The processing of polar quantifiers, and numerosity perception

Isabelle Deschamps [a,b,1], Galit Agmon [c,1], Yonatan Loewenstein [c,d,e], Yosef Grodzinsky [b,c,f,g,*]

[a] Department of Rehabilitation, Laval University, Canada
[b] Department of Linguistics, McGill University, Canada
[c] The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Israel
[d] Department of Neurobiology, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Israel
[e] The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel
[f] Language, Logic and Cognition Center, The Hebrew University of Jerusalem, Israel
[g] INM-1, Forschungszentrum Jülich, Germany

ARTICLE INFO

ABSTRACT

We investigated the course of language processing in the context of a verification task that required numerical estimation and comparison. Participants listened to sentences with complex quantifiers that contrasted in Polarity, a logical property (e.g., *more-than-half*, *less-than-half*), and then performed speeded verification on visual scenarios that displayed a proportion between 2 discrete quantities. We varied systematically not only the sentences, but also the visual materials, in order to study their effect on the verification process. Next, we used the same visual scenarios with analogous non-verbal probes that featured arithmetical inequality symbols (<, >). This manipulation enabled us to measure not only Polarity effects, but also, to compare the effect of different probe types (linguistic, non-linguistic) on processing.

Like many previous studies, our results demonstrate that perceptual difficulty affects error rate and reaction time in keeping with Weber's Law. Interestingly, these performance parameters are also affected by the Polarity of the quantifiers used, despite the fact that sentences had the exact same meaning, sentence structure, number of words, syllables, and temporal structure. Moreover, an analogous contrast between the non-linguistic probes (<, >) had no effect on performance. Finally, we observed no interaction between performance parameters governed by Weber's Law and those affected by Polarity. We consider 4 possible accounts of the results (syntactic, semantic, pragmatic, frequency-based), and discuss their relative merit.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Numerical tasks and instructions that drive them

This paper describes an attempt to get a glimpse at the manner by which natural language quantifiers are processed in the context of numerical comparison tasks. The study of these processes is important because it might shed light on the nature of the representations that are maintained as such tasks are carried out, and may also provide information about possible interactions between linguistic analysis and numerical comparison.

The rich literature on numerical estimation and comparison in humans typically features paradigms where the task is preceded by a verbal preamble: in many instances, participants are verbally instructed, prior to the beginning of the test session, on how they should perform the task – on how they should respond to each stimulus type. Verbal instructions require linguistic analysis. As numerosity experiments typically focus on non-linguistic processes, they seek to minimize the impact of instructions on processing and performance. As we shall see below, the implicit assumption appears to be that instructions, and representations thereof, are immaterial.

The present study, by contrast, focuses on the impact of verbal instructions on processing, in order to investigate their possible contribution to processing in numerosity tasks. That is, we sought to obtain evidence regarding the interaction (or lack thereof) between on-line linguistic analysis and numerical comparison.

Some details might help to make our goal clear. Numerosity experiments typically feature sequences of quantities. The

* Corresponding author at: Edmond and Lily Safra Center for Brain Sciences (ELSC), Silberman Building, Wing 3, 6th Fl., Givat Ram Campus, Jerusalem 91904, Israel.
E-mail address: yosef.grodzinsky@mail.huji.ac.il (Y. Grodzinsky).
[1] Equal contribution.

instructions given are often global.[2] Each trial features a sequence, beginning with an image of a fixed *reference* numerosity *r*, which is followed by another image that contains a *comparandum* numerosity *c*, that is varied systematically around *r*. The task requires a comparison between *r* and *c*. For example, Piazza, Izard, Pinel, Le Bihan, and Dehaene (2004) habituated participants to triplets of numerosities of a particular value of *r*; they then presented a fourth numerosity *c*, which varied from one trial to the next. Instructions, given prior to testing, also varied: in one condition, they asked participants to indicate "*whether the fourth set was larger or smaller than the preceding ones*" (Piazza et al., 2004, p. 548).[3] Discrimination depended on both the size of the quantities perceived, and the distance between them. Performance graphs in all conditions were "asymmetrical and better fitted by the integral of a Gaussian on a log scale than on a linear scale" (Piazza et al., 2004, p. 548), leading Piazza et al. to conclude that our internal number line, against which quantity estimations are made, is compressed logarithmically (as predicted by Weber-Fechner's Law, Dehaene, 1997; Dehaene & Changeux, 1993; Nieder & Miller, 2003), where *r*, *c*, are internally represented as means of a normal distribution with a variance that is fixed across all choices of *r*, *c*. Importantly, Piazza et al. (2004) report no effect of instructions on performance.

1.2. Instructional-symmetry and breaks thereof

If numerosity judgments are fully described as the comparison of the internal representations of the reference and comparandum sets, one expects our cognitive system to carry out the same calculation process whether the perceiver is instructed to verify statements that require comparison of *r* to *c*, or *c* to *r* (e.g., *compare r to c* vs. *compare c to r*). Call this property *I(nstructional)-symmetry*.

Now, consider the form and content of verbal instructions. As standard tasks require the estimation of quantities and comparison between them, instructions often feature quantifiers – linguistic elements that express quantity. These words and expressions have long been subject of intense study by linguists, philosophers, psychologists and mathematicians (Barwise & Cooper, 1981; Keenan & Westerstahl, 1997; Lewis, 1970; Mostowski, 1957; Oaksford & Chater, 2007). To see how quantifiers relate to numerosity, we consider the role of quantifiers in the evaluation of truth in the following sentences:

(1) a. She wears **at least 3** rings
     b. Is **every** man in the room holding a flag?
     c. At least half of the women here are wearing a scarf

In (1a), estimation of the minimal number of rings worn in the scenario must precede truth-value judgment. In (1b), a listener returns "*no*" if there is at least one man without a flag in the room, and "*yes*" otherwise.[4] Sentence (1c) is true just in case the proportion of scarf-wearers among the women in the vicinity of the speaker is half

or more. The use of quantifiers is thus intimately related to perceived (sometimes reported or even imagined) numerosity. Experiments with quantifiers indeed involve quantities, and tap both linguistic and numerosity processes (Hackl, 2009; Heim et al., 2012; McMillan, Clark, Moore, Devita, & Grossman, 2005; Moxey & Sanford, 1986; Pietroski, Lidz, Hunter, & Halberda, 2009).

Next, we note that the sentences in (2a–b), that contain *contrary* quantifiers, have the same meaning when the scenario contains circles of 2 colors and nothing else – they are made true and false by the same scenarios of the *r*/*c* variety:

(2)	*Sentence*	*Scenarios*	
		A	B
		2 red circles; 14 black circles	2 black circles; 14 red circles
	a. **More-than-half** of the circles are red	*False*	*True*
	b. **Less-than-half** of the circles are black	*False*	*True*

Indeed, this equivalence has led many studies to treat verbal instructions as a necessary, yet impertinent, component of numerosity experiments, one that merely needs to be properly balanced. For example, Barth, Kanwisher, and Spelke (2003) balanced the comparative quantifiers *more . . . than* with *fewer . . . than* in the sentences that they used in a task that required verification against scenarios (Experiment 3). No subsequent analysis attempted to separate performance by the *more/fewer* manipulation, presumably because like Piazza et al. (2004)Barth et al. (2003) assumed I-symmetry, namely that equal numbers of sentence tokens of each type renders this contrast orthogonal to the goals of their numerosity test.

However, the quantifiers in (2) *do* contrast in Polarity, a logical property: *More-* and *less-than-half of the circles* license inferences in opposite directions (*many* and *few of the circles*, as well as the comparative quantifiers *more . . . than* and *fewer...than*, are likewise opposed, as illustrated):

(3) Inferences licensed by Monotone Increasing (*a k a*
     positive) quantifiers
     a. **more-than-half** of the students ran
         fast ⇒ **more-than-half** of the students ran
     b. **many** of the students ran fast ⇒ **many** of the
         students ran
     c. there are **more** small circles **than** squares ⇒ there
         are **more** circles **than** squares

(4) Inferences licensed by Monotone Decreasing (*a k a*
     negative) quantifiers
     a. **less-than-half** of the students ran ⇒ **less-than-half**
         of students ran fast
     b. **few** of the students ran ⇒ **few** of the students ran
         fast
     c. there are **fewer** circles **than** squares ⇒ there are
         fewer small circles **than** squares[5]

The set of students who ran fast is a subset of the set of students who ran. The quantifiers in (3a,b,c) license inferences from the former to the latter are therefore *Monotone Increasing* (or upward entailing), positive quantifiers henceforth. Their *Monotone Decreasing* (or downward entailing) negative counterparts (4a,b,c) license the reverse

---

[2] Global are given once at the beginning of the experiment; local ones are provided on each trial. Though these different manipulations may have different performance consequences, participants must activate the instructions on every trial, or else they would not know what task they are performing. As we compare between different instruction types within the same mode of presentation, we are legitimized in suppressing the difference between global and local instructions.

[3] Piazza et al. put little emphasis on instructions. They are not entirely clear on whether they gave declarative sentences that called for a True/False response (*the fourth set is smaller*), yes/no questions (*is the fourth set smaller?*), or embedded disjunctive questions (*indicate whether the fourth set is larger or smaller*) that called for a Smaller/Larger response. These differences may have consequences to verification. Yet no cross-instructional difference is reported.

[4] It has been argued that if no man is in the room, the sentence is also true. This position, however, has been contested. In the foregoing, we steer clear from such issues.

[5] Comparatives introduce further complications, but nonetheless feature the Polarity contrast (see below. Cf., also (cf. Schwarzschild, 2008 for a recent review)).

inference – from sets to subsets. Contrasts such as (4) are used by linguists to diagnose Polarity (cf. Fauconnier, 1975; Klima, 1964; Ladusaw, 1980; *passim*).[6]

Exploring Polarity might help us reveal something important about the language/numerosity interactions, as it may be used to crack I-symmetry. In fact, there are scattered experimental hints that Polarity leads to *I-symmetry breaks*, affecting performance in verification against numerosity-containing scenarios.[7] In other words, one might use the equivalence in (2), which helped Piazza et al. (2004) and Barth et al. (2003) to counterbalance instructions, and putatively neutralize their impact, in order to probe the nature of the representations that are involved.

