TITLE: The Mental Representation of Universal Quantifiers

Tyler Knowlton*
Department of Linguistics
University of Maryland, College Park

Paul Pietroski
Department of Philosophy
Rutgers, The State University of New Jersey-New Brunswick

Justin Halberda
Department of Psychological and Brain Sciences
Johns Hopkins University

Jeffrey Lidz
Department of Linguistics
University of Maryland, College Park

*Correspondence should be addressed to
Tyler Knowlton
1401 Marie Mount Hall
College Park, MD 20742
tzknowlt@umd.edu

Keywords: natural language quantifiers; semantics; meaning; logic

ABSTRACT

The meaning of sentences like *every circle is blue* could be represented in speakers' minds in terms of individuals and their properties (e.g., for each thing that's a circle, it's blue) or in terms of relations between groups (e.g., the blue things include the circles). Formally, both the tools of first-order logic and the tools of second-order logic can be used to represent the meaning of universally quantified statements. We offer evidence that this formal distinction is psychologically realized in a way that has detectible symptoms. Specifically, we argue that, despite the truth-conditional equivalence of statements with universal quantifiers, *each*-statements are represented in first-order terms but *every*- and *all*-statements are represented in second-order terms. The crucial finding is that participants have a better estimate of a set's cardinality – a fundamentally group property – after evaluating statements with *every* or *all* than after evaluating statements with *each*. Our results support the idea that quantifier meanings are mentally represented at a finer grain size than truth-conditions.

1. Introduction

Words connect pronunciations with meanings. Different English speakers might think different thoughts about rabbits, but to the extent that they can talk about rabbits and understand each other, it's plausible that the pronunciation "rabbit" is hooked up to some sort of shared meaning. A common idea in semantics is that these meanings are abstract entities like the set of rabbits, or a function mapping each possible world **w** to the set of rabbits at **w**.

Similarly, one might think that the meaning of the quantifier *every* is an abstract entity that can be described equally well with any of (1-6) below; where 'R' and 'F' range over sets that correspond to predicates like *rabbit* and *furry*.

$\lambda R.\lambda F.$

(1) $\forall x{:}x \in R(x \in F)$
≈for each thing that's a rabbit, it's furry

(2) $\sim\exists x{:}x \in R(x \notin F)$
≈there is no thing that is a rabbit but not furry

(3) $R \subseteq F$
≈the rabbits are included in the furry things

(4) $R = R \cap F$
≈the rabbits are the furry rabbits

(5) $R - F = \varnothing$
≈the set of rabbits minus the set of furry things is an empty set

(6) $|R - F| = 0$
≈the cardinality of the set of rabbits minus the set of furry things is 0

There is a sense in which (1)-(6) all mean the same thing: given values for 'R' and 'F' (e.g., the rabbits and the furry things), each of these will be TRUE just in case the others are. If

each thing that's a rabbit is furry (like in (1)), then the furry things include the rabbits (like in (4)) and vice versa. Such an equivalence at the level of truth conditions means that different speakers might internalize different, unique, ways of representing these thoughts. That is, one might think that if a word is understood as a *universal* quantifier – like *each*, *every*, and *all* in English – that word's meaning can be described equally well in these and countless other ways. It might be that hearing *every rabbit is furry* causes one English speaker to think something like (1) but causes another to think something like (2).

Here, we offer experimental evidence that the meanings of *each*, *every*, and *all* are not representationally neutral in the above sense. In particular, we focus on an important contrast between two ways of specifying the relation named by the universal quantifiers: in terms of a first-order quantifier, as in (1) and (2), or in various second-order ways that describe it as a genuine relation between two sets, as in (3)-(6). We argue that this formal distinction is psychologically realized in a way that has detectable symptoms. To advertise: if a quantifier *Q* has a second-order meaning, the phrase *Q blue dot(s)* should prompt speakers to represent the blue dots taken together, in a way that promotes representations of "group" properties (e.g., cardinality). This expectation is confirmed for the undeniably second-order *most*; and the relevant pattern of data is reproduced for *all* and *every* but not for *each*.

In one sense, the conclusion that distinct words have distinct meanings will come as no surprise. After all, there are plenty of differences between the three universals, which we review in section 1.2. Grammatical differences, however, aren't enough to show that *each*, *every*, and *all* do not share a representationally neutral meaning. Perhaps the differences reflect grammatical quirks — or as Szabolcsi (2015) puts it, "annotations on the pertinent lexical items". It's an empirical question whether words like *each*, *every*, and *all* are represented by speakers in particular formats, and whether the formal distinctions in (1)-(6) correspond to genuinely different psychological hypotheses.

The remainder of this section clarifies the distinction between first-order and second-order ways of specifying the shared content of expressions like (1-6). We'll also discuss some grammatical differences between *each*, *every*, and *all*, which suggest that any such shared content does not exhaust their meanings. With this background in place, section 2 details the logic of our task, including the prior work on visual number estimation (Halberda, Sires, & Feigenson, 2006) and the relationship between meaning and verification (Lidz et al., 2011; Pietroski et al., 2011) that is crucial to the experiments. Sections 3-5 present the experiments and section 6 addresses potential alternative explanations for our results.

1.1 First- and second-order specifications

The formal analogues of *each, every*, and *all* given in (1-6) are only a handful of the infinitely many ways of specifying the relevant function in extension (i.e., the subset relation). But though (1-6) share truth conditions, each specification represents a different function in intension in that they make use of different logical vocabulary (see e.g., Church, 1941; Tichý, 1969; Horty 2007; Pietroski, 2018). For example, (6) is the only function listed that involves cardinality, and (2) the only one that incorporates negation. One can imagine a mind that could perfectly well think (6), but be unable to think (2), even though both minds would always reach identical conclusions about whether every rabbit is furry. As a result, these specifications might

represent distinct psychological hypotheses about how the lexical item in question is encoded in speakers' minds.

A somewhat coarser difference between these specifications is that while (1) and (2) invoke the tools of first-order logic ($\forall x$ and $\exists x$), (3-6) are second-order in that they describe relations between sets. Intuitively, the difference between first- and second-order specifications is this: first-order specifications characterize the generalization in terms of individual rabbits; second-order specifications do so by comparing the rabbits (taken together) with the furry things (taken together). In (1), individuals and their properties are considered: for each thing that's a rabbit, that thing is furry. In (3), entire sets are related: the set of rabbits is a subset of the set of furry things.[1]

To further highlight this distinction, consider a proportional quantifier like *most*, as in *most rabbits are furry*. This relation can be specified in various ways, including (7-9).

