subjects longer to process than *most* (see Figure 4.3).

However, parametric tests like the t-test and ANOVA cannot be reliably applied to our data: we found big differences in mean RTs between subjects and high standard deviations throughout the experiment (see the Appendix for details).[5] For our current purposes, we repeated the analysis with non-parametric tests to take the high variance of our data into account; however, we will explore possible reasons behind it in section 4.1.7.

Wilcoxon rank-sum test confirmed that reaction times were significantly affected by quantifier, $W = 957740, p < 0.001$. We also confirmed that there were no significant overall differences in accuracy, $W = 81576, p = 0.1204$.

We conducted a Kruskal-Wallis test to explore whether there were significant differences in reaction times per screen. RTs were significantly affected by screen, $H(3) = 140.39, p < 0.001$. Focused comparisons of the mean ranks between screens

[5] This also raises the question about whether the data in Hackl 2009 followed the normal distribution, and if so, why the subjects behaved differently in our experiment.

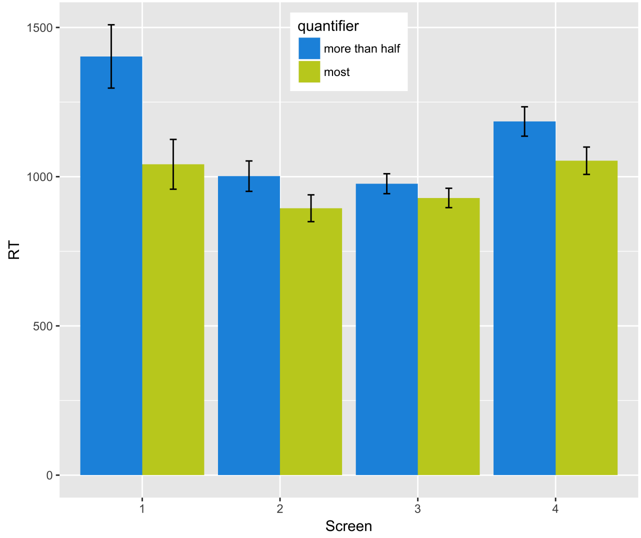


FIGURE 4.3: Reaction times per screen

showed that RTs were not significantly different when screens 2 and 3 were compared (*difference* $= 103$). However, RTs in screen 1 were significantly higher than in screen 2 (*difference*$= 224$), screen 3 (*difference* $= 327$) and screen 4 (*difference* $= 484$). We also found significant differences in RTs between screen 4 and screen 2 (*difference* $= 260$) and between screen 4 and screen 3 (*difference* $= 156$).

We further investigated whether there were significant differences in reaction times between *most* and *more than half* on each screen. 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 results were found (see table 4.1 for details).

TABLE 4.1: Differences in mean RTs between *most* and *more than half* per screen.

screens	result
1-2	$W = 551, p = 0.9391$
1-3	$W = 602, p = 0.4673$
1-4	$W = 521, p = 0.7695$
2-3	$W = 618, p = 0.3515$
2-4	$W = 492, p = 0.5074$
3-4	$W = 414, p = 0.09565$

To investigate the relationship between working memory capacity and accuracy and RT effects, we assigned a memory score to every subject. The memory score was equal to the number of digits a subject could reliably remember – i.e., the maximal number of digits a subject remembered on each trial. If, for instance, she remembered both sequences of 7 digits, and only one sequence of 8 digits, she would receive score 7. In some cases, subjects were able to remember longer sequences of digits than their reliable score: for example, they would reliably remember 5-digit sequences, but could also reproduce one of the two 7-digit strings, one of the two

8-digit strings, and one of the two 9-digit strings (i.e. one or more sequences in each category). In these cases, we boosted their memory score by 1 point.

To calculate accuracy effect, we subtracted the number of errors subjects made when verifying *most* from the number of errors with *more than half*. Likewise, to calculate RT effect, we subtracted each subject's mean RT for *most* from their mean RT for *more than half*. We found some negative correlation between memory score and accuracy effect (Pearson's $r = -0.22$), suggesting 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.

Further, we found that the total number of dots was a significant factor for overall accuracy ($H(2) = 8.54, p = 0.01398$) as well as for reaction time ($H(2) = 12.104, p = 0.0023$).

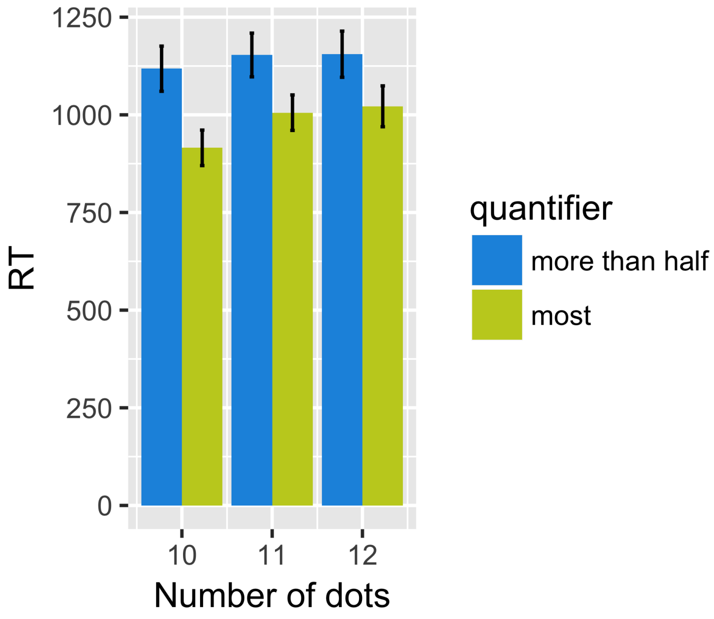
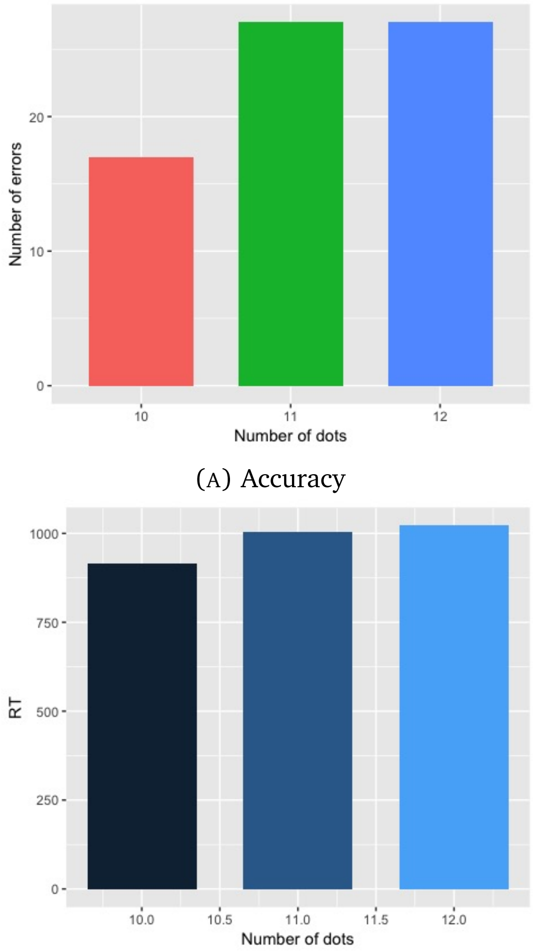


FIGURE 4.4: RTs with respect to the number of dots

Subsequently, we analyzed whether the odd condition was significantly different from the even condition – i.e., whether oddness or evenness of the overall dot number in a trial affected subjects' accuracy and RTs. As the distribution of trials between these two conditions was uneven in our dataset, we first analyzed a subset of our data that excluded the trials in the 10-dot condition. We found no effects on accuracy or RTs neither for *most*, nor for *more than half*.