1.3. I-symmetry breaks – linguistic representations in verification tasks

Preliminary hints about the involvement of instructions in numerical comparison tasks come from an experiment reported in Just and Carpenter (1971). Participants verified sentences that contained the comparative quantifiers *many* and *few* against 2 numerosity-containing scenarios (<2 red dots, 14 black dots>; <14 red dots, 2 black dots>). Sentences with *many* were verified faster than those with *few*, suggesting that I-symmetry can be broken. If true, this may suggest that at a minimum, certain properties of verbal instructions may affect behavior in tasks that require numerical comparison.[8]

Just and Carpenter tested different linguistic probes against 2 numerosities. A more complete test would broaden the perspective by featuring a variety of numerical relations, in order to study the relation between instruction probes that contain expressions of quantity (e.g., quantifiers), and scenarios that require comparisons between quantities.

Below, we describe a series of RT experiments that sought to detail the manner and degree to which linguistic representations affect processing in numerical comparison tasks. In particular, we demonstrate (i) that I-symmetry breaks are attested when linguistic, but not symbolic, instructions are administered in a numerical comparison experiment; (ii) that these I-symmetry breaks affect RT additively, and in particular – their effect is independent of the RT signature of numerical comparison processes. We conclude in a discussion of the architectural significance of these results, and possible explanations for the effects we documented.

2. Experiment

We describe an experiment that featured a verification paradigm, coupling several probe types with visual scenarios whose numerosity properties were parametrically varied.

2.1. Probes

The expressions we used as auditory probes were the sentence pair in (2) above in which the proportion is fixed (at ½), as well as the degree quantifiers *many/few* and the comparatives *more … than* and *fewer … than*[9]:

(5)   a.   **More-than-half** of the circles are blue
       b.   **Less-than-half** of the circles are yellow

(6)   a.   **Many** of the circles are blue
       b.   **Few** of the circles are yellow

(7)   a.   There are **more** blue circles **than** yellow circles
       b.   There are **fewer** yellow circles **than** blue circles

Next, we reasoned that if the expected effect stems from the linguistic contrast, then it should vanish when equivalent non-linguistic expressions are verified. To test this, we used a pair of quasi-algebraic inequalities:

(8)   a.   🟨 > 🟦 ; 🟦 > 🟨
       b.   🟦 < 🟨 ; 🟨 < 🟦

Cf. Appendix A for more details on these probes.

2.2. Visual scenarios and the verification task

Consider now the visual scenarios against which the probes in (5)–(8) were verified (Fig. 1). For each sentence or visual combination, r is the numerosity of the circles of the *reference* color, namely the one mentioned in the sentence, whereas c, the *comparandum*, is the numerosity of the other color. Weber's Law tells us that performance is influenced by the proportion r/c. If r and c are far apart, the task is easy; otherwise, it is difficult, affecting error rates as well as RT.

We thus knew that Proportion affects behavior; we further suspected that quantifier Polarity also affects behavior (as Just and Carpenter (1971) taught us). But would there be an interaction between these two parameters? That is, do linguistic and numerosity factors interact? Moreover, would non-verbal instructions bring about similar behavioral effects? An answer to these questions is likely to be informative about the nature of verification strategies, which would bear on the relation between language and numerical cognition. We therefore constructed well-controlled images that feature 7 different r/c proportions, and tested them with the 8 expression probes described above.

---

[6] Note that although to the extent the quantifiers we discuss are negative, this property in rather abstract. That is, we are not claiming that they contain an actual negation as they do not. Thus, we are not concerned here with effects of overt negation – a factor long known to affect performance (starting with Wason, 1959, 1965). Long ago, Clark and Chase (1972) showed that sentences like "*theplusisbelowthestar*" take more time to process than "*theplusisnotbelowthestar*" (see also Clark, 1974 passim). These studies – as well as the recent discourse-related study by Tian, Breheny, and Ferguson (2010) – measure effects of *overt* negation on processing. This, however, is *not* our issue. We are asking, rather, whether an *abstract* negative marker that may interact with the rest of the sentence (as we saw above), impacts performance in perceptual tasks, and if so, how. See Section 4 for elaboration.

[7] Just and Carpenter (1971) and Geurts, Katsos, Cummins, Moons, and Noordman (2010) compared performance in sentences with negative and positive comparative and superlative quantifiers (*at most/at least n* vs. *fewer than/more than n ± 1*). Yet they were interested in the comparative/superlative comparison, not negative/positive. For this reason, their study was not analyzed along lines that are of interest in the present context.

[8] Other results in this study, e.g., a quantifier type by truth-value interactions, are also of interest, but are beyond the scope of the present paper.

[9] For *more-* and *less-than-half*, verification takes place as the ratio between 2 perceived quantities is compared against an internally represented number (i.e., ½). With degree quantifiers, this is somewhat different: Following Geurts et al. (2010), Heim (2000), Hackl (2000), and Just and Carpenter (1971), we take *many* and *few* to denote a set of individuals whose numerosity is to some degree d:

(i) $[[many]] = \lambda d \cdot \lambda x \cdot |x| \geqslant d_{many}$
(ii) $[[few]] = \lambda d \cdot \lambda x \cdot |x| \leqslant d_{few}$

Verification of a sentence with *many* must involve a comparison between d_{image}, the numerosity observed in the scenario (=the cardinality of the set of circles in the image) and d_{many}. Due to noise, this comparison is carried out not on 2 points (or a point and an interval), but rather, on 2 distributions – one for the observed numerosity d_{image}, and an internal one for d_{many}. Each of these distributions comes with its own noise function (with its particular σ). Similar considerations hold for *few*. But d_{many} and d_{few} may not be the same. The latter may be vaguer than the former (on possible reasons for this difference, see the discussion). This vagueness would manifest as a larger noise function (a gaussian with a larger σ) around d_{few}. Thus comparing d_{few} to d_{image} would be harder than comparing d_{many} to d_{image}, which would result in $\Sigma RT_{few} > \Sigma RT_{many}$.

Verification strategies may or may not keep linguistic and numerical processes separate. Separation – independence of Polarity from proportion – would have a distinct effect on error rates and RTs. A lack of interaction between these factors, manifesting as an additive Polarity effect, would be indicative of such a separation. Conversely, if language and numerosity processes are intertwined, then verification strategy would be affected by both Polarity and Proportion, resulting in a Polarity/Proportion interaction. Here is one example of many forms that such an interaction can take: when perception is easy (r and c are distant from one another), perceivers may process all instructions uniformly; it is only when r and c are close ($r/c \rightarrow 1$) and cardinalities are difficult to discern, that they may differentiate between negative and positive.

2.3. Paradigm

The paradigm we used was a version of the Parametric Proportion Paradigm (PPP, Heim et al., 2012), that features expressions and scenarios with a reference numerosity r, and a comparandum c (Fig. 1). Each stimulus thus consisted of an expression probe, followed by an image with r blue and c yellow circles. Scenarios had 2 values of r (16, 24); each r featured 7 r/c (blue/yellow) proportions. Every scenario was coupled with 2 pairs of negative and positive expression probe (5)–(8). Crucially, meaning within each pair of expression is identical (5), (7), (8) or very close (6); other surface parameters are identical in all pairs. Participants were requested to determine whether the statement is true or false against the scenario (cf. Appendix A for more details on images and methods).

The main features of the PPP are:

 (i) *Fixed r, parameterized c:* each PPP trial presents an image with 2 quantities, r and c, as above. One quantity, r comes in 2 different values – 16, 24. For each value of r, a range of values of the other quantity, c, is presented.[10]
 (ii) *Simultaneous presentation of r and c:* the PPP deploys single images that contain 2 quantities per scenario, each containing a different proportion between discrete quantities of blue and yellow circles of systematically varied radii.
 (iii) *Local instruction probes:* each stimulus contains a probe-scenario pair, where the probe precedes the scenario. Probes may be linguistic – auditory sentences with quantifiers as above, or non-linguistic – symbolic, visual, quasi-algebraic expressions that contained inequality signs flanked by squares of each color (e.g., ▢ > ▪, ▪ < ▢), which participants were asked to verify against the scenarios. Instruction probe pairs did not differ in structural complexity.
 (iv) *Truth-Value-Judgment Task (TVJT):* following the probe-scenario pair in each trial, participants are asked to indicate the truth-value of the probe by pressing a true/false button.

3. Results

3.1. Response accuracy

3.1.1. A Polarity effect only in the linguistic conditions

Error rates were relatively low (except on the condition in which $r/c = 1$, in which performance was around chance). A significant Polarity effect was found in the linguistic conditions, such

that more errors are made on trials with negative quantifiers than on their positive counterparts ($p < .002$ in a 2-tailed test). Exploring the source of this effect, we found that sentences with negative quantifiers induce more error across the board, and that this effect is independent of task difficulty (that grows as $r/c \rightarrow 1$).[11] No such effect exists in the non-linguistic conditions ($p < .24$ in a 2-tailed test). There was an Instruction type × Polarity interaction effect on performance ($F(1, 16) = 15.295$, $p < .001$) (Fig. 2).

3.1.2. Weber's Law

To test whether Weber's Law applies in the error domain, we plotted the mean proportion of the responses that indicated subjects' belief that the scenario they saw contained more yellow circles than blue ones ("yellow responses" henceforth) as a function of the number of yellow circles.[12] Crucially, here we only cared about participants' belief about the truth-value of the sentence, and therefore ignored its actual truth-value.

When r, the number of blue circles, is fixed, the proportion of yellow responses is an estimator of a participant's ability to discriminate the yellow from the blue numerosity – the farther it is from 0.5 in either direction, the higher discriminability is. This response pattern can be approximated by a normal Cumulative Distribution Function (nCDF). Assuming Gaussian noise, Weber's Law predicts that this pattern be better approximated when the proportion of yellow responses is placed on a logarithmic, rather than on a linear scale (Fig. 3).