$\lambda R.\lambda F.$

(7) $|R \cap F| > |R \cap {\sim}F|$

(8) $|R \cap F| > |R| - |R \cap F|$

(9) $|R \cap F| > \frac{1}{2}(|R|)$

But *most* can't be specified in first-order terms (Rescher, 1962; Wiggins, 1980; Barwise & Cooper, 1981). What matters for whether *most rabbits are furry* is the ratio of rabbits to furry rabbits (i.e., $|R|$ to $|R \cap F|$); and this can't be captured with a description in terms of the individuals in the domain, as opposed to a description in terms of all of those elements taken together. For this reason, *most* is not first-orderizable.[2]

In contrast, a numerical relation like the one expressed by *four*, is first-orderizable. Namely, the second-order (10) is equivalent to a spelled-out version of the first-order (11).

$\lambda Y.\lambda X.$

(10) $|Y \cap X| = 4$

(11) $\exists x \exists y \exists z \exists w \{Yx \,\&\, Xx \,\&\, \ldots\, Yw \,\&$
$(x \neq y) \,\&\, (x \neq z) \,\&\, \ldots\, (w \neq z) \,\&\, \forall s[Ys \supset (x = s) \vee (y = s) \vee \ldots\, (w = s)]\}$

It might seem implausible that *four* is represented as the cumbersome (11) in the minds of speakers, but it is an empirical question whether the meaning of *four* is more like (10), more like (11), or more like the representationally neutral shared content of them both. Likewise, for the universal quantifiers the relevant relation can be captured with either type of description:

---

[1] Formally, the requirements on being first order are that (i) the domain consists only of individuals (not sets of them) and (ii) each variable quantified over receives only one value per assignment. Changing the ontology by adding sets (e.g., Scha, 1984) or mereological sums (e.g., Link, 1983) is one way to depart from first-order logic. Changing the assignment function by allowing variables to be associated with multiple values at a time (e.g., Boolos, 1984) is another.

[2] Though the relation can also be specified without appeal to cardinalities, given a relation of one-to-one correspondence. Pietroski et al. (2009) offer an argument against *most* being understood this way and Hackl (2009) offers an argument against *most* being understood as *more than half*, as in (9). Finally, Lidz et al. (2011) offer evidence in favor of a specification like (8) as opposed to (7).

in (1)-(2) the relation is stated in terms of individuals; in (3)-(6) the relation is stated in terms of groups of them. And it is an empirical question whether the meaning of e.g., *every* is more like (1)-(2), more like (3)-(6), or more like the representationally neutral shared content of (1)-(6).

Our aim here is not just to pin down the meanings of specific words though, but to develop empirical methods that can aid in discovering which formal distinctions correspond to genuine semantic distinctions, and which are merely notational variants. Our claim is that the first-/second-order distinction has psychological – and hence empirical – consequences.

1.2 Semantic and grammatical properties of *each*, *every*, and *all*

Before turning to some claims about vision and psychology that our experiments rely on, we want to review some independent reasons for suspecting that the meanings of *each*, *every*, and *all* differ in various respects, one of which may be a first-/second-order distinction akin to the contrast between (1) and a representation more like (3) or (4).

The universal quantifiers differ along several dimensions. These include ease of generic construals (Gil, 1992), speaker preferences regarding scope (Ioup, 1975; Kurtzman & MacDonald, 1993; Feiman & Snedeker, 2016), interactions with negation (Beghelli & Stowell, 1997), and compatibility with collective predicates (Vendler, 1962; Dowty, 1987). A commonly reported feeling is that *each* directs attention to individuals, while *all* typically — and perhaps always — invites representations of groups.[3] For example, even if (12) and (13) are truth-conditionally equivalent,

(12) All rabbits are animals.
(13) Each rabbit is an animal.

the generic (12) invites the thought that the category of rabbits belongs to the more inclusive category of animals whereas (13) invites the thought that the property of being an animal applies to each and every thing that we're willing to call a rabbit. Likewise, although (14) is not generic,

(14) The preacher looked at {each/all} of the members of his flock.

using *each* seems to imply many glancing-events, each targeting a different member of the flock, while using *all* conjures an image of a preacher looking out at his congregation with one prolonged stare (Beghelli & Stowell, 1997). This is in line with the observation discussed by Vendler (1962) that *each* forces distributivity and is thus incompatible with collective predicates, as in (15a). But *all* is easily used with collective predicates; though it doesn't force collectivity, as illustrated with the distributive (15b).

[3] This is not to say that it invites representations of *sets*, or groups of any particular kind. For our current purpose of differentiating first- from second-order specifications, nothing turns on such distinctions. But appeal to sets can be replaced with appeals to second-order quantification construed as plural quantification; see Boolos (1984); Pietroski (2005).

(15) a. {*Each / All} of the soldiers surrounded the fortress.
b. {Each / All} of my students can sing that song well as a solo piece.

Grammatically, it seems that *each* requires the partitive *of* in sentences like (16a), but *all* can take the plural noun phrase *the dogs* or the full prepositional phrase *of the dogs*.

(16) a. Each *(of) the dogs barked.
b. All (of) the dogs barked.

Though in *all the dogs*, it may be that *all* intensifies *the*, as opposed to serving as a quantificational determiner (*cf*. (17), which seems synonymous with (16b)).

(17) The dogs all barked.

It may be more significant that *one* can be freely added to sentences with *each*, without a change in meaning, as in (18a). But with *one*, (18b) is highly marked, and a special prosody/context is required to convey its meaning (i.e., that there is only one dog in the domain and it barked).

(18) a. Each (one) of the dogs barked.
b. All (?one) of the dogs barked.

Every tends to pattern with *each* rather than *all*. This is especially clear with regard to compatibility with collective predicates, as shown in (19); see Vendler (1962) and Dowty (1987).

(19) a. {*Each/?Every} dot is alike.[4]
b. All dots are alike.

Likewise, *every* requires a singular count noun. But it differs from *each* in abhorring the partitive without support by *one*, as seen in (20b).

(20) a. Every {dog is / *dogs are} brown.
b. Every *(one) of the dogs barked.

There are, however, a few respects in which *every* patterns with *all*. Beghelli & Stowell (1997) note that both words can occur with *almost*, as if they both indicate the end point of a scale, and with negation. In these respects, *each* is the odd universal quantifier out, as show in (21) and (22).

[4] Moreover, as Vendler notes, *All of those dots are similar* can be heard as true even if there is no single dimension on which each pair of dots is similar. Though it may be worth noting that *Every soldier surrounded the fortress* does not sound as disastrously bad as the variant that starts with *each.* There are also examples where sentences with *each* can give rise to distributive readings, but those with *every* can't; see section 6.

(21) a. One kid ate almost {all the cookies/every cookie}
b. *One kid ate almost each cookie
(22) a. Not {all the kids/every kid} ate a cookie
b. *Not each kid ate a cookie

Perhaps relatedly, *every* is friendly to generic interpretations in a way that *each*, as we saw above, is not. To take another example, even though (23a) seems like a distributive generalization that is TRUE iff (23b) is TRUE, (23c) carries no generic implication of the sort conveyed with (23a) or (23b).

(23) a. Every rabbit hops.
b. All rabbits hop.
c. Each rabbit hops.