We further analyzed a subset of our dataset that only included the 10-dot and the 11-dot arrays. For *more than half*, we found an effect of dot number on accuracy ($W = 7854, p = 0.0186$) such that subjects made fewer errors in the 10-dot condition; no effect on reaction time was found. We found no significant effects on accuracy for *most* ($W = 8052, p = 0.0995$). Reaction time was significantly affected by the number



(A) Accuracy

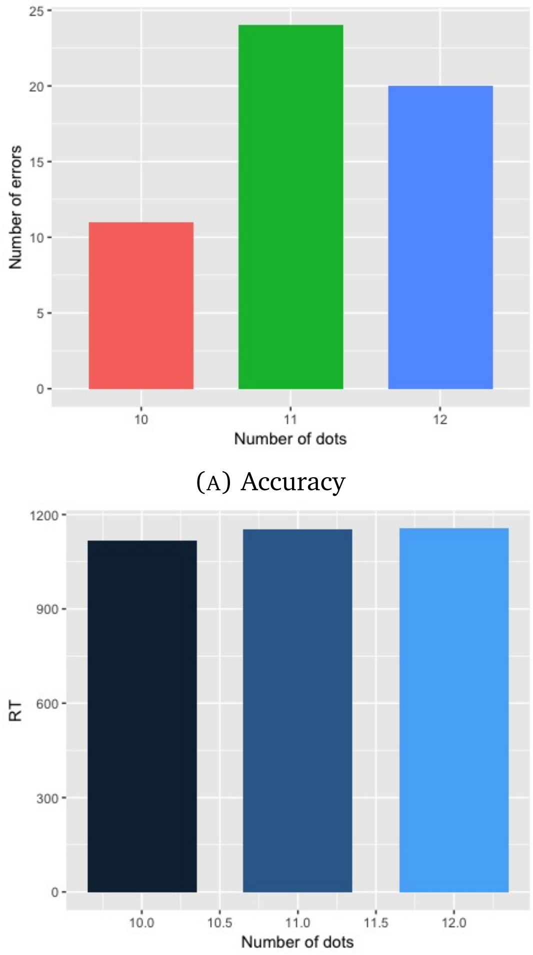
(B) RTs

FIGURE 4.5: Accuracy and reactions times with respect to dot number for MOST

of dots ($W = 82800, p = 0.0016$), as participants verified *most* faster in the screen where there were 10 dots. However, there was a confound with the difficulty due to a higher number of dots in the odd condition; see Discussion for more.

4.1.7 Identifying strategies

We have noted that there is a wide spread of means across of our data, as well as high standard deviations throughout (see Appendix for mean RTs per participant). At the same time, the selection procedures we described in this section precluded any dishonest participants who solved the task at random or didn't follow the instructions. This leads us to suspect that the observed differences are actually an indication that participants did not adhere to one particular strategy while solving the task; instead, it looks like different subjects approached the problem in different ways. The question remains whether these differences boil down to distinct verification strategies subjects used (that is, subjects consistently used one of several possible strategies) or they mixed and matched strategies as they went through the task.



(A) Accuracy

(B) RTs

FIGURE 4.6: Accuracy and reactions times with respect to dot number for MORE THAN HALF

We have noted that the reaction times in the first screen were considerably higher than in other screens (or than RTs in the first screen in Hackl 2009); we suggested 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 4.7, not all participants behaved in this way: while some spent over 3000 milliseconds looking at the screen, others were fast and 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 (> 2000 ms) who we suspect used a precise strategy throughout the whole experiment, the "mixed" group

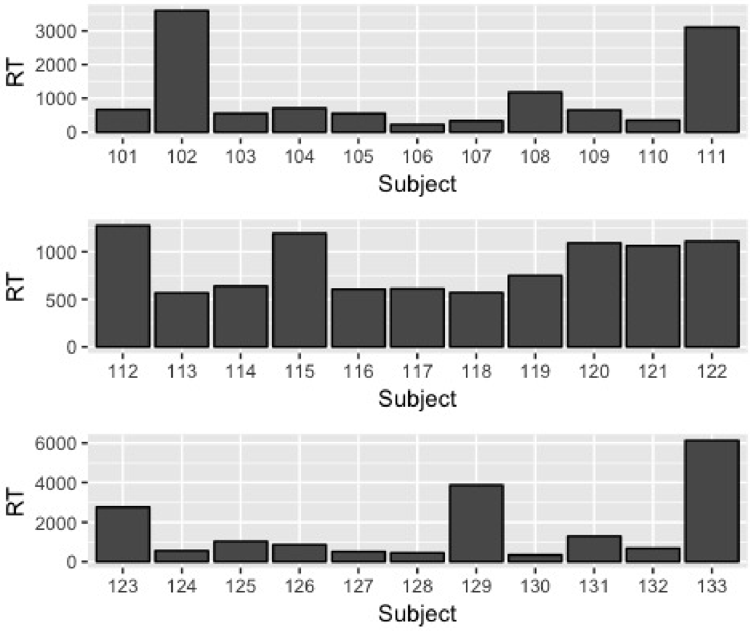


FIGURE 4.7: Mean RTs by participant in the first screen

($1000 - 2000$ 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 ($W = 20757, p = 0.4057$). Further, no significant effect of dot number was found either on RT ($H(2) = 0.9656, p = 0.6171$) or on accuracy ($H(2) = 0.119, p = 0.9422$). Note, however, that due to the low number of participants, these results are not conclusive, but rather exploratory and indicative of a potential effect.

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 ($W = 81854, p = 0.012$). We also found a significant effect of dot number on overall reaction times ($H(2) = 8.9493, p = 0.0113$) and accuracy ($H(2) = 7.6839, p = 0.0214$).

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 ($W = 298730, p = 0.00042$). We also found that dot number was not significant neither for accuracy ($H(2) = 4.2809, p = 0.1176$), nor for reaction time ($H(2) = 4.2679, p = 0.1184$).