As the figure indicates, the fit on a linear scale was high, hence improvement on logarithmic compression could only be slight. It was slight indeed, but found in all cases (compare **A** to **B**, and **C** to **D**). Standard deviations of the fitted nCDFs (on a logarithmic scale) were similar across expression types and both values of r, as Weber's Law predicts.

3.2. RT

3.2.1. A Polarity effect only in the linguistic conditions

In the linguistic conditions, a significant Polarity effect was found in the averaged RT for both r values (Fig. 4) ($F(1, 16) = 54.118$, $p < .001$ for **A**; $F(1, 16) = 39.361$, $p < .001$ for **B**);

---

[11] To uncover the source of the higher error rate for negative quantifiers (i.e., whether it stems from quantifier Polarity or scenario Proportion), we modeled the probability to choose "yellow" with two parameters:

$$P(choose\ yellow) = \varepsilon + (1 - 2\varepsilon) \int_{-\infty}^{\#yellows} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\log(ratio)}{2\sigma^2}}$$
$$= \varepsilon + (1 - 2\varepsilon)nCDF(\#yellows|\sigma).$$

The parameter ε, the baseline of this function, reflects *constant* load across proportions in different scenarios; σ, the width of this function, reflects load that is *dependent* on proportion. Fitting this equation enabled us to extract ε and σ for each instance (i.e., each combination of polarity * referendum * quantifier-type). We then ran a permutation test on these data (2^{17} permutations for each combination), and then took the difference between the ε's and σ's of negative and positive quantifiers for each referendum * quantifier-type combination. We found a significant effect for the difference in Polarity of ε ($p = 0.0378$, <0.001, <0.001, =0.0263 for fixed-proportion quantifier * $r = 16$, fixed * $r = 24$, degree quantifier * $r = 16$ and degree * $r = 24$ respectively). The difference between the negative σ and the positive σ was far from significant (except for fixed * $r = 16$, $p = 0.0233$).
The results of this test therefore indicate that the higher error rate of negatives is not dependent on the scenario but arises only from the linguistic polarity.

[12] Yellow responses, indicating the belief that |{yellow}| > |{blue}|, were scored as follows: when subjects marked as true "*more than half of the circles are yellow*" and "*less than half of the circles are blue*", and when they marked as false "*less than half of the circles are yellow*" or "*more than half of the circles are blue*". All other responses were taken to indicate subjects' belief that (|{yellow}| > |{blue}|) = |{blue}| ⩾ |{yellow}|. We chose this statistic due to the fact that our referendum numerosity (whether $r = 16$ or $r = 24$) was blue across the experiment (to avoid unmanageable growth in number of stimuli). While this choice could not possibly create a response bias, as it was balanced across all conditions, it forced us to use "yellow responses" when we plotted errors for the present test.
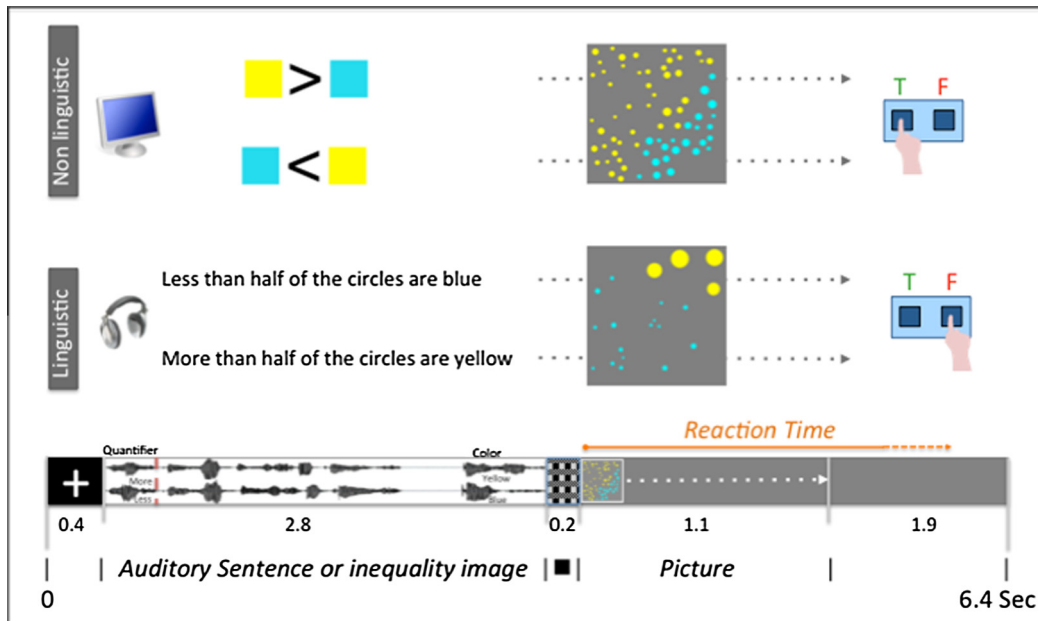
---

[10] This is not the only possible implementation of the PPP: let $r + c = T$. We kept r constant, hence T varied with c. In Heim et al. (2012), we took a different approach: T was fixed at 50, hence r and c co-varied – when r grew, c became smaller by the same amount.

Fig. 1. The PPP trial structure. After fixation, an auditory sentence or a visual quasi-algebraic probe unfolds, followed by a time-locked image. Verification (Truth-Value Judgment) is performed by a button-press.
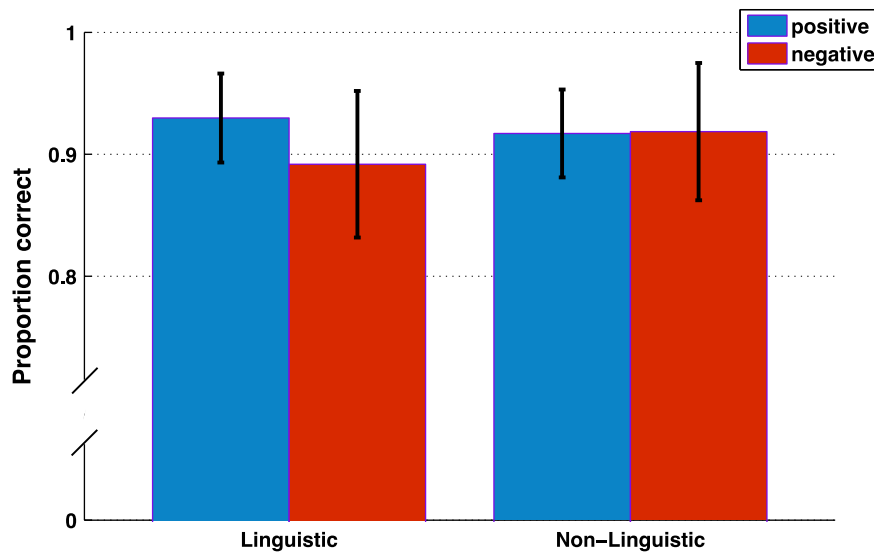

Fig. 2. Proportion of correct answers of all subjects. Columns collapse data for (r = 16, 24). Error bars represent one SD of subject distribution. The r/c = 1 ratio was removed from this analysis, because chance performance is expected on it. Performance was high. Only the linguistic condition resulted in more errors on the negative quantifiers than on the positive ones. The more errors are due to a change in the baseline of the error rate independently of the specific ratio.

as well, there was an Instruction × Polarity interaction effect ($F(1,16)$ = 108.026, p < .001 for **A**; $F(1,16)$ = 57.659, p < .001 for **B**) for both numerosities. Of special interest is the absence of a main effect of the value of r on RT ($F(1,16)$ = .179, p = .678).

3.2.2. Weber's Law

We approximate the relation between the RT and the complex stimuli that consist of instructions and numerosity scenarios by the equation in (9):

$$(9) \quad RT = B + Ae^{\frac{-(\log r - \log c)^2}{2\sigma^2}} = B + Ae^{\frac{-\log\left(\frac{r}{c}\right)^2}{2\sigma^2}}$$

A, B, and σ are parameters. For each r (=16, 24), we plotted mean RT against c, the number of yellow circles, and fitted Gaussian curves (Piazza et al., 2004, supplement). As predicted by Weber's

Law, R^2 as a measure of goodness of fit improved on logarithmic compression of the abscissa (Fig. 5).