These differences are suggestive of a difference in meaning between *each*, *every* and *all*. But instead of taking these grammatical and semantic facts as the starting point and attempting to account for (some of) them, our approach is to look for independent evidence of a formatting distinction (in this case, first- or second-order specifications). If such a distinction exists, then it can be added to the store of primitives that theorists might turn to when trying to derive such differences with as few stipulative grammatical features as possible.

2. Logic of the task

2.1 Representing ensembles

The logic of our experiments requires us to draw on research from the visual cognition literature. The basic idea is that a first-order format highlights individuals, whereas a second-order format privileges groups. So, if a given quantifier has a first-order format, then evaluating statements with it should, all else equal, direct attention to individuals. On the other hand, if a given quantifier has a second-order format, then evaluating statements with that quantifier should, all else equal, direct attention to groups and cause participants to represent groups.

One consequence of attending to (and representing) a group is that knowledge of its summary statistics – center of mass, density, average size, approximate cardinality, etc. – becomes available (e.g., Ariely, 2001; Chong & Treisman, 2003; Halberda, Sires & Feigenson, 2006; Burr & Ross, 2008; Alvarez, 2011). In fact, representing a group and encoding knowledge of its summary statistics does not require explicitly representing each individual constituting that group (Ariely, 2001; Alvarez & Oliva, 2008). Merely attending to and representing individuals though doesn't afford the same access to summary statistics of whatever group(s) those individuals belong to.

The important point for our task is that attending to and representing a group – compared to attending to the individuals within that group and representing them as individuals – enhances one's sensitivity to that group's cardinality. It's worth mentioning that verbally or mentally counting is an exception to this principle. Here, individuals are attended to and labeled, and an inference is made that the largest numeral reflects the cardinality of the

group. But absent counting, we represent large numerosities with the Approximate Number System (see Dehaene, 2011 and Feigenson, Dehaene, & Spelke, 2004 for review). This system operates over set – or at least ensemble – representations. For this reason, numerosity is often talked about as a kind of perceived property of groups (e.g., Burr & Ross, 2008). On analogy, color and shape are perceived properties of individual objects and one cannot help but encode them when attending to an individual. Viewed this way, it is perhaps unsurprising that attending to and representing groups should yield better estimates of those groups' cardinalities than merely attending to individuals.

We leverage this apparent fact about groups and their properties in our task. Suppose participants are given the sentence in (24) and asked to evaluate it with respect to a scene of different sized and colored dots.

(24) All of the big dots are red.

If *all* has a second-order format, we might expect participants to be biased to attend to the set of big dots when evaluating the sentence. In doing so, they should have a good estimate of the cardinality of the set of big dots. If *all* has a first-order format, we might instead expect them to consider individual dots (and decide whether both predicates apply to each one). In using this strategy, they should have a worse estimate of the big dots. Our main diagnostic then, will be how well participants know the cardinality of the restrictor set after evaluating some quantificational statement.

2.2 Meaning and verification

The second assumption required to get our experiments off the ground is that meanings of sentences carry some weight in determining how those sentences are verified. Lidz et al. (2011) call this the Interface Transparency Thesis (ITT). The ITT states that, all else being equal, people are biased to evaluate a given statement with a procedure that transparently reflects that statement's meaning. For example, Pietroski et al. (2009) show that participants verify *most*-statements with a cardinality-based strategy, even given displays that invite one-to-one correspondence strategies. Indeed, when participants were asked to find the leftover dot in these displays, they used a one-to-one correspondence strategy that is faster and more accurate than relying on cardinality comparisons. But given the exact same display and asked whether *most of the dots are blue*, they resorted to the sub-optimal cardinality comparison.

What explains this variation in verification? The idea is that if nothing else can explain the change in verification strategy, the change in meaning must be to blame. Specifically, the representational format must be to blame: *most* is specified in terms of cardinality, not correspondence (both of which are perfectly good *a priori* candidates for the format of *most*'s meaning). Cross-linguistic work confirms that this particular prediction is borne out for majority determiners in Polish (Tomaszewicz, 2011) and Cantonese (Knowlton et al., in prep).

That the meaning of an expression carries some weight in determining verification and that its weight is measurable is not meant to be a contentious claim. It does not state that meanings *are* verifications or that people *always* use a certain strategy to evaluate a certain statement. Rather, it is meant as a modest point in the spirit of Marr (1982). Namely, the

format of a given thought highlights certain information, which makes certain ways of evaluating that thought more natural than others (see Pietroski et al., 2011 for discussion). To be sure, there are other considerations that go into determining what verification strategy some individual might deploy in a certain situation. The claim is just that the meaning carries some detectable weight. This linking hypothesis, or something like it, is needed to make inferences from observed performance on psychological tasks to the format of the underlying representations.

2.3 Measuring cardinality knowledge

For this task to serve as a metric of a first- or second-order representational format, we need a way of accurately measuring how well a participant knows the cardinality of the restrictor set given some quantificational statement. There are two things to consider when trying to judge the extent of someone's cardinality knowledge. The first is accuracy: on average, how much do they over- or under-estimate (Stevens, 1964)? The second is precision: how internally consistent are their responses (Laming, 1997)? These two parameters are captured by the standard model of magnitude estimation on offer in psychophysics, namely, Gaussian tuning curves ordered along an internal scale (Figure 1) (see Feigenson et al., 2004 for review).

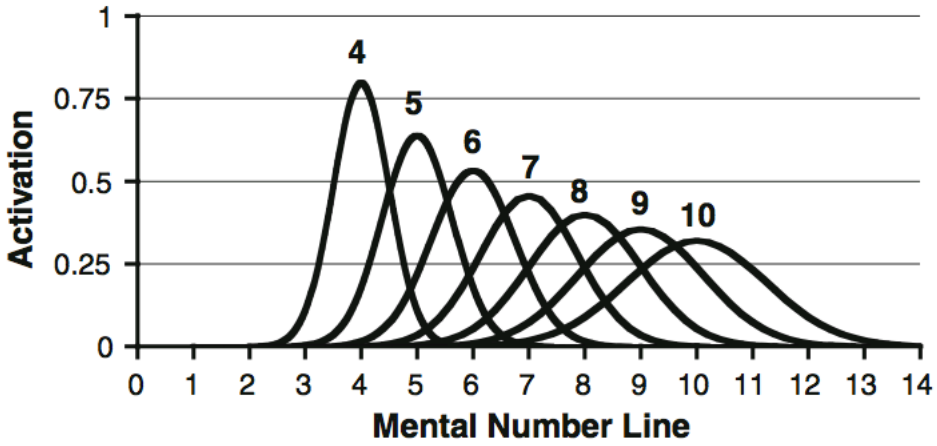

Figure 1: The standard model of the Approximate Number System.