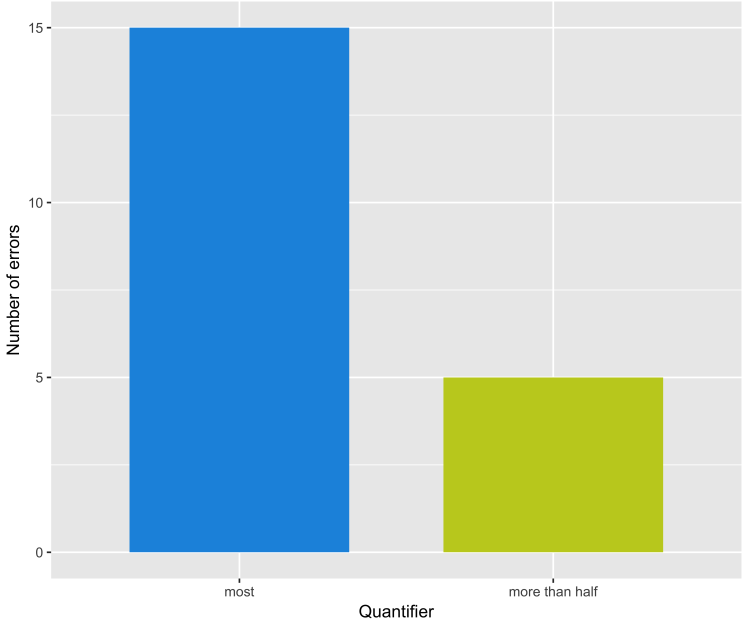


FIGURE 4.8: Number of errors per quantifier in the "counters" group

4.1.8 Discussion

The results of the present experiment differ in a number of respects from Martin Hackl's study. First of all, recall that in Hackl's 2009 experiment no significant differences in RTs and accuracy between *most* and *more than half* were found, which, the author 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 overall difference in RTs between the two quantifiers was significant. While this fact alone is not enough to claim that *most* and *more than half* have distinct default verification profiles, it is a starting point to look into some differences between the verification strategies subjects used in our experiment.

As we have have mentioned previously, there was a consistent increase in reaction times throughout a trial in Hackl's experiment: the further subjects went through screens, the longer it took them to press the relevant key (the space bat or one of the response keys). In the present study, the picture is quite different: mean reaction time for all participants together was highest in the first screen, then dropped in screen 2 and remained stable on screen 3, increasing again in screen 4. Although this pattern is quite different, we will argue that it can be explained by underlying differences in verification profiles of *most* and *more than half* that became more relevant in our design.

The first screen is particularly interesting not only as the screen on which subjects spent the longest time; it is also characterized by the biggest difference in RTs between the two quantifiers. One of the possible reasons for this is the presentation of visual stimuli in our experiment. As we have noted, contrary to Hackl's setup, the dots in our study were not located in rows next to each other, but spread around the

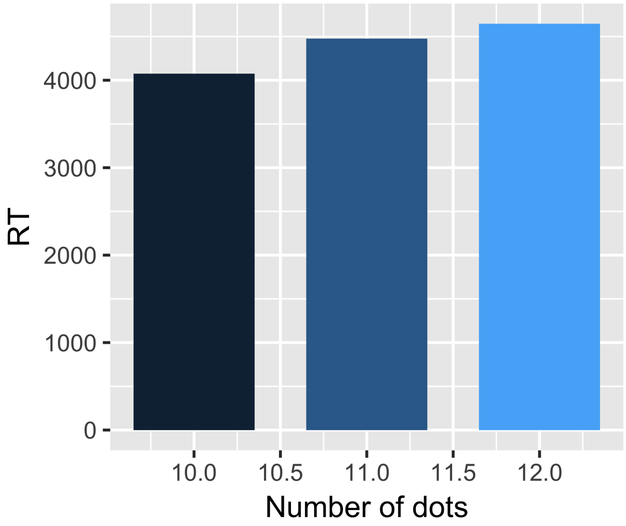


FIGURE 4.9: Reaction times per number of dots in the "mixed" group

screen in chunks of 2-3 dots. This design makes it easier for subjects to count or estimate the total number of dots: to do so in Hackl's experiment, subjects would have to visually discern every dot in a row, which would be time-consuming and error-prone, while estimating that two chunks of 3 dots and two chunks of 2 dots would yield 10 dots is relatively easy. This extra step, however, is not effortless, and the fact that subjects chose to take it when verifying *more than half* suggests that this quantifier employs a more precise verification strategy compared to *most*.[6]

Hackl has also argued that the difference in screen-by-screen RTs for *most* and *more than half* is related to the way in which stimuli is presented in the experiment, with the information being uncovered gradually. Hackl argues that this mode of presentation is particularly well suited for a lead counting strategy, which he believes is triggered by *most*. However, as we have argued above, if indeed 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 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. We will also note that the leading strategy described by Hackl is reminiscent of the OneToOnePlus strategy that Pietroski and colleagues discussed among other possible default verification for *most*. Pietroski et al. (2009) concluded that it is unlikely that reasoners rely on the OneToOnePlus strategy, as it would be incompatible with the ANS effects they have found.

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

[6] Alternatively, the spike in reaction times in the first screen could have been a spill over, as subjects were processing the statement they had seen in the previous screen. This, however, is unlikely: no such effect was found in Experiment 2, even though the designs of the experiments were almost identical.

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.

Although the correlation effect is not big enough to make any strong conclusion about the role of working memory in verification profiles, it is important to note that the memory scores our subjects received were not consistently reflective of their working memory capacity. As we did not have control over how honestly participants followed our instructions, they were free to employ whatever strategy they found the most efficient when solving the task, including remembering the sequence in the order that it appeared, putting it into a text field and only then changing the order, or not even attempting to memorize the sequence and instead writing down the numbers as they appeared on the screen. To estimate whether any such practices could have impacted the relationship between memory score and accuracy effect, we punished the subjects whom we suspected of solving the digit span task dishonestly by assigning them a score of 5, the mean memory score in our pool. The subjects we picked solved the entire task without making any errors and initially received the highest score of 9. This resulted in an even higher negative correlation between memory and accuracy effect (Pearson's $r = -0.36$), on the one hand, and between memory and RT effect, on the other (Pearson's $r = -0.39$).

More investigation into the relationship between working memory and verification is needed before any conclusions can be made, but if indeed, as our preliminary results suggest, higher WM capacity allows for a more precise verification for *most*, this would add further evidence that the two quantifiers have different default verification profiles. If subjects pursue a more thorough strategy for *most* only when they have sufficient executive resources, but consistently rely on it when verifying *more than half*, we have reasons to believe that precision is not part of the default verification strategy for *most*.

Significant differences in accuracy between the odd and even condition suggest that it's indeed easier for subjects to verify statements involving *most* and *more than half* when the total number of dots is even. We hypothesized that the total number of dots is divisible by two, judging how many dots in the target color there need to be in order for the statement to be true becomes more straightforward. Although we expected this difference to be relevant only for *more than half*, it turned out to be important for verifying *most* as well. This finding, however, is not conclusive, as there is a confound of the number of objects subjects had to process: it is possible that the high accuracy we have observed is simply due to the fact that 10 objects are easier to count than 11 or 12.