3.3. The Polarity effect is additive

Consider now possible verification strategies and the way they would affect RT. As indicated above, Polarity and Proportion (r/c) may be distinct, or they may interact. Separate effects would lead to an additive effect – a difference in baseline B between conditions, but not in amplitude A in formula (9). Such a result would imply that linguistic processing is modularized from numerical cognition. An interaction effect (due perhaps to the interactive verification strategy illustrated above) would lead to a difference in A (and possible in B and σ too).
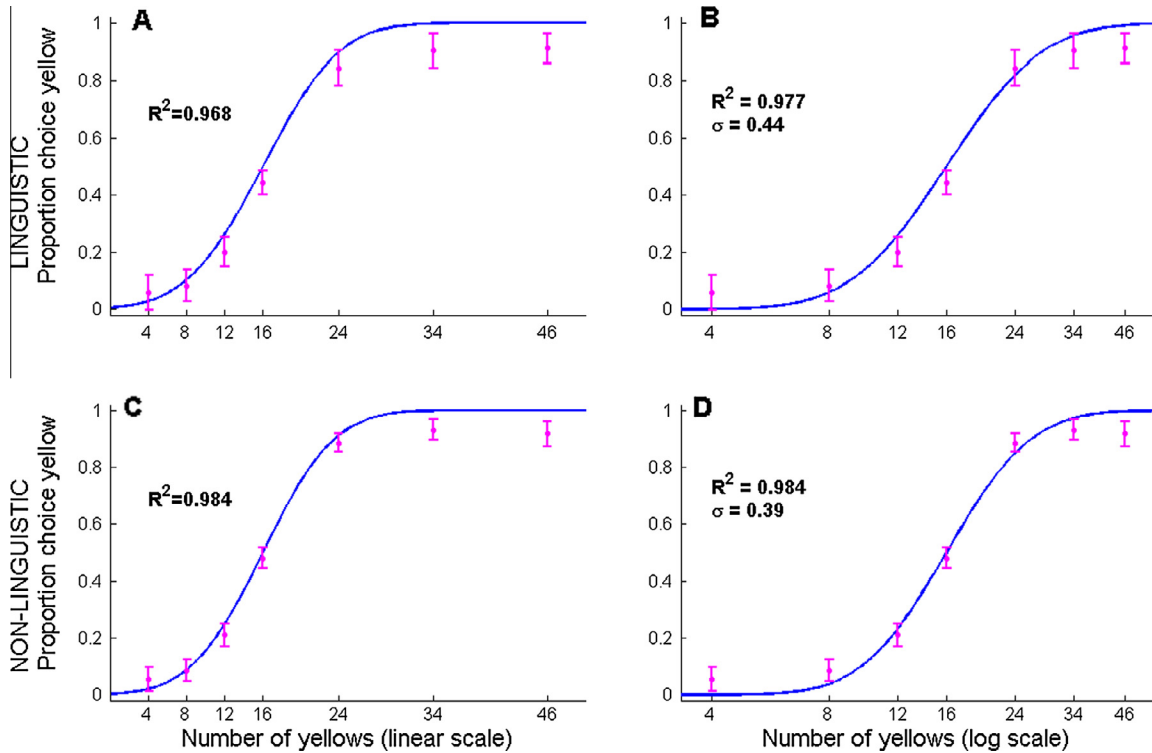
Fig. 3. Performance graphs and Weber's Law. Normal Cumulative Distribution Functions (nCDFs) fitted to the proportion of "choice of yellow" (regardless of whether the sentence contains the predicate *blue* or *yellow*), as a function of the number of yellows in the scenario, *c*, for the linguistic (**A, B**) and non-linguistic conditions (**C, D**) (illustrated for *r* = 16). Error bars represent a confidence interval of 95%. Logarithmic compression (**B, D**) either slightly improves (**A**), or does not change (**C**), the fit relative to a linear scale, as shown by the R^2 measure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
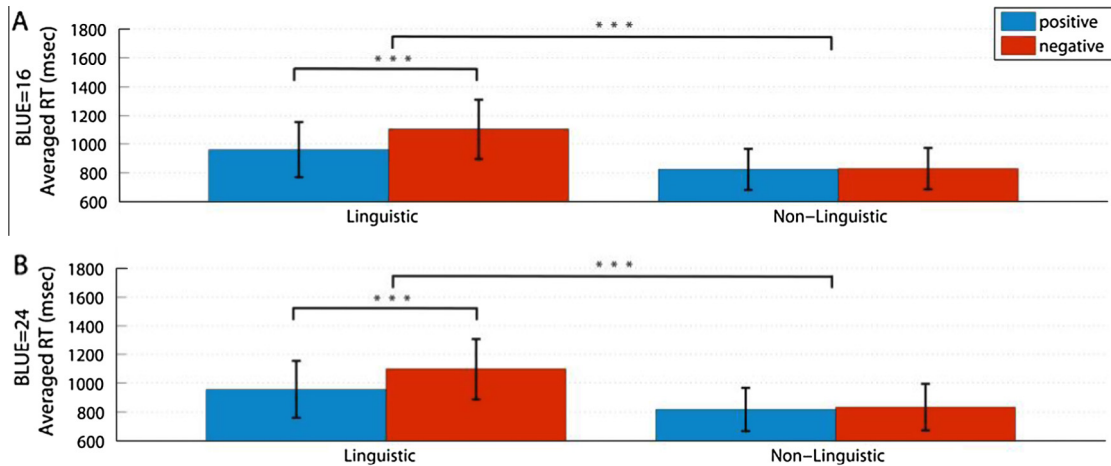

Fig. 4. Mean RT for *r* = 16 (**A**) and for *r* = 24 (**B**), for the linguistic and non-linguistic conditions. Included are RTs for both correct and incorrect responses. Error bars represent one SD of the population. ***$p < .001$.

The hypothesis that linguistic analysis and numerical estimation and comparison do not interact predicts that the difference in the mean RT between positive and negative trials be independent of the r/c ratio. To test this, we averaged the RT for each r/c ratio, separately for the positive and negative quantifiers, and for the 2 reference numerosities, $r = 16$ and $r = 24$. We then computed the RT difference between the two polarities – $RT_{diff} = RT_{neg} - RT_{pos}$.

In keeping with equation (9), $RT_{diff} = \Delta B + \Delta A e^{\frac{\log(\frac{r}{c})^2}{2\sigma^2}}$, where ΔA and ΔB represent the difference between the coefficients in the equations for RT_{pos} and RT_{neg}, respectively. If r/c ratio and the Polarity interact, then ΔA should differ from zero. Independence,

by contrast, implies that $\Delta A = 0$. Noting that ΔA is simply the regression coefficient of RT_{diff} against $e^{\frac{-\log(\frac{r}{c})^2}{2\sigma^2}}$, we ran a permutation test and found that ΔA was not significantly different from zero in any condition ($r = 16$: $p = .469$; $r = 24$: $p = .326$, in a 2-tailed-test), and ΔB was significantly different from zero in both conditions ($r = 16$: $p = .024$; $r = 24$: $p < 0.001$, in a 1-tailed-test), implying that the significant ΔRT between negative and positive quantifiers (Fig. 6) was a global effect that is independent of the r/c proportions. In other words, the effect of quantifier Polarity and the perceptual effect of numerosity comparison on RT are independent.
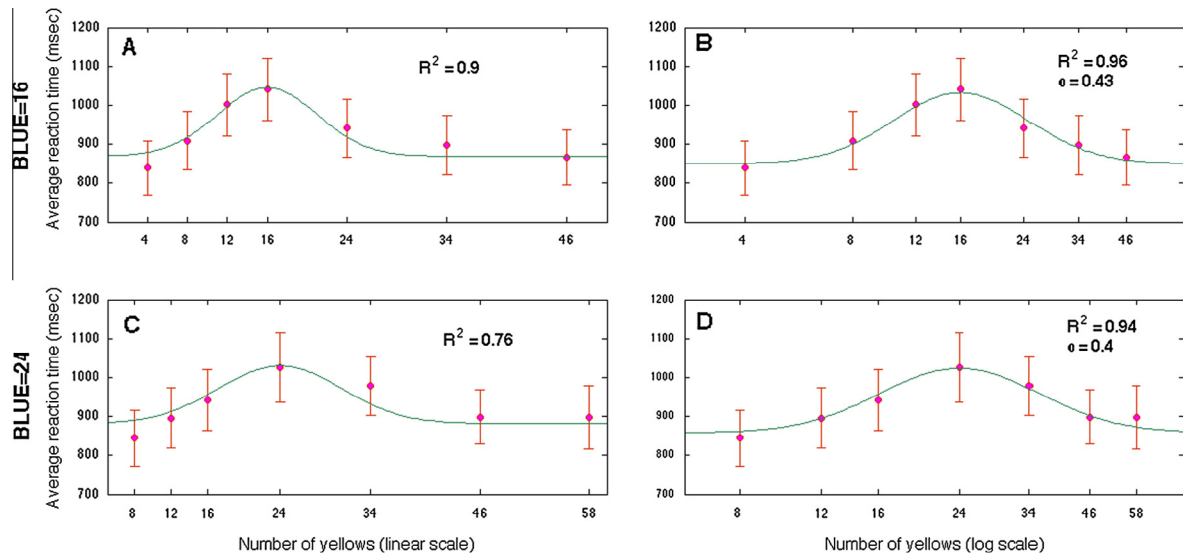
Fig. 5. Improvement of R^2 on logarithmic compression. Top row: all data averaged (linguistic and non-linguistic) for $r = 16$ on a linear (A) and a logarithmic (B) scale. Bottom row: same for $r = 24$, on a linear (C) and a logarithmic (D) scale. Error bars represent a confidence interval of 95%.
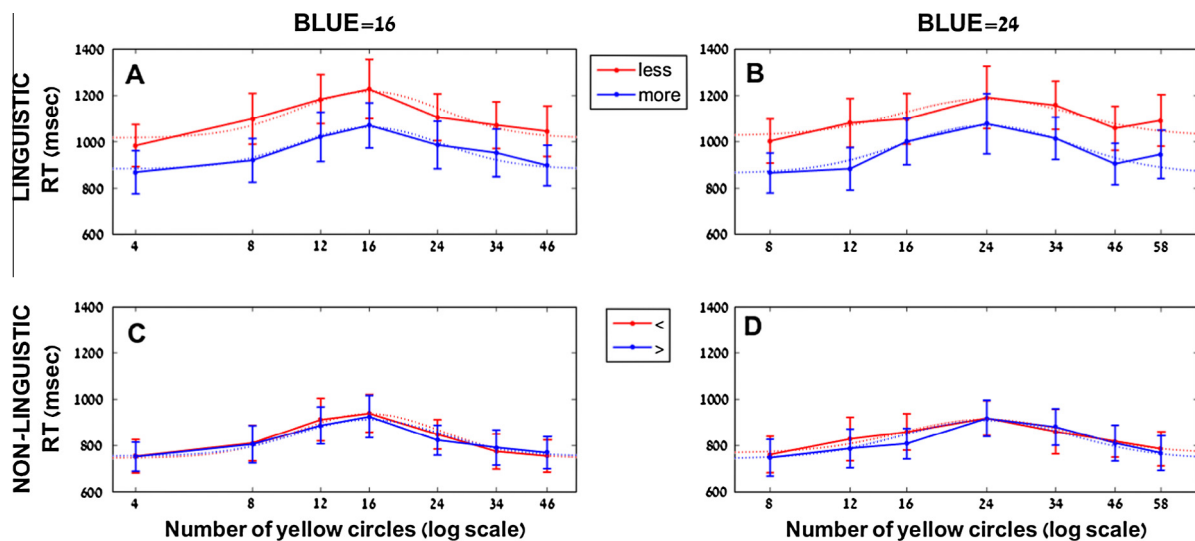

Fig. 6. The breaking of an I-symmetry. Mean RT in the linguistic condition for $r = 16$ (**A**) and $r = 24$ (**B**), show that the negative quantifiers take much longer ($\sim$140 ms on average) than the positive quantifiers. Such effect does not appear in the non-linguistic conditions ($r = 16$ in **C**, $r = 24$ in **D**). Error bars represent confidence interval of 95%. The difference in RT observed in Fig. 4 comes about from a difference in the baseline of each Gaussian and not the amplitude. This means that the linguistic effect of Polarity is modular and is not affected by the specific ratio/scenario.