Under this model, numerosities in the world "activate" different parts of the scale, here called the "mental number line".[5] These distributions are linearly ordered such that larger numerosities are represented by activation farther to the right. In Figure 1, each Gaussian is centered over the corresponding numerosity on the mental number line. If the participant modeled by Figure 1 were shown an image of say, 10 dots on a computer screen, and asked to

[5] Though it's worth noting that this model does not only apply to numerosities perceived visually. It has been applied to many other psychological dimensions as well (e.g., loudness, brightness, distance) across multiple modalities (e.g., vision, audition, touch) (Stevens, 1964; Cantlon, Platt, & Brannon, 2009; Lu & Dosher, 2014; Odic et al., 2016).

estimate (without counting) how many there were, they would answer "ten" most often, sometimes answering "nine" or "eleven", less frequently "eight" or "twelve", and so on.

This is the ideal case, but in reality, the numerosity signal from the world might be compressed or expanded in the process of building an internal representation of that numerosity within the ANS. This leads to over- or under-estimation. Informally speaking, the distribution of activation that arises from seeing 10 items in the world might, for a different participant, systematically result in Gaussian activation centered over 8 on their mental number line. Such a participant will systematically under-estimate if shown 10 dots, usually answering "eight", sometimes answering "seven" or "nine", and so on. This tendency to over- or under-estimate is captured by the parameter β.[6] In Figure 1 β = 1, meaning (roughly) that every distribution is centered over the "correct" value on the mental number line. A participant with a β < 1 would systematically underestimate, whereas someone with a β > 1 would systematically overestimate. Figure 2 provides a few examples of β's impact on numerical estimates.

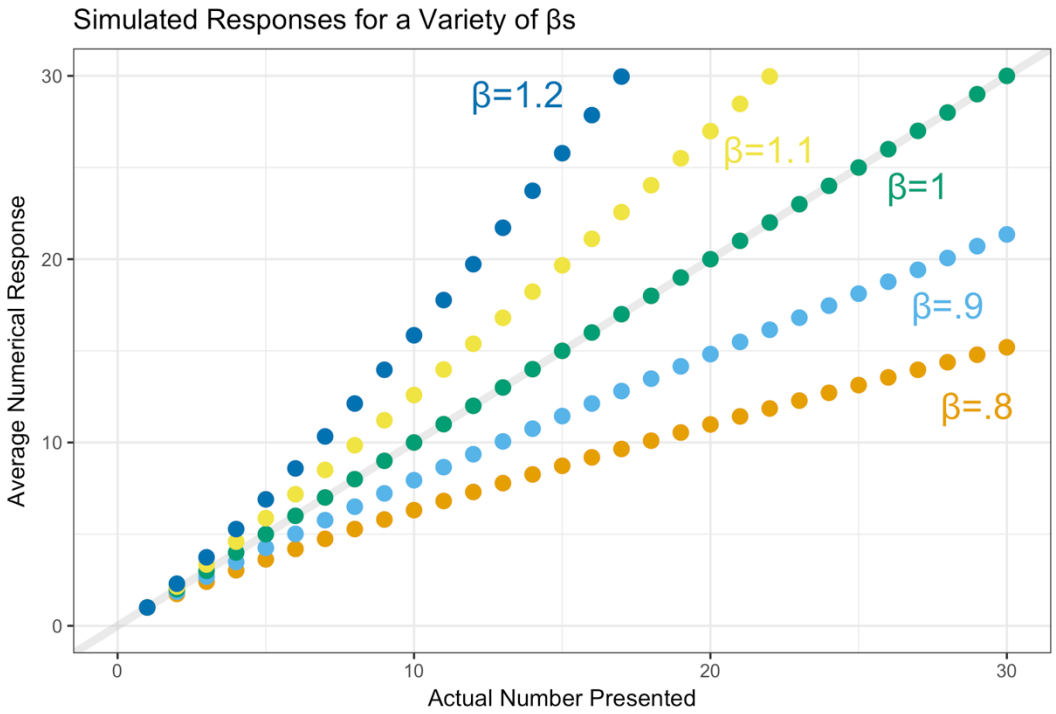

Figure 2: Different values of β and their effect on numerical estimation. *Average Numerical Response* is the mean answer an individual with that β would give when shown *Actual Number Presented*. The grey line represents ideal performance (i.e., β = 1).

[6] The parameter β represents the degree of signal compression/expansion in the equation $y = \alpha x^{\beta}$ where y is the percept (i.e., the participant's numerical estimate, on average), x is the actual number of objects presented, and α is a scaling factor. Alternatively, β may indicate the compression or expansion of a response code that the Gaussian humps of Figure 1 are mapped to (e.g., output compression/expansion) (Izard & Dehaene, 2008); but either a compressed signal or compressed response code would be consistent with our results and theorizing. We will rely on the notion of a compressed signal throughout.

Another one of the model's hallmarks is scalar variability: the fact that the standard deviation of the Gaussian "activation patterns" increases linearly with the mean. The distribution labeled 10 in Figure 1, for instance, is much wider than the distribution labeled 4. This captures the fact that larger numbers come with more internal "noise" (alternatively, less confidence; see Halberda & Odic, 2014 for discussion of this distinction). Scalar variability gives rise to the Approximate Number System's well-documented ratio-dependence (e.g., 9 and 10 are just as difficult to distinguish as 90 and 100) and explains why one is more confident about answering "five" after seeing 5 dots than about answering "fifty" after seeing 50. Intuitively, the distribution activated when experiencing the numerosity 50 has a larger standard deviation and thus overlaps with more numbers on the mental number line.

The rate of increase in standard deviation as number increases – which we'll call σ – develops throughout the lifespan (Halberda & Feigenson, 2008) and is subject to individual differences (Halberda et al., 2012; Libertus, Odic, & Halberda, 2012). If a participant has a large σ, their estimates will be more variable and that variability will grow rapidly as numerosities increase (i.e., they will be less precise). If a participant has a small σ, their estimates will have less variability, and that variability will grow at a slower rate as numerosities increase (i.e., they will be more precise). To take the ideal case, a participant with a σ of 0 will give the same answer every time they are shown some given numerosity. If shown forty dots one hundred different times, our ideal participant would always respond "forty" (assuming their β = 1). Some example values of σ – holding β constant at .9 – are given in Figure 3.