What we can say for sure is that reducing the number of objects makes judgments about truth or falsity of statements about both *most* and *more than half* much easier. In turn, this leads us to suggest that verifying *most* may also involve precise techniques, when the conditions are favorable. Indeed, although we suggested that *more than half* requires more precise verification, a perfectly strict division of labor between the two quantifiers is unlikely – while most exhibits more approximative patterns and more than half relies more on counting, subjects probably use a combination of these strategies when they are presented with a quantity judgment task.

We have provided evidence that reasoners in our study did not consistently use one specific strategy, but instead relied on several possible options. The three main strategies where we have outlined also have distinct patterns:

1. The "counting" strategy takes equally long to verify *most* and *more than half*, as speakers seem to be precise when verifying both of these quantifiers.
2. The "mixed" strategy requires a longer time to verify *more than half*, suggesting that this quantifier is verified with higher precision.
3. The "fast" strategy is also characterized by different reaction times for *most* and *more than half*.

In summary, only Prediction 1 (the effect of working memory capacity on accuracy and RT effects) got confirmed. According to Prediction 2, only *more than half* would have been affected by the oddness or evenness of the total number of dots, but instead, we found an effect on both quantifiers. Prediction 3 wasn't confirmed, as we found no significant differences in RTs between *most* and *more than half* on the target screen.

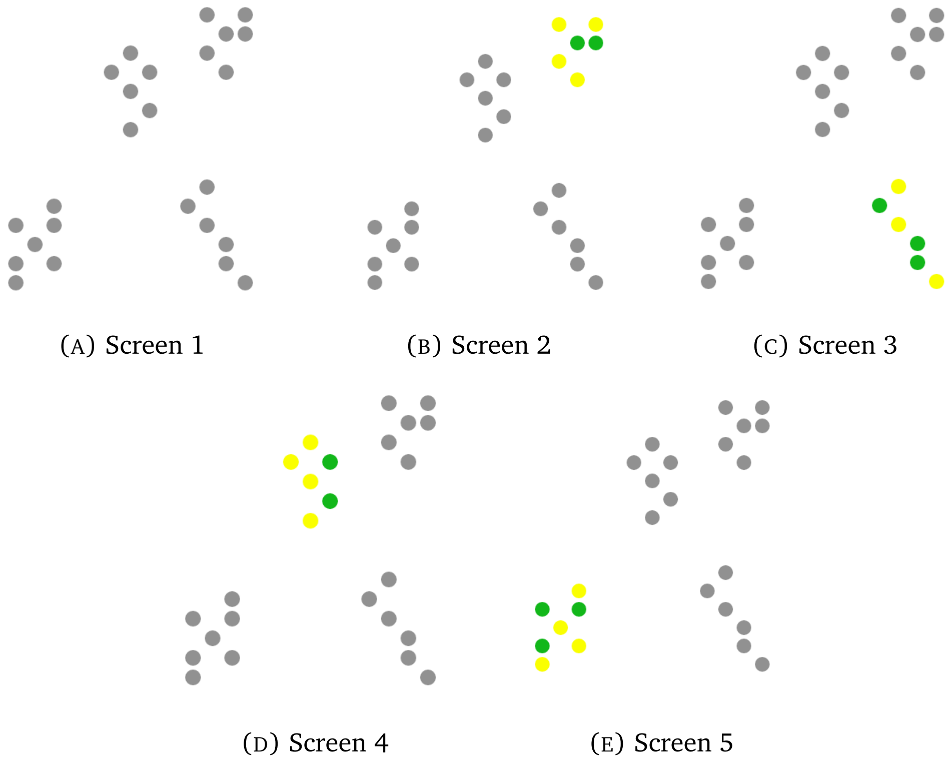
While we have discerned some verification patterns that distinguish *most* and *more than half*, it is unlikely that they sum up to default verification profiles. Throughout our experiment, subjects used a number of different strategies that could have been chosen based on multiple factors, such as cognitive resources, the type of linguistic input and personal preferences.

Given these considerations, we propose that instead of triggering a default verification strategy, quantifiers trigger a *collection of strategies* that can overlap for *most* and *more than half*. This idea has been previously put forward by Suppes (1982), among others, who proposed treating the meaning of a sentence not just as one procedure, but as a collection of those. In the next section, we will show that manipulations with the design of the experiment trigger both *most* and *more than half* to rely more on approximation, which further supports our intuition.

4.2 Experiment 2

4.2.1 Motivation

The results of Experiment 1 have provided some support for the claim that *most* and *more than half* trigger different verification strategies. However, as we have discussed earlier, this does not necessarily imply that the predictions of GQT are incorrect and that their meanings are inherently different: the relationship between meaning and verification is a complex one. It could imply that *most* and *more than half* have default verification profiles that are distinct from each other – but before we can make any conclusions about it, we need to ascertain that the differences we discovered in Experiment 1 are not conditioned on the type of the task the subjects had to solve.



(A) Screen 1

(B) Screen 2

(C) Screen 3

(D) Screen 4

(E) Screen 5

FIGURE 4.10: Sequence of events in the “high” condition.

To follow up on the results of the first study, we will investigate whether the differences we observed between *most* and *more than half* will persevere in a slightly modified experimental setup that is designed to elicit more ANS-like effects. More specifically, we will compare the processing of *most* and *more than half* against two conditions: the “low” condition, in which stimuli will contain 11- and 12-dot arrays, and the “high” condition, in which the total number of dots will vary between 25 and 30 dots (see Figure 4.10).

4.2.2 Predictions

One of the biggest differences between our results from Experiment 1 and Hackl’s study was the spike in reaction times on the first screen. We suggested this could be explained by subjects relying on a more precise strategy in our experiment and trying to estimate or count the total number of dots. In the present study, however, doing so in the “high” condition would be difficult, and we would expect subjects not to rely on this strategy systematically.

Prediction 4. We will not observe a spike in RTs in the first screen.

We have also argued that both *most* and *more than half* warrant a combination of precise and approximative strategies. The former are easier to apply when the total number of objects subjects have to count is relatively low – recall we have observed that accuracy was significantly higher in the 10-dot condition in the previous experiment. As the results of Experiment 1 suggest that *more than half* triggers a more precise strategy, we expect that reasoners would tend to use it when conditions are favorable, as it yields more accurate results. At the same time, we expect that accuracy will be higher in the "high" condition for both quantifiers, as the 3:2 ratio of colors makes it easier to discriminate and estimate the number of dots. On the other hand, we have argued that *most* relies more on approximation even when the number of dots is relatively small, and we would not expect a significant effect of the "high"/"low" condition on accuracy in this case.

Prediction 5. Accuracy will be higher in the "high" condition. The difference in accuracy between "high" and "low" conditions would be greater for *more than half*.

4.2.3 Participants

Thirty two subjects (11 female, 21 male) were recruited via Prolific.ac. Participants were native speakers of English and received a compensation of £2.50. All of them also had success on at least 70% of catch trials. All participants spent over 10 seconds on the instructions page, and their overall success rate was at least 80%.

4.2.4 Materials

The experiment did not measure subjects' working memory and only consisted of one section – the quantity judgment task. Materials were similar to Experiment 1, but the trial items varied between two conditions: in the "high" condition, trials contained 25 or 30 dots, and in the "low" condition, there were 11 or 12 dots.