3.4. Generality of effects – other quantifiers

3.4.1. Degree quantifiers

Most results for (5) were replicated for (6) – the degree quantifiers *many* and *few* (Fig. 7, see note 8 for discussion of their properties). Like in Section 3.1.1, error data were collapsed over ($r = 16, 24$); the results for the 1:1 ratio were removed, because chance performance is expected on it. Error rates for sentences with *many* and *few* were relatively low ($\mu_{many} = 88.26\%$ correct, SD = .067; $\mu_{few} = 83.87\%$ correct, SD = .075), and a significant Polarity effect was found by a paired 2-tailed t-test [$t(8.854, 16)$, $p < .001$]. All the other tests reported for the quantifiers with a fixed standard were performed here, too. The same (or higher) levels of significance were obtained.

RT: on the data for $r = 16$, a 2-way repeated measures ANOVA (Instructions × Polarity) on mean RT yielded a significant effect

of Polarity ($F(1, 16) = 26.853$, $p < .001$) and a significant interaction effect ($F(1, 16) = 38.368$, $p < .001$). A similar ANOVA was run on the data for $r = 24$, and the same effects were obtained ($F(1, 16) = 28.712$, $p < .001$ and $F(1, 16) = 22.551$, $p < .001$ respectively). Here, too, ΔA was not found to be significant in any of the conditions ($r = 16$: $p = .08$; $r = 24$: $p = .147$, in a 2-tailed-test), lending further support to the hypothesis that language and numerical perception are modular. $\Delta B \neq 0$ approached significance for the condition where $r = 16$, and was not significant for $r = 24$.

3.4.2. Comparative quantifiers

This test was conducted with 2 goals in mind: (i) to test the closest linguistic analogue to the non-linguistic condition; (ii) to extend the scope of the effect to the comparative quantifiers *more than* and *less than*, that contrast in Polarity (8). This part had an identical structure to the previous ones (i.e., $8 \times 2 \times 2 \times 4$), but
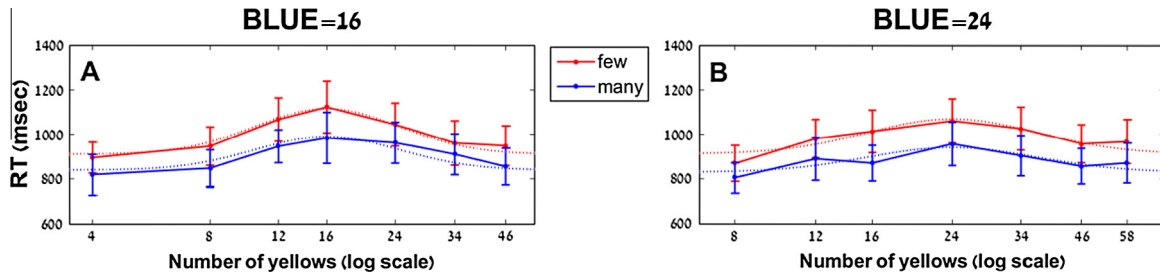
Fig. 7. An I-symmetry break in degree quantifiers. Sentences with degree quantifiers that contrast in Polarity replicate the effect in Fig. 6, for both $r = 16$ (**A**) and $r = 24$ (**B**). Error bars are confidence intervals of 95%.
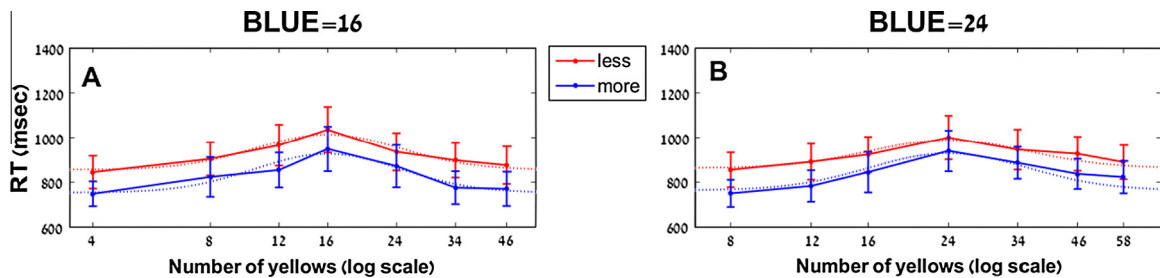

Fig. 8. An I-symmetry break in comparatives. Sentences with comparatives that contrast in Polarity replicate the I-symmetry break seen in Figs. 6 and 7, for both $r = 16$ (**A**) and $r = 24$ (**B**). Error bars are confidence intervals of 95%. (Comparative sentences (8) appear to have yielded shorter RTs, a point of potential linguistic relevance that awaits further exploration.)

was conducted on a different group of participants ($n = 22$). Once again, error data were collapsed over $r = 16$, 24, and the results for the 1:1 ratio were removed, as chance performance is expected on it. Error rates were low ($\mu_{more\ than} = 90.49\%$ correct, SD = .064; $\mu_{less\ than} = 81.85\%$ correct, SD = .104), but a significant Polarity effect was detected through a paired 2-tailed *t*-test [$t(6.859, 21)$, $p < .000$]. All other tests carried out for the other quantifiers obtained the same (or higher) level of significance.

The independence hypothesis receives further support from patterns of RT (Fig. 8): first, a paired *t*-test on the mean RT for both values of *r* revealed a significant Polarity effect (for $r = 16$: $t(21) = 6.647$, $p < .001$); for $r = 24$: $t(21) = 8.014$, $p < 0.001$). Second, we calculated the significance of ΔA and ΔB in keeping with Eq. (9). ΔA was not significant for both values of *r* ($r = 16$: $p < .2687$; $r = 24$: $p < .1188$, in a 2-tailed-test). ΔB, however, was significant for both values of *r* ($r = 16$: $p < 0.001$; $r = 24$: $p < .003$, in a 1-tailed-test).[13]

4. Discussion

Let us first summarize the results of our experiments: (i) Weber's Law is followed in most conditions; (ii) sentence verification exhibits an I-symmetry break, as negative quantifiers produce more errors and longer processing times than positive ones; (iii) this break does not carry over to analogous non-linguistic expressions, which preserve I-symmetry; (iv) the Polarity effect is additive, indicating a lack of interaction between Polarity and Proportion; (v) results are replicated across 3 quantifier pairs and for both numerosities ($r = 16$, 24)[14]; (vi) set size (manifest through the different values of *r* and their accompanying *c*'s) has no effect on RT.

4.1. Four alternative accounts of the Polarity effect and their limitations

We have thus far reflected on the implications of the I-symmetry break we documented to our view on the verification of numerical scenarios. But why are negative sentences slower than positive ones? We consider 4 possible explanatory paths: 1. <u>Frequency-based</u>: *few* and *less* are not as frequent in the ambient language as *many* and *more*, hence the latter are processed faster ($f_{many} < f_{few} \Rightarrow RT_{many} < RT_{few}$). 2. <u>Syntactic</u>: a negative quantifier may force covert syntactic movement of some sort, which a positive one does not. Thus (5–7b) may involve an extra syntactic operation, compared to (5–7a). This operation may takes time[15]; 3. <u>Pragmatic</u>: Polarity correlates with discourse contrasts that may lead to elevated RTs in the negative case; 4. <u>Semantic</u>: the monotonicity properties of a quantifier determine processing difficulty. Sentences with positive, monotone increasing quantifiers (*many*, *more-than-half*, *more ... than*) are easier to process than those with negative, monotone decreasing ones (*few*, *less-than-half*, *fewer ... than*). We briefly consider these possibilities in order. Before doing so, we note that preliminary evidence seems to suggest that negative quantifiers, while sharing certain properties with negation, are not processed as simply containing a negation (Agmon, Loewenstein, & Grodzinsky, 2015).

4.1.1. Frequency of occurrence

According to this account, the RT differences between quantifier types stem from differences in relative frequencies of the words from which the sentences are built. *More* and *many* are more

---

[13] For ΔA, we had no specific hypothesis, hence we ran a 2-tailed test. For ΔB, by contrast, we tested the hypothesis that $\Delta B = B_{neg} - B_{pos} > 0$, hence a 1-tailed test.

[14] As a methodological observation, we note that the improvement in Gaussian fit in all conditions upon logarithmic compression validates the PPP as a numerical comparison task.

[15] Geurts et al. (2010) put forward a related proposal, that negative quantifiers are more complex, hence their processing is longer. We are not clear about how the term complexity needs to be construed here. In the present case, increased *overt* structural complexity of the negative quantifiers is not attested. We also note that their proposal shifts the burden of explanation from the verification process to the internal structure of the negative quantifiers.

frequent than *less* and *few*, respectively,[16] and the result follows. However, this difference cannot account for the results, for three reasons: First, if frequency *f* is negatively correlated with RT, other parts of the data should exhibit this relation as well. This, however, is not the case: e.g., as $f_{more-} > f_{many}$, we expect that $RT_{more-than-half} < RT_{many}$, but the opposite is true.[17,18]

Second, the manner by which RT is measured in the framework of the PPP diminishes the plausibility of an account based on word frequency: each stimulus in this study always began with a quantifier, whose duration is 0.311–0.339 s; all the other words were identical across sentence pairs. Subsequently, an image was presented, at *t* = 2.8 (see Fig. 1 and Appendix A for details). The measurement of RT was time-locked to the onset of the image, and not to the sentence. And yet, the account in question seeks to explain an effect (whose size is rather large) that occurred 2.5 s past the word at issue – participants were in a position to make a judgment only once the image appeared. Words that contrasted in frequency were thus quite remote from the behavior that was measured. Thus, while logically possible, the plausibility of this account seems rather low, and is certainly not in the spirit of the usual frequency-based accounts in psycholinguistics.