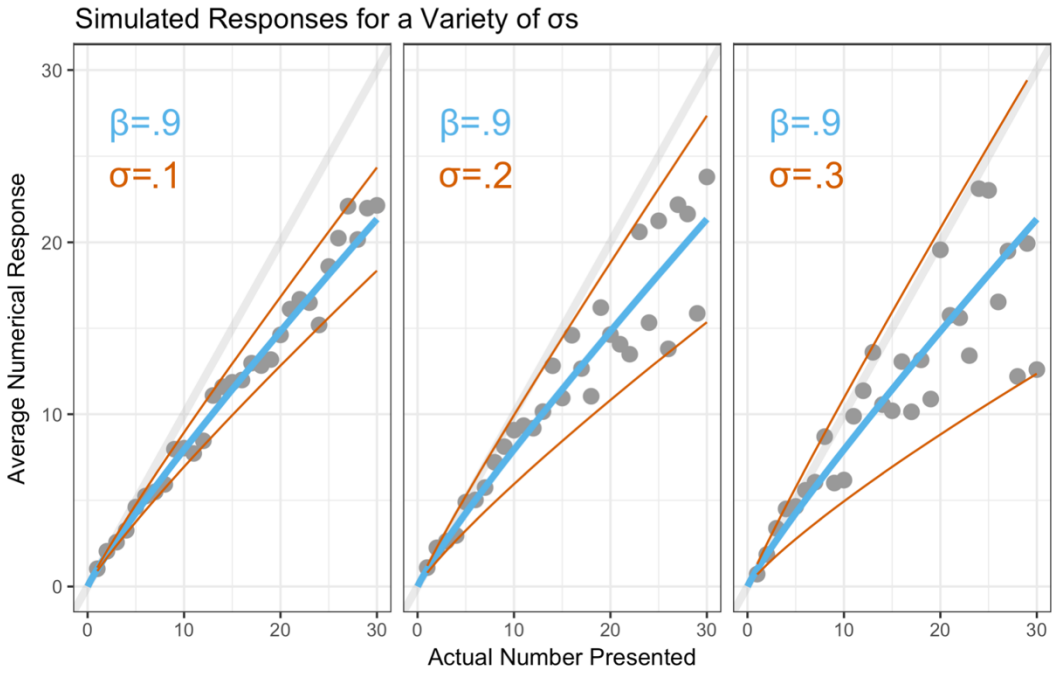

Figure 3: Different values of σ and their effect on numerical estimation holding β constant. *Average Numerical Response* is the answer given in one particular instance by an individual shown *Actual Number Presented*. The grey line represents ideal performance (i.e., β = 1, σ = 0). The blue line is the mean response given β = .9. The orange lines represent the standard deviation of the given numerosity representation.

We're left with two measures of cardinality estimation ability. For simplicity, we can call the level of over- or under-estimation (β) the participant's *accuracy*, and the amount of noise in their estimates (σ) their *variability*. Good performance on a numerical estimation task will result in high accuracy and low variability. To make our graphs easier to interpret and our predictions easier to discuss, we'll use *precision*, 1 - σ, instead of variability. Ideal performance would then be an accuracy and a precision of 1 (i.e., σ = 0). Of course, people often deviate from theoretically ideal performance. For numerical enumeration of visually presented stimuli in particular, previous studies demonstrate that humans show signal compression such that they slightly underestimate the true number (e.g., Krueger, 1984; Odic et al., 2016).[7] This results in best possible performance being an accuracy (β) and a precision (1 - σ) of approximately .8.

For each participant in our task, we can determine their accuracy and their precision after evaluating one quantificational statement or another. To preview: if participants look at a display of dots and evaluate (25),

(25) Most of the big dots are red.

they should attend to the group of big dots, which in turn should lead to higher accuracy and higher precision when estimating the cardinality of the big dots.

On the other hand, suppose participants evaluate (26).

(26) There is a big dot that's red.

They are likely to attend to the big dots individually. Without representing the group, we should expect lower accuracy and lower precision when asked to estimate the cardinality of the big dots. And, if the same holds for statements like (27) but not statements like (28) and (29), this plausibly reflects a symptom of *each* having a first-order format.

(27) Each of the big dots are red.
(28) All of the big dots are red.
(29) Every big dot is red.

In other words, we use attention to groups or individuals as a probe into the mental representation of quantifier meanings. The leading idea is that the mental representation of the quantifier meaning will have detectible consequences on the information that is gathered during the verification process. Importantly, this work capitalizes on a body of research from visual cognition and ensemble representation that provides a rich understanding of the cognitive mechanisms that will be engaged in our verification tasks.

First, we establish accuracy and variability baselines by testing cases when participants should know a cardinality well versus when they should not (Experiment 1). Specifically, we

[7] Along other psychological dimensions – like perceived distance of finger spread – human participants have been shown to systematically overestimate (Gaydos, 1958; Stevens & Stone, 1959; Odic et al., 2016).

show a difference between when participants are told which set to attend to before seeing the display (likely resulting in the set being attended and in memory) versus shown the display before knowing which set they would be asked about (making it less likely to be in memory). Then, using the same kinds of displays, we encourage participants to attend to groups or individuals by asking them to evaluate statements like (25) and (26) before estimating a set's cardinality (Experiments 2a-c). Finally, we repeat the same task pitting *each*, *every*, and *all* against each other with statements like (27)-(29) (Experiments 3a-d).

3. Experiment 1: Cardinality knowledge baseline

Before probing the possible first- or second-order formats of different expressions, we sought to document both accuracy (β) and precision (1 - σ) for human performance under conditions where they would be very likely to attend to and represent the relevant set (i.e., the question highlighting the relevant set appears before the dot display – Question First) and where they would be unlikely to attend the relevant set because too many sets are presented (i.e., the dot display is shown before the question that highlights the relevant set – Dots First). These two cases will allow us to place our later tests in context. Namely, quantifiers that inspire second-order construals (along with attention to the set) should result in performance more resembling "Question First" performance while quantifiers that invite first-order construals (along with attention to the individuals) should result in performance more closely resembling that on the "Dots First" trials.

3.1 Method

3.1.1 Procedure

Following Halberda, Sires, & Feigenson (2006), we relied on a "Dots First" versus "Question First" manipulation, presented across two blocks, to obtain our baseline for accuracy and precision in estimating numbers of attended and less-attended sets. Participants were shown 50 dot displays per block consisting of big, medium, and small dots that could be red, blue, or yellow (see Figure 4). Medium dots had black holes in the middle, to make them more distinguishable from the other two sizes (Chen, 1982; 2005). Participants were shown all three dot sizes during the training portion of the experiment to ensure that they could be correctly identified. Six size/color combinations were shown on the screen at once and a one second viewing time was enforced for all displays. Between 28 and 44 dots were shown on screen in each trial. Each subset present (e.g., big blue dots) contained a minimum of 3 dots and a maximum of 9 dots. Participants were never asked about subsets for which no dots were shown (on every display, three possible subsets were absent).

Their task on each trial was to enumerate one subset (e.g., "How many big dots were there?") by typing in a number on the keyboard and pressing 'enter'. In one block of 50 trials, the dot display was shown for one second, then the question was displayed (Dots First). In the other 50-trial block, the question was displayed prior to the dot display being shown (Question First). There were three question types: Dot Size ("How many {big / medium / small} dots were there?"), Dot Color ("How many {red / blue / yellow} dots were there?"), and Total Dots ("How

many total dots were there?"). Participants were never asked about size / color combinations (there were no trials asking about e.g., "big blue dots").