There were 32 trial items overall (16 most and 16 more than half), with 8 target items in each dot array (2 false *most*, 2 true *most*, 2 false *more than half* and 2 true *more than half*). As before, the difference between true and false conditions was 1-2 dots in the 11- and 12-dot arrays. In the "high" condition, the ratio of dots in the target to the dots in the nontarget color (or vice versa when the statement was false) was 3:2, which is a much easier ratio than the minimum that adults can discriminate (cf. Pietroski et al. 2009). As before, it wasn't clear whether the statement was true or false in the first four screens.

4.2.5 Procedure

As in Experiment 1, subjects viewed the experiment in their web browsers. They had to press the space bar to uncover the dots. In the "low" condition, two or three dots were uncovered at a time. In the "high" condition, they were uncovered in chunks of 6 to 8 dots. When subjects uncovered a new screen, the dots they had previously seen were covered again. Participants could respond during any part of the trial by pressing the “Y” key (for “yes”) on their keyboard if they thought the statement was true or the “N” key (for “no”) if they thought the statement was false.

4.2.6 Results

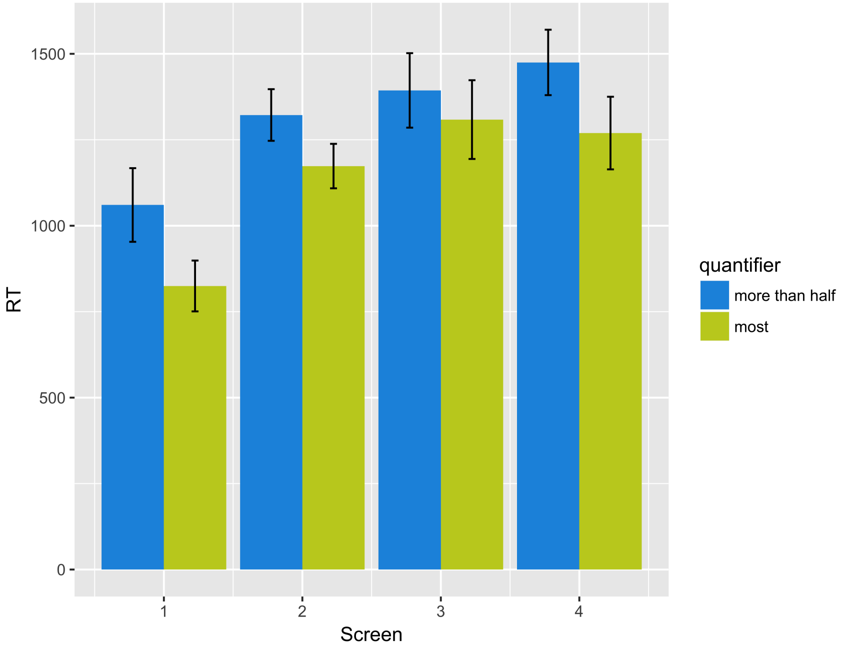


FIGURE 4.11: Mean RTs per screen in Experiment 2

As before, all subjects performed successfully on more than 80% of trials, and none of them needed to be excluded from the analysis.

We found there was no significant difference in overall accuracy between *most* and *more than half* ($W = 133910, p = 0.2109$), but the quantifiers differed with respect to overall reaction time ($W = 834020, p < 0.001$), as *most* was consistently faster. There was also a significant effect of screen ($H(3) = 316.59, p < 0.001$) such that reaction times increased the further subjects proceeded with the trial (see Figure 4.11).

Next, we looked at the differences in RTs and accuracy between the low and high conditions. We found that the two conditions differed significantly with respect to accuracy ($W = 146190, p < 0.001$) as well as overall reaction time ($W = 749240, p < 0.001$). Reaction times were higher in trials with the 25- and 30-dot arrays (see Figure 4.13); accuracy, however, was higher in the “low” condition (see Figure 4.12).

Further, we took a closer look at differences in the “low” and “high” conditions for each quantifier separately. For *most*, there was a significant difference in accuracy ($W = 28282, p < 0.001$), but not in reaction time ($W = 459420, p = 0.096$). Subjects made fewer mistakes in the “high” condition.

For *more than half*, we found a significant difference in accuracy and reaction time. As in the case with *most*, accuracy was higher in the “high” ($W = 29440, p <$

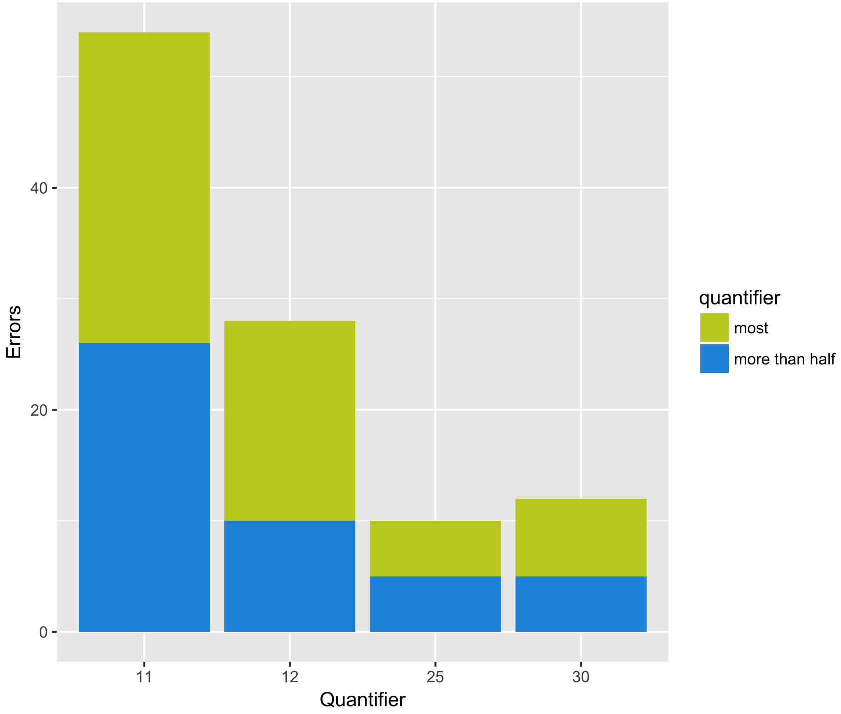


FIGURE 4.12: Accuracy with respect to overall number of dots in Experiment 2

0.001) conditions, and reaction times were shorter in the "low" condition ($W = 502860, p = 0.0032$).

To investigate whether gaps in RTs between *most* and *more than half* were different across screens in each condition, we collected mean reaction times for every subject per every screen, and calculated the difference between *most* and *more than half*. We followed this procedure for the "high" and "low" conditions separately. In the "high" condition, no significant differences were found ($H(3) = 2.4547, p = 0.5$). However, in the "low" condition we found that the gaps in RTs across screens were significantly different from each other ($H(3) = 10.999, p = 0.0117$). As shown in Figure 4.15, the difference in RTs between *most* and *more than half* was bigger in screen 1 than in other screens, which mirrors our finding from the previous experiment.