Finally, even if error rates and RT were related to frequency, it is difficult to articulate an account that relates them. We would still need *a reason* for the correlation between performance and *f* differentials. That is, observations about frequency and RT differentials may both stem from the same factor(s), but cannot follow from one another.

4.1.2. A syntactic account

It has been argued that Polarity effects do not stem from the monotonicity properties of quantifiers, and that negative quantifiers contain a separable monotone decreasing degree phrase (an abstract negative particle) that may be displaced at some syntactic level by a movement operation (Abels & Martí, 2010; Bech, 1955; de Swart, 2000; Heim, 2001; Penka, 2010; Zeijlstra, 2007). If a piece of a negative quantifier (e.g., a degree phrase) undergoes movement, then the real time processing of sentences that contain these quantifiers – analysis and comparison to a scenario, and subsequent verification – may require more mental operations than their positive counterparts, and their analysis is expected to take longer. This prediction is borne out.

Moreover, the syntactic account connects the present data to independent neurological results: a recent fMRI study (Heim et al., 2012) found that the contrast between sentences with negative and positive quantifiers is only manifested in Broca's region. That is, signal intensity obtained for negative sentences is higher than for positive ones. Broca's region has long been known to support overt syntactic displacement operations (Grodzinsky, 1986; Grodzinsky & Santi, 2008). If negative (but not positive) quantifiers contain a negative operator that must be displaced at some syntactic level, then the anatomical juxtaposition between the negative-/positive and movement/no-movement contrasts is explained by

the assumption that Broca's region is the main brain locus that supports syntactic movement (whether overt or covert). Preliminary studies with Broca's aphasic patients seem to corroborate this view: in a pilot experiment, 3 diagnosed patients failed with negative, but not positive instructions (Grodzinsky et al., 2012). Admittedly, though, the judgment facts that motivate the movement perspective on abstract negative particles are more elaborate than those that we used in our test.[19] A linguistic skeptic might thus doubt the reasons to posit a movement operation in simple syntactic environments such as those we used. The neuro- and psycho-linguistic evidence above should put these doubts to rest, in our opinion.

4.1.3. Discourse properties of quantifiers

Moxey and Sanford (1986, Moxey, 1993; see Nouwen, 2010 for a recent review) discuss an interesting contrast that regards so-called complement anaphora (e.g., Kamp & Reyle, 1993).[20] Certain quantifiers enable pronouns in subsequent discourse to refer to the *complement set* of Q (i.e., to $D_e - Q$), as in (10), where italicized expressions corefer, and "#" marks an odd discourse:

(10) a. *A few senators* came to Congress last week. #*Their* absence was noted
      b. *Few senators* came to Congress last week. *Their* absence was noted

The pronoun in the second sentence of (10a) is interpreted as referring to all senators who came to Congress last week. Thus, the proposition expressed by the first sentence – that that they came to Congress (hence present) – contradicts the presupposition of the second sentence – that they were absent. Curiously, no such contradiction is noted for (10b). Here, the pronoun may either refer to the set of individuals that the generalized quantifier *few senators* denotes (the REFSET, which would lead to the same contradiction as in (10a)), or to its complement set – the rest of the individuals, namely the senators who did not come to Congress ("COMPSET"). The latter reading, they note, is not contradictory. The same contrast is observed for other quantifier pairs.[21]

The contrast in (10) has led to discoveries of processing contrasts (Moxey, Sanford, & Dawydiak, 2001; Sanford, Dawydiak, & Moxey, 2007), which have been taken to suggest that "denial" – presumably a component of the interpretive process of negative quantifiers – is the crucial factor that distinguishes between pairs such as *few* and *a few* in terms of their discourse function. The

---

[16] The COCA frequency count (http://www.wordandphrase.info/frequencyList.asp) returned the following values: *more* 1081183; *less* 166789; *many* 437597; *few* 226602 (occurrences in a large web corpus).

[17] A within subjects paired *t*-test of difference between mean RTs for each quantifier came out highly significant (*t*(7.494, 16), *p* < .000, 2-tailed).

[18] Note that when the units of frequency measurement are changed, and instead of a single word, ones measures the frequency of whole quantifier, the frequency relation changes. That is, $f_{more-than-half} < f_{many}$ and the results follow. But such an account would require a radical change in the nature of frequency accounts: these would no longer be based on neither words not sentences as units of analysis, but rather, on word strings whose size is determined by principles that are not related to frequency. The present suggestion seems to relate unit size to a linguistic principle (by looking at quantifier in its entirety), however no motivation for this choice is provided. It is therefore difficult to evaluate this approach before it is more clearly articulated, and some of its consequences are derived.

[19] The motivation for this abstract movement typically comes from so-called split scope – instances in which an abstract negation assumes scope over another scope-bearing element in the same sentence (e.g., a modal verb), like in this famous German example (Bech, 1955):

(i)   Du mußt keine   Krawatte anziehen
       You must not-one tie   wear

(i) is ambiguous: one meaning, similar to English, implies that a tie is forbidden; a second, more accessible, meaning reads: you don't have to wear a tie. It is the latter meaning (unavailable in the English counterpart) that is of interest, because it shows that a negation, implicit in the negative existential *kein*, takes scope over the modal verb *müßen*. Movement of an abstract negative particle is therefore a suitable explanation.

[20] In their papers, no formal definition of a negative quantifier is proposed. For simplicity, we assume that their criterion is similar to ours, namely, that a negative quantifier can be an NPI licensor and participates in inferences from supersets to subsets.

[21] Kamp and Reyle (1993) also point to cases in which the pronoun refers to the set denoted by the quantifier's restrictor (MAXSET = REFSET ∪ COMPSET):

(i)   Few women from this village came to the feminist rally. No wonder. They don't like political rallys very much.

Presupposition Denial Account (PDA, Moxey, 2006) claims that reference to a COMPSET arises when "the shortfall between a previously expected amount and a smaller amount, denoted by a natural language quantifier, is made salient" (Ingram & Moxey, 2011). If a hearer expects an assertion involving a large reference set, then a sentence referring to a smaller-than-expected set creates a salient "shortfall", that makes the construction of a COMPSET possible. In (10b), the expectation is, roughly, that a large set of senators would arrive. As the sentence asserts that only few did, a shortfall between expectation and assertion is created, which helps build a COMPSET to which the pronoun in (10b) may refer. No such shortfall exists in (10a), and so pronominal reference is blocked.

This account can be easily extended to the temporal domain: it makes sense to say that the process of denial, said to occur in the context of negative quantifiers, incurs a processing cost, which would predict a processing speed difference between sentences with negative quantifiers and those with positive ones (and perhaps beyond). Our results, that seem to live on the Polarity contrast, are thereby accounted for.

A long discussion of this issue is admittedly beyond the scope of this paper, yet there is an interesting issue we would like to raise in this context, one that concerns comparatives – a construction that as we have seen, evinced an RT difference in one of our experiments (Section 3.4.2). Consider the following contrasts:

(11) a. More *students* than professors showed up for the party last night. #*They* stayed home to study for the exam
   b. More students than *professors* showed up for the party last night. *They* stayed home to grade the exam

(12) a. Fewer *students* than professors showed up for the party last night. *They* stayed home to study for the exam
   b. Fewer students than *professors* showed up for the party last night. #*They* stayed home to grade the exam

These judgments are obtained under the assumption that only students study for exams, and only professors grade them. Peculiar to these facts is the split within each sentence pair, one that cuts across Polarity. That is, (11b) indicates that COMPSET coreference is possible in the absence of a negative quantifier; and (12b) indicates that a negative quantifier does not guarantee the availability of COMPSET coreference.

Can the PDA account for these facts? We think that a refined version, readjusted for comparatives, might do the job. The refinement, however, cannot be directly linked to the negative quantifier itself. Such linking would leave the availability of COMPSET coreference (to the COMPSET of *professors*) in (11b) unexplained, as it contains no negative quantifier that would license a "shortfall"; it would also leave unavailability of COMPSET coreference (again, to the COMPSET of *professors*) in (12b) unaccounted for. A more promising account, it would appear, would attribute this licensing to the (monotone decreasing) environment of the comparative *than professors* in (11b), and its monotone increasing counterpart in (12b). While these considerations obviously fall outside the scope of the present study, they merit careful consideration.

Importantly, even a refined PDA, that bases the shortfall on the monotonicity of the environment is unable to account for the RT difference we measured. The sentences we used (7) are reproduced below as (13):

(13) a. There are **more** blue circles **than** yellow circles
   b. There are **fewer** yellow circles **than** blue circles

We found that $RT_{(14a)} <^{sig} RT_{(14b)}$. According to the refined PDA, there is shortfall of yellow circles in both instances. In (11a), we saw that a pronoun in subsequent discourse can corefer with the COMPSET of the *than* phrase in the comparative (parallel to *yellow* in (13b)); in (12a) we saw a possibility for such COMPSET coreference in the DP that contains the quantifier (parallel to *yellow* in (13a)). A shortfall in both (13a–b) is therefore expected. If it leads to processing costs, no RT difference is expected between (13a–b), contrary to what we found. A refined PDA thus seems to fail here.[22]

Note that the syntactic account we considered (4.1.2), while accounting for the RT data, stops short of handling the discourse coreference facts in (10)–(12), as it is not designed to handle such facts.

4.1.4. A semantic account

Consider the sentences in (2), repeated here as (14):

(14) a. More-than-half of the circles are blue
   b. Less-than-half of the circles are yellow

What would it take to make (14) true in a scenario featuring blue and yellow circles and nothing else? An influential theory (Barwise & Cooper, 1981) proposes an algorithm for successful verification, which distinguishes monotone increasing from monotone decreasing quantifiers which we illustrate for (14). For sentence (14a), which contains a monotone increasing, namely positive, quantifier, first locate all sets of circles in the scenario that satisfy the property *more-than-half of the circles*, namely the family of sets of circles, w, whose cardinality exceeds half of that of C, the set of all circles in the scenario ($|w| \subseteq |C| > \frac{1}{2}|C|$). These are known as WITNESS SETS. Next, find in w at least one set w_B whose members are all blue. Even one w_B is sufficient to make (14a) true – additional sets of blue circles with the right cardinality do not affect the verification process (Fig. 9a).