Following Halberda, Sires, & Feigenson (2006), the Total Dots trials were expected to result in equivalent performance whether they appeared as Dots First or Question First trials. This is because the superset of Total Dots appears to behave as a "default set" for visual processing; roughly, humans seem to always attend the superset of Total Dots on both Dots First and Question First trials. Importantly, if we do not see a difference between the Dots First and Question First conditions on these Total Dots trials, this will demonstrate that observers are indeed capable of attending and enumerating a set during the Dots First trials, just as well as they can on Question First trials.

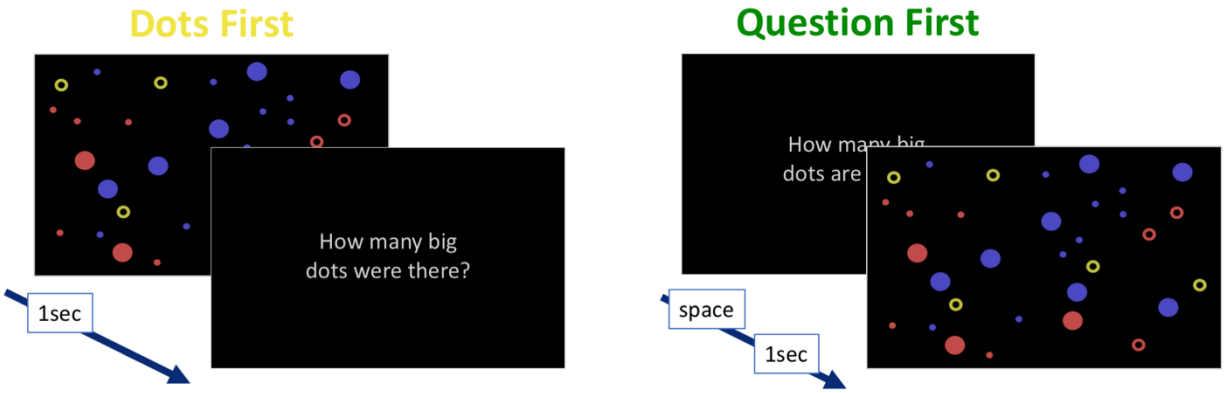

Figure 4: The trial structure of the two blocks of Experiment 1. Participants responded to the "how many x dots were there?" questions by typing a number and pressing enter, upon which they were automatically advanced to the next trial.

3.1.2 Participants and analytic approach

This and all following experiments were built using PsychoPy2 (Peirce et al., 2019). 12 University of Maryland undergraduates took part in this experiment in exchange for course credit (and, in general, we aimed for 12 participants per group in all subsequent experiments). Informed consent was obtained from these participants, and from the participants in all following experiments.

We fit six parameters for each participant: an accuracy (β) and a variability (σ) for each of the three question types.[8] Analysis for this, and all subsequent experiments, was carried out in R (R Core Team, 2017), and model fits were obtained using the PsiMLE package (Odic et al., 2016). Figure 5 shows the mean of these fitted parameters; error bars represent standard error and stars represent the results of planned t-tests. As mentioned, precision (1 - σ) is plotted instead of variability (σ), to make the figures easier to read (i.e., higher values correspond to better performance).

[8] We put no upper limit on the values β could take, but we did not allow accuracy fits to result in a β<0. Our reasoning was that a β of 0 or below reflects no sensitivity to number information. This choice was mostly aesthetic however, as refitting the models without any restrictions (or with an upper and lower bound on β values) led to the same pattern of results.

3.2 Results

As expected, participants were equally good at estimating the total number of dots in the display whether they knew that the superset would be probed ahead of time (Question First) or not (Dots First); there was no difference in either accuracy (β) or precision (1 - σ) between the Dots First and Question First conditions (β : t_{11} = 0.32, p = .752; σ : t_{11} = 0.32, p = .756) on Total Dots trials (Figure 5). This is the predicted null result for these trials. Because subjects did not know that the Total Dots would be probed on these particular Dots First trials, their performance reveals that they always selected and enumerated the superset of all dots on Dots First trials, in addition to whatever other subgroups they may have attended. This indicates that subjects had no problems with our Dots First trials, and they were able to select and enumerate a set from these displays.

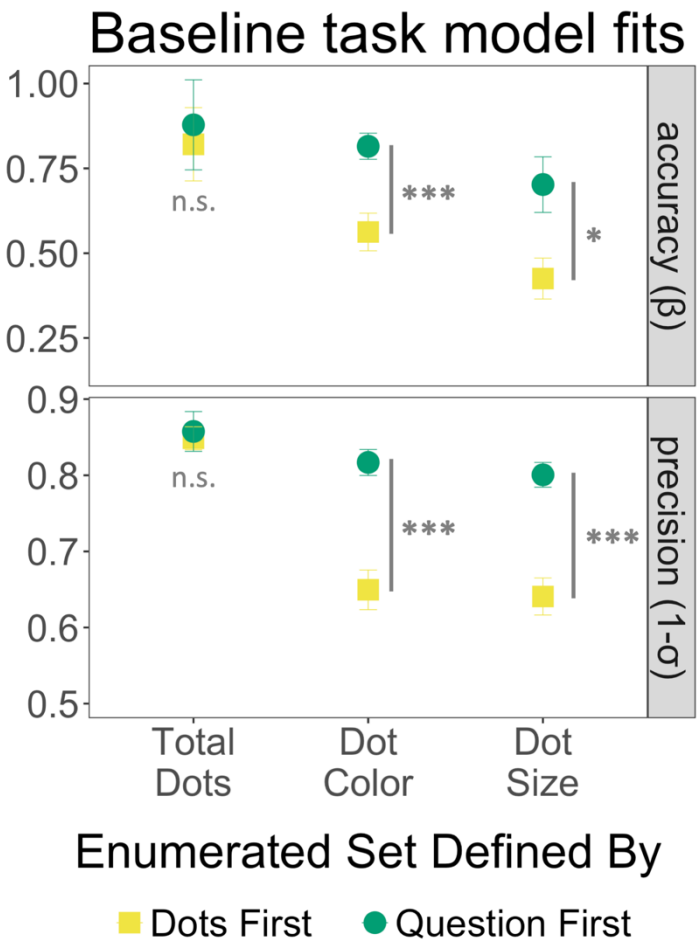

Figure 5: Experiment 1 average parameter fits.

Next, we looked at performance when the to-be-enumerated set was defined by a Dot Color or Dot Size (see Figure 5). Recall that, on these trials, we expected that subjects would be able to select and enumerate these subsets on Question First trials, but that they would

struggle to attend all possible subsets on Dots First trials, resulting in worse accuracy and worse precision relative to Question First trials.