4.2.7 Discussion

In line with our prediction, there was no increase in reaction time in the first screen. Instead, the distribution of RTs per screen was similar to the pattern observed by Hackl (2009). As Hackl explicitly claims that subjects in his study did not count the total number of dots in the first screen, our results also suggest that subjects did not count the total number of dots while solving the tasks. This implies that a more approximative procedure was used throughout the experiment.

Generally, it appears that the difference between the "high" and "low" conditions impacted verification of *most* and *more than half* to the same extent: the differences

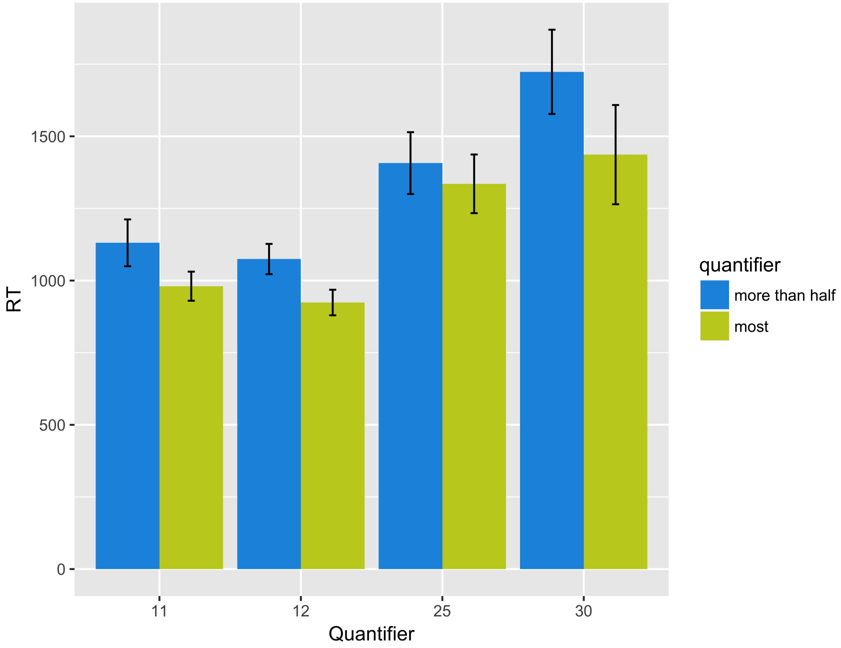


FIGURE 4.13: RTs with respect to overall number of dots in Experiment 2

we have found between the two conditions were there systematically for both quantifiers. We have observed that both *most* and *more than half* took subjects longer to verify in the "high" condition. Likewise, success rates were higher in the "high" condition for both quantifiers.

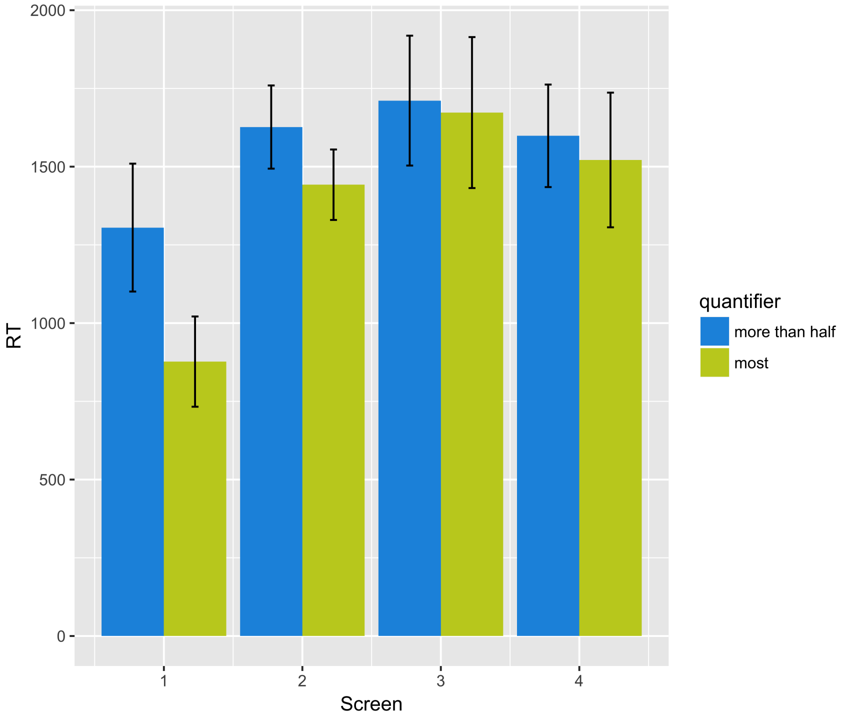


FIGURE 4.14: RTs per screen in the "high" condition of Experiment 2

The latter finding suggests that subjects were more error-prone when there were only 11 or 12 dots on the screen (although in both conditions, *more than half* had higher accuracy than *most*). This adds further support to the claim that subjects adhered to an approximative verification strategy throughout the experiment: in the "low" condition, the difference between target and non-target color was limited to one or two dots, while in the "high" condition, the colors were always in a 3:2 ratio – a ratio that can be reliably discerned by subjects using ANS (cf. Pietroski et al. 2009). This does not, however, explain the higher reaction times in the "high" condition: if subjects approximated more, wouldn't it have take a shorter time to go through trials? A possible explanation is that, as we have argued in the Discussion section following Experiment 1, subjects most likely switch between strategies, combining both precise and approximative techniques. The increased RTs and accuracy in the "high" condition could be the result of subjects doing just that.

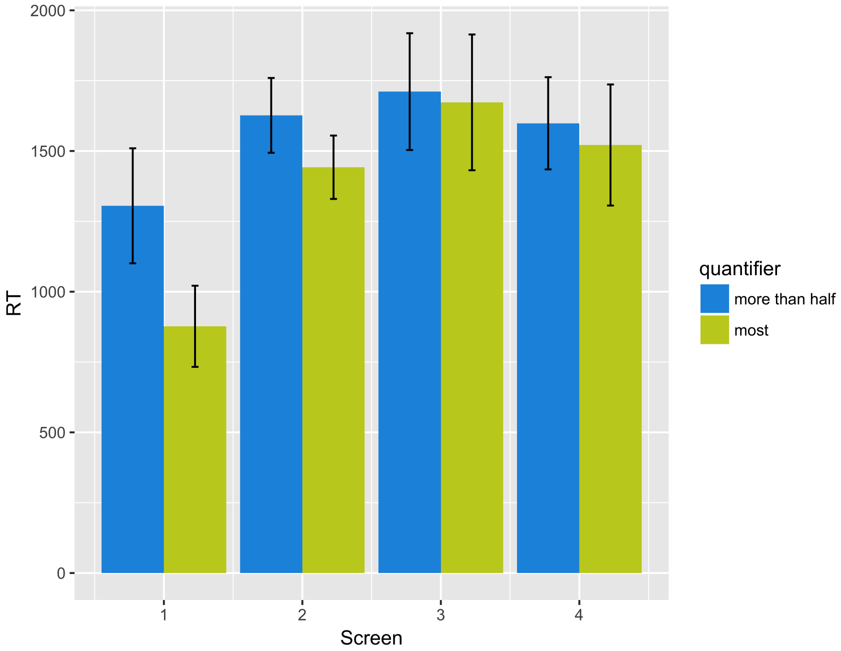


FIGURE 4.15: RTs per screen in the "low" condition of Experiment 2

Finally, we have found that only in the "low" condition the gaps in RTs between *most* and *more than half* were significantly different screen by screen. This offers further support to our previous conjecture that reasoners tend to use a combination of precise and approximative strategies when solving a self-paced counting task. If subjects used a more precise verification strategy in the "low" condition, knowing the exact number of dots would be more relevant for verifying the meaning of *more than half*. That, in turn, would contribute to the differences in RTs between *most* and *more than half*. Conversely, relying on more approximative strategies would make counting less preferable, which would contribute to the more even distribution of RTs between *most* and *more than half* that we have observed in the "high" condition.