Compare this process to the one involved in the verification of the equivalent negative sentence (14b) against the same scenario. The search might begin as before, in an attempt to locate in the scenario a set of WITNESS SETS of circles w, whose cardinality is less than half of that of C ($|w| \subseteq |C| < \frac{1}{2}|C|$). Yet unlike before, finding one set w_Y whose members are all yellow would not suffice for verification, because the scenario may still feature sets of yellow circles that are not in w, as their cardinality is $\geqslant \frac{1}{2}|C|$. But to judge sentence (14b) as true, we must ensure that the scenario contains *no* such set – that the cardinality of *every* set of yellow circles w_Y in the scenario is $<\frac{1}{2}|C|$. This process requires more steps than before (Fig. 9b). This WITNESS SET based algorithm predicts a negative/positive *Processing Differential*.[23]Barwise and Cooper (1981) themselves suggest that under their model, "response latencies for verification

---

[22] Another possible idea may be that as *few* implies a shortfall, this implication carries over to *fewer*. This may be the case, yet as the text demonstrates, this assumption is insufficient for an explanation of the phenomena at hand. In particular, the reason for the unavailability of COMPSET coreference in (12b), and its availability in (11b), cannot be explained by these assumptions. Moreover, while *few* and *fewer* are related, it is not clear how. For example, while the meaning representation of *few* contains a contextually determined degree, this may not be the case for *fewer* (cf. note 8 and references cited therein).

[23] The foregoing only pertains to the verification process of true sentences. When the scenario makes the sentence false, new considerations are introduced. Such issues are therefore beyond the scope of our present study. See Koster-Moeller, Varvoutis, and Hackl (2007) for elaboration and experiments.
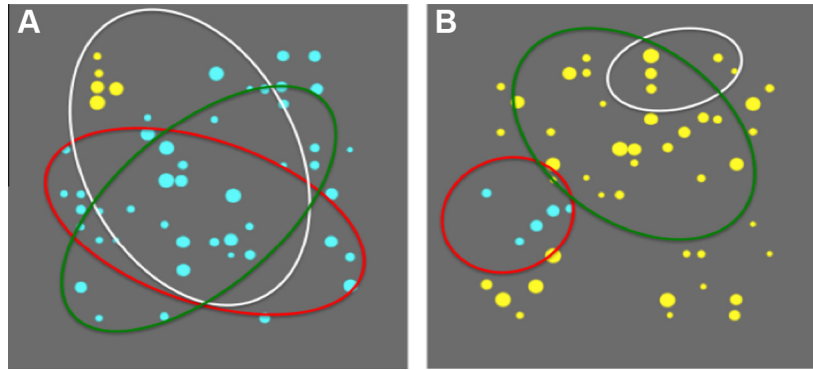
Fig. 9. Verification strategies for positive and negative instruction sentences. (A) Verification algorithm for *more-than-half of the circles are blue*: (i) find the family of sets of circles C that contain each more-than-half of the circles (e.g., white, red, green $\in C$); (ii) if you find one WITNESS SETw_B in C whose members are all blue (e.g., red, but not white), indicate "True". Here, we find another set that satisfy the requirements (green), but that does not affect truth-value. (B) Verification algorithm for *less-than-half of the circles are yellow*: (i) find the set $w_{<½}$ that contains <½ of the circles; (ii) find the family of sets of yellow circles C (e.g., white, green $\in C$; red $\notin C$); (iii) if $C \subseteq w_{<½}$, indicate "True". (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tasks involving decreasing quantifiers would be somewhat greater than for increasing quantifiers" (p. 192, see also Nouwen (2003))[24].

While this approach accounts for the Polarity effect, we note 2 predictions that are not borne out: (a) the ease of identification of a witness set depends on the cardinality of the set of objects in the scenario that make it either true or false – the larger the set of objects, the harder it is to find a witness set. Therefore, an increase in the value of r (and correspondingly, c) should results in an RT increase. This, however, is not what we found (see also Koster-Moeller et al. (2007), for an earlier attempt to detect such an effect of cardinality on RT). (b) It is not clear how the semantic account would account for the Polarity effect found for comparatives, because as stated, the WITNESS SET-based procedure seems to be an inapplicable verification strategy in this case. For these, a natural verification strategy would be to home in on the set in one of the colors (e.g., the set of blue circles), estimate its cardinality, and compare to the cardinality of the set in the other color (yellow). This algorithm predicts *Processing Uniformity*: perceivers verify non-verbal instructions in a uniform fashion, and no effect of instructions on performance is expected.[25] Indeed, this strategy appears to be used for the analogous symbolic, non-verbal, instruction probes, where instead of complex quantifiers, contain inequality symbols like "<", ">", that denote relations between cardinalities.

Finally, we consider the judgment data in (10)–(12), which the PDA successfully accounts for. The semantic account would also be successful, and work along lines suggested in Nouwen (2003). On this view, positive quantifiers do not allow for a COMPSET to be constructed, because propositions that contain these quantifiers are entailed by (hence consistent with) logically stronger alternatives – as *all circles are blue* entails *a few circles are blue*, a COMPSET cannot be defined. By contrast, negative, monotone decreasing, quantifiers allow for the construction of COMPSETS: we find no logically stronger alternative to the propositions expressed by sentences that contain

them (that is, *all circles are blue* does not entail *few circles are blue*). This allows for the construction of a COMPSET.

What about the Polarity contrast in the RT domain, for which we have seen that the PDA fails to account for? The semantic account does not fare any better in this case: it, too fails to explain the RT data for comparatives (unlike the syntactic account, which is successful). The reason for this failure distinguishes it from the PDA, though: witness sets cannot be defined in comparatives, and thus a verification strategy that relies on them critically cannot get off the ground.

5. Coda

The upshot of this study is simple: the linguistic analysis of sentences with quantifiers, and the parsing of quantity-containing visual scenarios, are mostly modular from one another, as real-time processing of such sentences is for the most part unaffected by properties of the scene, and vice versa, in keeping with Weber's Law. There is, however, one point of contact – the Polarity of quantifiers gives rise to an I-symmetry (in keeping with generalized quantifier theory); and while none of the perspectives we considered fully accounts for our complex results (as noted in Section 4), some of them do provide initial clues regarding the intricate relations between quantifiers and quantities, one that deserve further exploration.

Appendix A. Materials and methods

A.1. Sentences and quasi algebraic inequalities

Four pairs of expressions were verified against visual scenarios (note that here, the probes are organized by their Polarity):

---

[24] Alternatively, participants could convert (14b) into its positive equivalent (14a), whose verification procedure is shorter. But the conversion itself constitutes at least one step, again resulting in increased processing complexity for the negative sentence. Yet such conversion is unmotivated. Moreover, versions of it have been considered and rejected in the past, with arguments that seem rather convincing (cf. Carey, 1978; Clark, 1970).

[25] Geurts et al. (2010) tested this type of contrast as part of a larger study that had a different focus. This experiment focused on inference patterns of comparative and superlative quantifiers, but had a small component in which expressions that contain ">2" and "<2" were also tested. This particular piece of their study is not well-controlled from the present perspective. Indeed, the authors merely report the absence of a difference between these 2 conditions, and do not dwell on its potential implications.
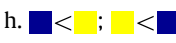
(I)	Polarity	+Linguistic	–Linguistic
	+	a. More-than-half of the circles are blue/yellow	g. ▨ > ▪ ; ▪ > ▨
	–	b. Less-than-half of the circles are blue/yellow	h. ▪ < ▨ ; ▨ < ▪
	+	c. Many of the circles are blue/yellow	
	–	d. Few of the circles are blue/yellow	
	+	e. There are **more** blue circles **than** yellow circles	
	–	f. There are **fewer** yellow circles **than** blue circles	

Table 1

List of proportions in the scenarios: two reference numerosities (r) and their comparanda (c).

r	c						
16	4	8	12	16	24	34	46
24	8	12	16	24	34	46	58

Our first test (carried out with cases (Ia–d) and (Ig–h)) consisted of 6 different probes. Each probe was presented with yellow and with blue as target color, yielding a total of 12 different probes (4 quantifiers × 2 colors + 2 inequalities × 2 colors). Sentences were auditory, with duration of ~2800 ms. Within each pair, sentences differed only in the comparative part of the quantifier (<**more**, **less**>, <**many**, **few**>). Duration of this was virtually identical in all cases: $t = 339$ ms in 3 instances, and 311 ms in the fourth (Fig. 1). Thus overall stimulus duration was the same. True and false items were counterbalanced.

Non-linguistic, symbolic expressions (Ig,h) were visual (Fig. 1). We used two alternative versions. In one, we presented a single visual instruction that contained the complete quasi-algebraic expression for the same duration (2800 ms). We also studied a display mode that more closely followed the sentences, in which the symbols unfolded sequentially in close analogy to the auditory part: first, the leftmost square was presented (1050 ms); next, the inequality sign was added (950 ms); finally, the other square was introduced (800 ms). This display mode not only mimicked the auditory part, but also prevented participants from reading non-linguistic instruction probes right-to-left or using any visual strategy other than left-to-right processing.[26] The results obtained by the two methods are almost identical. For clarity, we report only the results with the non-piecemeal presentation.

All probes – whether auditory or visual – disappeared prior to the projection of the image that contained the blue/yellow proportion of circles.

A.2. Proportion-containing images

Each instruction probe was followed by an image, depicting a scenario in which circles of the reference numerosity ($r = 16, 24$) was coupled with a comparandum number that was varied. For each r, seven different values of c were generated (Table 1). For each r/c pair, eight different visual displays were created by a dedicated Mathematica™ script. Thus, each stimulus was created de novo – participants never saw the same image twice. The process of image creation was as follows: for circles in each color, a quantity that would amount to the total length of radii was picked at random from within a rather broad range. Then, radii were iteratively picked from this quantity, and circles (whose number was predetermined) in that color were drawn (against gray background whose RGB values were exactly midway between blue and yellow). The algorithm guaranteed that the initial quantity was used up. The process was then repeated for the other color.