This is indeed what we found. Subjects showed better performance – higher accuracy and higher precision – in the Question First block than in the Dots First block when asked about a subset defined by Dot Color (β: $t_{11} = 4.95$, $p < .001$; σ: $t_{11} = 6.03$, $p < .001$) or by Dot Size (β: $t_{11} = 2.63$, $p < .05$; σ: $t_{11} = 7.24$, $p < .001$). Overall, participants performed somewhat better on Dot Color compared to Dot Size trials. This is likely because color was more visually salient in our stimuli and was thus more easily attended. The important thing is that the difference in performance across these trials shows us what to look for in the upcoming experiments. If a linguistic prompt (e.g., "most of the big dots are blue") leads participants to focus on the set of relevant items, then both accuracy and precision should be higher than when a different prompt (e.g., "there is a big dot that's blue") leads them to focus on the individual items.

3.3 Discussion

Experiment 1 demonstrates how powerful attention to a set can be in terms of improving estimates of that set's cardinality. It also establishes a rough benchmark for the kinds of values we might expect our accuracy and precision parameters to take. Clearly though, participants always know something about a set's cardinality. Strategies like consulting visual working memory or reasoning on the basis of how numerous sets in this experiment tend to be likely helped participants guess, roughly, the cardinality of a set even if they were not attending to it. Importantly, they were able to provide relatively better estimates when they were prompted to attend to the set. In Experiments 2 and 3, we look for the same relative differences. But instead of explicitly instructing participants to attend to a particular group or neglecting to do so, we encourage them to attend to a group or to individuals by evaluating some quantificational statement.

4. Experiment 2: *Most of the* vs. Existential-statements

Armed, from Experiment 1, with a metric for determining the extent to which a cognizer focuses on (and represents) the ensemble or set of items versus the individuals (i.e., both higher accuracy and higher precision of numerical estimates), we now turn to asking how quantificational statements may induce one or the other state.

A natural starting point for testing the idea that quantifiers with a second-order format will encourage attending to groups whereas quantifiers with a first-order format will invite attending to individuals, is with *most*. After all, as mentioned in section 1, proportional *most*'s claim to fame is that it cannot be stated with first-order logic alone and thus requires a second-order specification (Rescher, 1962; Wiggins, 1980; Barwise & Cooper, 1981).[9] There are also empirical reasons, discussed earlier, for thinking that *most*'s representational format highlights cardinalities specifically, and that the default verification strategy for evaluating *most*-statements involves a cardinality comparison (Pietroski et al., 2009; Lidz et al., 2011). If any

[9] Although, see Hackl (2009) for a proposal to cash out *most* without appeal to sets.

quantifier will cause participants to attend to and represent the restrictor set – and thus afford participants good estimates of its cardinality – *most* is our best bet.

Statements with *there is a* and *there are* are sensible linguistic comparisons, as it is at least intuitively plausible that they have first-order formats. To be sure, anything that can be represented with first-order logic can be represented with second-order logic as well. But it would strike us as surprising to learn that their mental representations highlighted groups more than individuals.

We predict that participants should be better at estimating the cardinality of the restrictor set following sentences with *most of the* compared to sentences with *there is a* or *there are*. All of our statements had the form in (30), so the restrictor set was always a size. For this reason, we focused only on size-based questions.

(30) a. Most of the {big, medium, small} dots are {blue, yellow, red}.
b. There is a {big, medium, small} dot that's {blue, yellow, red}.
c. There are {big, medium, small} dots that are {blue, yellow, red}.

If a participant is asked to evaluate the statement *most of the big dots are blue* for example, they should have a good estimate of the number of big dots, but not of the number of medium or small dots. If they are asked to evaluate *there is a big dot that's blue*, on the other hand, they should not have a great estimate of any set's cardinality. This may not be particularly surprising. After all, one might only need to attend to a single big dot to evaluate a sentence like *there is a big dot that's blue*. For this reason, we also analyze the FALSE trials in Experiment 2b. If there are no big blue dots, a participant would have to look at each big dot to successfully evaluate *there is a big dot that's blue*. Even in this case, we expect participants not to attend to the set of big dots, but rather, to attend to individuals (e.g., by cycling through each big dot, and checking its color). As such, we predict that even on FALSE trials in the *there is a* block, participants should still show inferior cardinality estimates compared to the *most of the* block.

4.1 General method

In Experiment 2 (Figure 7), participants first read a quantificational statement, then viewed a dot display and evaluated whether that statement was TRUE or FALSE with respect to the display. After responding, they were asked to give a cardinality estimate of one of the subsets present in the previous display (e.g., "How many big dots were there?"). The dot displays used were generated in the same way as those used in Experiment 1. Participants had as long as they wanted to read each statement. When ready, they pressed 'space' to view the display. Their viewing time was either limited to one second (Experiments 2a&c), as in Experiment 1, or unconstrained (Experiment 2b). They were instructed to press 'J' or 'F' to judge the statement as TRUE or FALSE as quickly and as accurately as possible. They were led to believe the "how many" questions were largely incidental to the task, and in any case, less important than getting the TRUE / FALSE portion correct as quickly as possible.

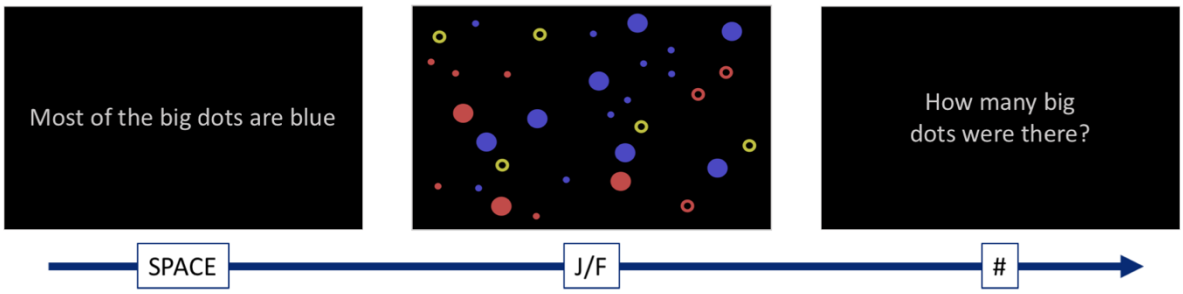

Figure 7: Trial structure of Experiment 2.

The fact that participants were asked about a random subset each time discouraged a strategy of trying to remember each group's cardinality in anticipation of the question (of the 136 trials within each block, only 30 asked about the restrictor set). Moreover, as discussed, Halberda, Sires, & Feigenson (2006) showed that participants can only enumerate three sets in parallel, one of which is the superset. Our displays always contained six subsets along with the superset, meaning that participants could not easily enumerate each subset to prepare for any possible "how many" question.