Although further investigation is required to compare strategies triggered by *most* and *more than half* in each condition, our current results suggest that in Experiment 2, manipulating the total number of dots affected both quantifiers in the same way: by pushing participants to approximate more. This result provides further evidence to our intuition that *most* and *more than half* do not trigger distinct default verification profiles, and a collection of procedures is available instead.

5 Conclusion

One of the core issues we have raised in this thesis is the relationship between meaning and verification. According to the procedural approach to semantics, meanings can be identified with algorithms for determining whether a given natural language expression is true or not. We have shown that this approach has some empirical weight: we have summarized results from several experimental studies that support predictions of semantic automata theory about the involvement of working memory in proportional quantifier verification.

Still, a growing body of research points to the fact that meaning cannot be completely equated with verification. Given a particular rendition of truth conditions for a natural language expression, we cannot guarantee reasoner will always verify that these conditions are met in a unified way. For instance, the meaning of the quantifier *most* can be formalized in Generalized Quantifier Theory as $|A \cap B| > |A - B|$. Even if this is exactly how all competent speakers interpret *most*, it's unlikely that every time they hear a sentence like *Most of the guests brought gifts*, they would count the guests who brought gifts, count those who came empty-handed, and subtract the latter from the former. Especially since the choice of a verification procedure can depend on so many factors: the cognitive resources of a reasoner, the nature of the task, even the mood of the reasoner.

Lidz et al. (2011) and Hunter et al. (2016) propose a compromise between identifying meaning with verification and completely divorcing the two. They argue that truth conditions associated with a natural language expression trigger default verification procedures best suited for judging whether the conditions in question are met. In line with this idea, Hackl (2009) argues based on experimental evidence that quantifiers *most* and *more than half* have distinct verification profiles. We have noted, however, that the interpretation of these results is not that straightforward, as several confounding factors can interact with the choice of a verification strategy.

As the results of our experiments suggest, there are indeed some differences between *most* and *more than half* – the former tends to rely more on approximation, and the latter tends to trigger a more precise strategy, but *tendency* is the key notion here. These differences do not prevail over the overall patterns we have observed – instead, the choice of a particular strategy in our task depended on various factors. We have shown, for example, that both *most* and *more than half* were impacted by individual working memory capacity, the overall number of dots, and the changes in experimental setup that elicited more approximative procedures.

More importantly, our exploratory analysis of three groups of participants based on their preferred strategy 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 made the point that the meaning of a sentence can be treated not just as one procedure, but as a collection of those. We can expect, then, that some of these procedures overlap for *most* and *more than half*.

Further work needs to be carried out to understand whether this intuition is correct. One of the motivations for the experiments we have presented was the effect context could have on the choice of verification strategy: we have argued that, since subjects had to solve a big number of similar tasks in a row, they could have developed a cognitive strategy that was the most efficient for the type of task at hand, such as trading time for accuracy. Although we have changed the original design of Hackl's study and manipulated it to elicit more approximation in Experiment 2, both of our studies were still part of the Self-Paced Counting paradigm. A possible direction for future research would be to collect empirical evidence from studies that involve multiple tasks and paradigms. Discerning verification patterns and procedures in this context would shed more light on what aspects of verification are intrinsic to meaning. Finally, comparing multiple quantifiers within the context of a verification study would add more clarity as to which verification patterns are shared, and which are triggered by the meanings of particular quantifiers.

6 Appendix

Means and standard deviations of participants' reactions times in Experiment 1

subject	quantifier	mean	sd
$101.csv$	more than half	2860.09	793.31
$101.csv$	most	2807.91	756.53
$102.csv$	more than half	7866.00	702.80
$102.csv$	most	6952.80	1579.63
$103.csv$	more than half	4158.75	1034.06
$103.csv$	most	3663.90	832.24
$104.csv$	more than half	4260.00	2220.21
$104.csv$	most	2485.45	566.21
$105.csv$	more than half	2239.43	330.07
$105.csv$	most	2411.83	715.27
$106.csv$	more than half	1287.33	355.92
$106.csv$	most	1145.00	227.91
$107.csv$	more than half	3071.64	1639.12
$107.csv$	most	2179.50	294.97
$108.csv$	more than half	5074.80	1091.70
$108.csv$	most	3775.00	920.69
$109.csv$	more than half	4168.92	568.62
$109.csv$	most	4732.58	1364.24
$110.csv$	more than half	3180.92	1255.20
$110.csv$	most	2746.17	1234.87
$111.csv$	more than half	10356.09	3042.10

111.*csv*	most	9835.14	1385.85
112.*csv*	more than half	4716.92	1794.00
112.*csv*	most	4102.64	1486.04
113.*csv*	more than half	3487.25	623.94
113.*csv*	most	3055.38	419.54
114.*csv*	more than half	1683.00	1054.53
114.*csv*	most	2675.43	1095.10
115.*csv*	more than half	4712.50	1626.09
115.*csv*	most	4584.40	1975.22
116.*csv*	more than half	3491.57	1093.69
116.*csv*	most	2857.33	926.75
117.*csv*	more than half	3028.75	700.77
117.*csv*	most	3051.80	643.29
118.*csv*	more than half	2013.00	557.39
118.*csv*	most	2006.73	430.04
119.*csv*	more than half	3816.00	991.01
119.*csv*	most	2867.90	674.95
120.*csv*	more than half	2778.00	746.21
120.*csv*	most	2218.56	183.18
121.*csv*	more than half	4642.55	1121.63
121.*csv*	most	4465.50	953.77
122.*csv*	more than half	4878.40	1828.07
122.*csv*	most	3736.33	1034.45
123.*csv*	more than half	7407.27	3026.72
123.*csv*	most	6846.20	2800.11
124.*csv*	more than half	3790.33	1257.43
124.*csv*	most	3094.09	657.64
125.*csv*	more than half	5039.60	3751.65
125.*csv*	most	3628.30	1279.73
126.*csv*	more than half	2635.11	1760.46
126.*csv*	most	3002.90	1127.80
127.*csv*	more than half	3464.30	522.12
127.*csv*	most	2917.56	386.05
128.*csv*	more than half	2157.50	323.17
128.*csv*	most	1999.91	297.78
129.*csv*	more than half	7620.83	4337.45
129.*csv*	most	5024.11	4547.12
130.*csv*	more than half	2613.33	533.80
130.*csv*	most	2346.75	536.20
131.*csv*	more than half	7414.36	2461.90
131.*csv*	most	6817.00	2676.58
132.*csv*	more than half	4671.18	1960.84