Crucially, the choice of the initial quantity from which radii were taken was done separately for each color, and was independent of the choice of radii for the other color. These 2 measures guaranteed that circle size/amount of space occupied in the visual display by one color are independent of the other color. As a result, circle sizes and total surface area of a given color could not reliably serve as a guide for judgment. Each image was virtually divided into an array of 100 squares, and circles were randomly placed inside these, with their center randomly placed inside the square. Circles in each color were clustered together in order to make estimation feasible, and to preclude counting.[27]

A.3. Overall trial structure

Each trial started with the visual presentation of a fixation cross for 400 ms. Then participants were presented with either an auditory sentence (linguistic experiment) or a visual display (nonlinguistic experiment), lasting 2500–2800 ms. A visual array of blue and yellow circles, time-locked to onset at $t = 3400$ ms, was then presented, and left on the screen for 1100 ms. For the linguistic conditions, the fixation cross was displayed while the audio sentence was playing to keep subjects' gaze focus on the screen. The total duration of a trial was 6400 ms (Fig. 1). Responses were time-locked to the onset of the visual array.

For the linguistic part of the experiment, eight tokens of each condition type (6a–d) preceded 7 different proportions, resulting in 896 trials [8 tokens × 2 polarities (positive/negative) × 7 comparanda × 2 reference numbers (16/24) × 2 reference colors × 2 quantifier pairs = 896]. For the non-linguistic part of the experiment, eight tokens of each condition type (Ie,f) preceded 7 different proportions, resulting in a 448 trials [8 tokens × 2 polarities (positive/negative) × 7 comparanda × 2 reference number (16/24) × 2 reference color = 448]. Thus, the entire experiment contained a total of 896 + 448 = 1344 trials, counterbalanced for nearly all elements.[28] Participants were instructed to make a TVJT on each probe-image pair, with examples given prior to testing. Each participant was tested in 3 separate 1-h sessions (each consisting of 4 "runs"), including breaks given on request.

A.4. Task

Participants performed the 2 types of TVJT (linguistic, nonlinguistic) by pressing the left or right mouse button. In the linguistic task, subjects had to decide whether the sentence with a quantified subject matched a subsequently presented visual array of blue and yellow circles. In the nonlinguistic task, subjects had to decide whether a visual display depicting a quantified expression using symbols matched a subsequently presented visual array of blue

---

[26] Thanks to Leon Deouell and Martin Hackl for this most helpful suggestion.

[27] This was done by painting all circles blue, and then making the program pick one at random, paint it yellow, then move to the closest circle, paint it yellow, with as many iterations as the desired number of yellow circles.

[28] One part of the design that was not counterbalanced is the color of r and c. That is, r was blue throughout the experiment, because switching it for yellow would have doubled the size of an already huge experiment. Thus the set of blue circles featured in 2 cardinalities (16, 24), whereas yellow featured in many more. During debriefing, participants said they had not noticed that blue featured in a smaller set of values than yellow. We can think of no reason that this imbalance would have affected the results, because *color* is not a factor in our design. Sentences that mentioned *blue* and *yellow* were perfectly balanced, and crossed with the other factors.

and yellow circles. To minimize differences between the two tasks, the temporal structure of a trial was kept constant across both tasks.

A.5. Instructions

You will sit in front of a computer screen and receive one of the following sentences as auditory instructions:

. . . *A LIST OF ALL SENTENCES GIVEN IN THIS EXPERIMENTAL SESSION . . .*

After each instruction, an image containing blue and yellow circles will appear on the screen. The number of circles will vary, as will their sizes.
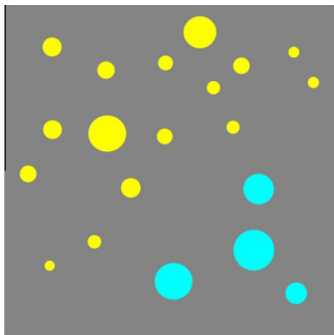
Your task is to determine whether the instruction matches the scenario in the image, and do so as quickly as you can. Press the left button if the sentence is TRUE, and the right button if the sentence is FALSE. Make sure to respond as quickly as you can. Again: left button if the sentence is TRUE, and right button if the sentence is FALSE.

Do not try to use a 'strategy' to perform the task, such as counting individual circles or relying on the approximate surface area of the colors. Such strategies won't work. Do your best to focus on your capacity to quickly estimate the number of each color of circle, and base your decisions on this.

A.6. Participants

21 right-handed (2003) native speakers of English participated in this experiment (mean age 22 years ± 2.2, 9 males), recruited from the McGill student community and paid $10 per hour. Three participants were excluded from the analysis because they did not complete the entire experiment (i.e., they did not return to the lab in order to complete the required 3 sessions) and one subject was excluded due to technical difficulties during testing. We therefore report results from 17 participants, recruited from the McGill community and paid for their participation. All participants had normal hearing and normal or corrected-to-normal vision, as self-reported and were tested to make sure that none was colorblind. All participants gave written informed consent in accordance with McGill University's Research Ethics Board. The comparatives experiment was run separately with 22 participants, selected by the same criteria as in the previous experiment, and with written informed consent as above.

References

Abels, K., & Martí, L. (2010). A unified approach to split scope. *Natural Language Semantics, 18*(4), 435–470.

Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). Why negative quantifiers are really negative. *Second Israeli conference on cognitive research of the Israeli society for cognitive psychology.* Akko (Acre), Israel.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*(3), 201–221.

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

Bech, G. (1955). *Studien über das deutsche verbum infinitum.* København: Det Kongelige Danske Akademie av Videnskaberne.

Carey, S. (1978). Less may never mean more. In R. Campbell & P. Smith (Eds.), *Recent advances in the psychology of language.* New York: Plenum Press.

Clark, H. H. (1970). The primitive nature of children's relational concepts. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 269–278). New York, NY: Wiley.

Clark, H. H. (1974). Semantics and comprehension. In T. A. Sebeok (Ed.). *Current trends in linguistics: Linguistics and adjacent arts and sciences* (Vol. 12, pp. 1291–1428). The Hague: Mouton.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

de Swart, H. (2000). Scope ambiguities with negative quantifiers. In K. Heusinger & U. Egli (Eds.). *Reference and anaphoric relations* (Vol. 72, pp. 109–132). Netherlands: Springer.

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

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5*(4), 390–407.

Fauconnier, G. (1975). Polarity and the scale principle. *Paper presented at the Chicago linguistics society.*

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

Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain and Language, 27*(1), 135–159.

Grodzinsky, Y., Love, T., Ferrill, M., Gutierrez, R., Deschamps, I., Pieperhoff, P., et al. (2012). Broca's region and aspects of negation: A new selectivity pattern in Broca's Aphasia, and a model. *Paper presented at the neurobiology of language conference.* San Sebastian, Spain.

Grodzinsky, Y., & Santi, A. (2008). The battle for Broca's region. *Trends in Cognitive Sciences, 12*(12), 474–480.

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

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

Heim, I. (2000). Degree operator and scope. *Paper presented at the proceedings of SALT X.* Cornell University, Ithaca.

Heim, I. (2001). Degree operators and scope. In C. Féry & W. Sternfeld (Eds.). *Audiatur Vox Sapientiae: A Festschrift for Arnim von Stechow* (Vol. 52, pp. 214–239). Berlin: Akademie Verlag.

Heim, S., Amunts, K., Drai, D., Eickhoff, S. B., Hautvast, S., & Grodzinsky, Y. (2012). The language-number interface in the brain: A complex parametric study of quantifiers and quantities. *Frontiers in Evolutionary Neuroscience, 4,* 4.

Ingram, J., & Moxey, L. M. (2011). Complement set focus without explicit quantity. *Journal of Cognitive Psychology, 23*(3), 383–400.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10*(3), 244–253.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory.* Springer.

Keenan, D., & Westerstahl, D. (1997). Generalized quantifiers in linguistics and logic. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language* (pp. 837–893). Amsterdam: Elsevier.

Klima, E. S. (1964). Negation in English. In: *The structure of language, 245323.*

Koster-Moeller, J., Varvoutis, J., & Hackl, M. (2007). Processing evidence for quantifier raising: The case of antecedent contained deletion. *Paper presented at the proceedings of SALT.*

Ladusaw, W. A. (1980). *Polarity sensitivity as inherent scope relations.* Garland Pub.

Lewis, D. (1970). General semantics. *Synthese, 22*(1–2), 18–67.

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

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

Moxey, L. M. (1993). *Communicating quantities: A psychological perspective.* Hove, UK: L. Erlbaum Associates.

Moxey, L. M. (2006). Effects of what is expected on the focussing properties of quantifiers: A test of the presupposition-denial account. *Journal of Memory and Language, 55*(3), 422–439.

Moxey, L. M., & Sanford, A. J. (1986). Quantifiers and focus. *Journal of Semantics, 5*(3), 189–206.

Moxey, L. M., Sanford, A. J., & Dawydiak, E. J. (2001). Denials as controllers of negative quantifier focus. *Journal of Memory and Language, 44*(3), 427–442.

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron, 37*(1), 149–157.

Nouwen, R. (2003). Complement anaphora and interpretation. *Journal of Semantics, 20*(1), 73–113.

Nouwen, R. (2010). Two kinds of modified numerals. *Semantics and Pragmatics, 3*(3), 1–41.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford: OUP.

Penka, D. (2010). *Negative indefinites.* Oxford: OUP.

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

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

Sanford, A. J., Dawydiak, E. J., & Moxey, L. M. (2007). A unified account of quantifier perspective effects in discourse. *Discourse Processes, 44*(1), 1–32.

Schwarzschild, R. (2008). The semantics of comparatives and other degree constructions. *Language and Linguistics Compass, 2*(2), 308–331.

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology, 63*(12), 2305–2312.

Wason, P. C. (1959). The processing of positive and negative information. *The Quarterly Journal of Experimental Psychology, 11*(2), 92–107.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior, 4*(1), 7–11.

Zeijlstra, H. (2007). Negation in natural language: On the form and meaning of negative elements. *Language and Linguistics Compass, 1*(5), 498–518.