132.*csv*	most	3579.92	1144.07
133.*csv*	more than half	10733.50	3079.60
133.*csv*	most	11564.00	5460.21

Means and standard deviations of participants' reactions times in Experiment 2

subject	quantifier	mean	sd
101.*csv*	more than half	2249.52	1725.73
101.*csv*	most	1763.40	1105.68
102.*csv*	more than half	1116.79	845.35
102.*csv*	most	647.80	517.44
103.*csv*	more than half	238.00	119.45
103.*csv*	most	790.81	522.29
104.*csv*	more than half	714.00	441.84
104.*csv*	most	821.08	778.09
105.*csv*	more than half	2863.34	3832.73
105.*csv*	most	2160.80	2065.67
106.*csv*	more than half	436.83	247.03
106.*csv*	most	616.08	321.95
107.*csv*	more than half	1138.36	840.04
107.*csv*	most	1110.77	1185.12
108.*csv*	more than half	854.94	304.93
108.*csv*	most	769.11	368.76
109.*csv*	more than half	828.66	550.47
109.*csv*	most	673.40	293.28
110.*csv*	more than half	803.75	413.55
110.*csv*	most	653.87	202.00
111.*csv*	more than half	1109.82	525.85
111.*csv*	most	2657.73	5360.20
112.*csv*	more than half	798.66	444.66
112.*csv*	most	618.36	243.86
113.*csv*	more than half	1335.47	1782.83
113.*csv*	most	836.93	477.09
114.*csv*	more than half	463.63	137.21
114.*csv*	most	505.61	84.07
115.*csv*	more than half	765.07	739.02
115.*csv*	most	781.82	905.07
116.*csv*	more than half	895.39	404.01
116.*csv*	most	775.43	362.18
117.*csv*	more than half	4885.95	4927.97
117.*csv*	most	4178.63	4560.29
118.*csv*	more than half	1014.04	535.43
118.*csv*	most	930.17	511.55

119.csv	more than half	1661.48	913.65
119.csv	most	1564.14	893.65
120.csv	more than half	1881.92	1748.27
120.csv	most	1327.05	554.46
121.csv	more than half	918.29	1069.05
121.csv	most	633.75	279.41
122.csv	more than half	649.79	292.06
122.csv	most	639.96	360.82
123.csv	more than half	819.04	363.02
123.csv	most	753.50	297.70
124.csv	more than half	1071.85	603.95
124.csv	most	780.33	487.99
125.csv	more than half	1108.15	609.69
125.csv	most	1094.41	1960.29
126.csv	more than half	1180.05	1986.50
126.csv	most	1235.70	1986.09
127.csv	more than half	687.28	313.86
127.csv	most	651.84	322.32
128.csv	more than half	1635.25	1320.93
128.csv	most	1066.07	430.50
129.csv	more than half	924.13	450.26
129.csv	most	966.71	490.83
130.csv	more than half	1379.28	740.86
130.csv	most	1129.35	727.45
131.csv	more than half	751.97	316.56
131.csv	most	725.88	324.00
132.csv	more than half	1178.23	734.09
132.csv	most	1145.13	701.91

Bibliography

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839.

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

van Benthem, J. (1986). *Essays in logical semantics*. D. Reidel Publishing Company, Dordrecht.

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition*, 12(01):3–11.

Bialystok, E., Craik, F. I., and Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, 45(2):459–464.

Bornkessel, I. D., Fiebach, C. J., and Friederici, A. D. (2004). On the cost of syntactic ambiguity in human language comprehension: An individual differences approach. *Cognitive Brain Research*, 21(1):11–21.

Daneman, M. and Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4):561.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Donlan, C., Cowan, R., Newton, E. J., and Lloyd, D. (2007). The role of language in mathematical development: Evidence from children with specific language impairments. *Cognition*, 103(1):23–33.

Dummett, M. (1973). *Frege: Philosophy of Language*. London: Duckworth.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314.

Geurts, B., Katsos, N., Cummins, C., Moons, J., and Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, 25(1):130–148.

Geurts, B. and Nouwen, R. (2007). At least et al.: the semantics of scalar modifiers. *Language*, pages 533–559.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695):496–499.

Hackl, M. (2000). *Comparative quantifiers*. PhD thesis, MIT.

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

Halberda, J. and Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, 44(5):1457.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., and Germine, L. (2012). Number sense across the lifespan as revealed by a massive internet-based sample. *Proceedings of the National Academy of Sciences*, 109(28):11116–11120.

Hale, C. M. and Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, 6(3):346–359.

Horty, J. (2007). *Frege on definitions: A case study of semantic content*. Oxford University Press.

Hunter, T., Lidz, J., Odic, D., and Wellwood, A. (2016). On how verification tasks are related to verification procedures: a reply to kotek et al. *Natural Language Semantics*, pages 1–17.

Izard, V., Sann, C., Spelke, E. S., and Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, 106(25):10382–10385.

Le Corre, M. and Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105(2):395–438.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1):1–12.

de Leeuw, J. R. and Motz, B. A. (2016). Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, 48(1):1–12.

Leslie, A. M., Gelman, R., and Gallistel, C. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, 12(6):213–218.

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

Lindström, P. (1966). First order predicate logic with generalized quantifiers. *Theoria*, 32(3):186–195.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Holt and Co., Inc. New York, USA.

McMillan, C. T., Clark, R., Moore, P., Devita, C., and Grossman, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia*, 43(12):1729–1737.

Miyake, A. and Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Mostowski, A. (1957). On a generalization of quantifiers.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics*, 8(1-2):107–121.

Newton, E. J., Roberts, M. J., and Donlan, C. (2010). Deductive reasoning in children with specific language impairment. *British Journal of Developmental Psychology*, 28(1):71–87.

Odic, D. (2014). *Objects and Substances in Vision, Language, and Development*. PhD thesis, Johns Hopkins University.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., and Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44(3):547–555.

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.

Pullum, G. K. (1991). *The great Eskimo vocabulary hoax and other irreverent essays on the study of language*. University of Chicago Press.

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.

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

Spaepen, E., Coppola, M., Spelke, E. S., Carey, S. E., and Goldin-Meadow, S. (2011). Number without a language model. *Proceedings of the National Academy of Sciences*, 108(8):3163–3168.

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. (1980). Procedural semantics. In Haller, R. and Grassl, W., editors, *Language, logic, and philosophy: Proceedings of the 4th International Wittgenstein Symposium*, pages 27–35.

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 Thorne, C. (2017). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences*, 60:80–93.

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

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.

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.

Zajenkowski, M., Szymanik, J., and Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research*, 43(6):839–853.