4.2 Experiment 2a: *Most of the* vs. *There is a*, one-second display time

The stimuli for Experiment 2 were generated in exactly the same way as the stimuli for Experiment 1. However, in this and all future experiments, participants saw 136 trials in each block for a total of 272 trials (compared to 100 total trials in Experiment 1). We used the same types of "how many" questions from Experiment 1 in each block: 30 questions probed the target size (big dots, in the example in Figure 7), 30 probed a distractor size (e.g., small or medium dots), 30 probed the target color (e.g., blue dots), 30 probed a distractor color (e.g., red or yellow dots) and 16 probed the total number of dots. We only report the results following target and distractor size trials here, as the others were included as filler trials (so participants could not guess in advance which set they would be asked about).

4.2.1 Participants

13 University of Maryland undergraduates participated in exchange for course credit. One participant was excluded for evaluating every statement as TRUE. This left us with the desired 12 participants. Of these, 6 started in the *most of the* block, and 6 started in the *there is a* block. Together, they correctly responded to 83.9% of the *most of the*-statements and 96.2% of the *there is a*-statements.

4.2.2 Results

As in Experiment 1, each participant was fitted with an accuracy (β) and precision (σ-1) for each question type (target, distractor) in each block (*most of the*, *there is a*). Recall that target questions probed the restrictor set (e.g., *big* in *most of the big dots are blue*) and

distractor questions probed the complement of the restrictor set (e.g., *small* or *medium* in *most of the big dots are blue*). The average parameter values are shown in Figure 8. Planned t-tests confirm significant differences between target fits for *there is a* and *most of the* along both dimensions (β: $t_{11} = 3.06$, $p < .05$; σ: $t_{11} = 3.86$, $p < .01$). Parameter values for distractor fits are included for reference.

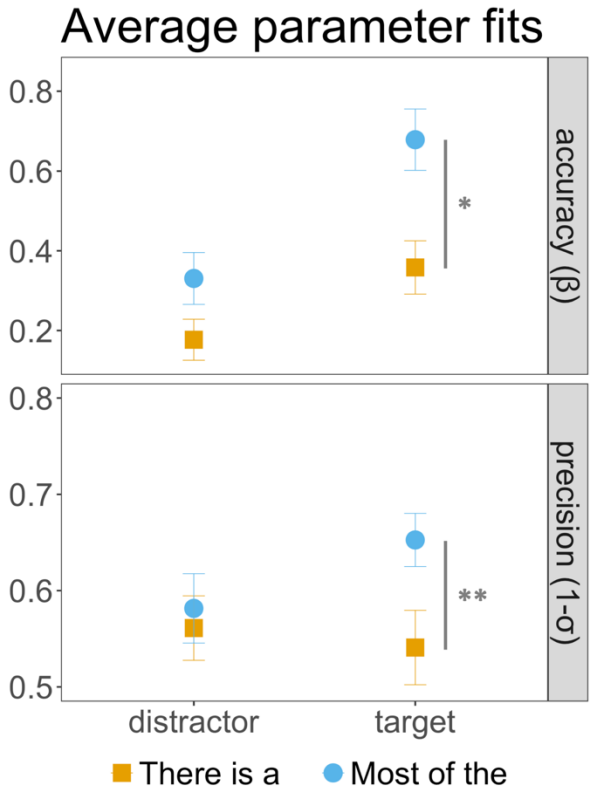

Figure 8: Average parameter fits for all participants from Experiment 2a. Significance bars represent results of planned t-tests.

4.2.3 Discussion

Our findings in Experiment 2a resemble our predictions. In particular, participants knew the cardinality of the target set better following (the manifestly second-order) *most*-statements than following *there is a*-statements. This accords with our intuition that *there is a* directs attention to individuals and shows that evaluating different quantificational statements can have a similar effect on cardinality estimation that showing the display before or after the question had in Experiment 1. Moreover, performance on the distractor trials confirms that participants are not better across the board at estimating cardinality on in the *most of the* block; they are better only when asked about the set denoted by the complement of the restrictor of the statement they evaluated. This is because, presumably, the statement *most of the big dots are blue* highlights the cardinality of the restrictor set (i.e., *big dots*) – that is, *most* is a quantifier with an underlying second-order structure.

4.3 Experiment 2b: *Most of the* vs. *There is a*, unlimited display time

One might be concerned that enforcing a one second viewing time is too constraining in terms of what kinds of strategies are viable for participants to use. For this reason, display times were unconstrained in experiment 2b. That is, the dot displays remained on the screen until participants gave their judgement by pressing 'J' or 'F'. Aside from this difference, Experiment 2b has virtually the same setup as Experiment 2a.[10]

4.3.1 Participants

We sought to double the number of participants in Experiment 2b to look for potential order effects. To this end, 28 University of Maryland undergraduates participated in exchange for course credit. Three participants were excluded for failing to finish both blocks within the allotted hour and one participant was excluded for performing at chance on the TRUE/FALSE portion. This left us with the desired 24 participants, 12 of whom started in the *most of the* block, and 12 of whom in the *there is a* block. They correctly responded to 88.7% of the *most of the*-statements and 97.6% of the *there is a*-statements.

4.3.2 Results

On average, participants viewed the dot displays for 3.6 seconds in the *most of the* block and 2.5 seconds in the *there is a* block. This was a significant difference ($t_{23} = 3.71$, $p = .001$). This makes sense: a difference in strategy is likely to result in a difference in viewing time. However, it's worth mentioning that excluding all trials slower than 3 seconds and re-fitting the models yielded very similar parameter estimates. Moreover, if viewing time alone could account for better cardinality estimates, one would expect performance to improve across the board (i.e., on distractor trials as well). As Figure 9 shows, this was not the case. And, similar concerns about differing viewing times do not apply to Experiment 2a where all trials were limited to one second.

In order to look for potential carryover effects from one block to the next, we separated participants into two groups based on which block of trials they completed first, *most of the* or *there is a*. In both cases, the pattern of results ended up being the same. Recall that if *most*-statements prime alinguistic thoughts about groups (and number), while existential-statements prime alinguistic thoughts about individuals, then we should see higher accuracy and higher precision when asked about the target set (i.e., complement of restrictor) following *most*-statements. These predictions were borne out (Figure 9). Planned t-tests confirmed that participants who started in the *most of the* block were more accurate and more precise on target questions following *most*-statements (β: $t_{11} = 2.34$, $p < .05$; σ: $t_{11} = 3.49$, $p < .01$) and that

[10] There was another minor difference: Some of our participants were given an image of one of the dots in the restrictor set before seeing the display (in both blocks). For example, if the statement for that trial was "most of the big dots are red" they would see a single big red dot below that statement. We thought this might guard against participants misreading the statements and, in general, make the experiment easier. We found no significant difference in performance on accuracy of target trials when these "reminder dots" were present though (*most of the*: $t_{21.1} = 0.9$, $p = .378$; *there is a*: $t_{21.1} = 1.23$, $p = .233$), so we collapsed across the two groups in the analysis presented here and discontinued use of "reminder dots" in all following experiments.

the same was true for participants who completed the *there is a* block first (β: $t_{11} = 4.57$, $p < .001$; σ: $t_{11} = 3.02$, $p < .05$).

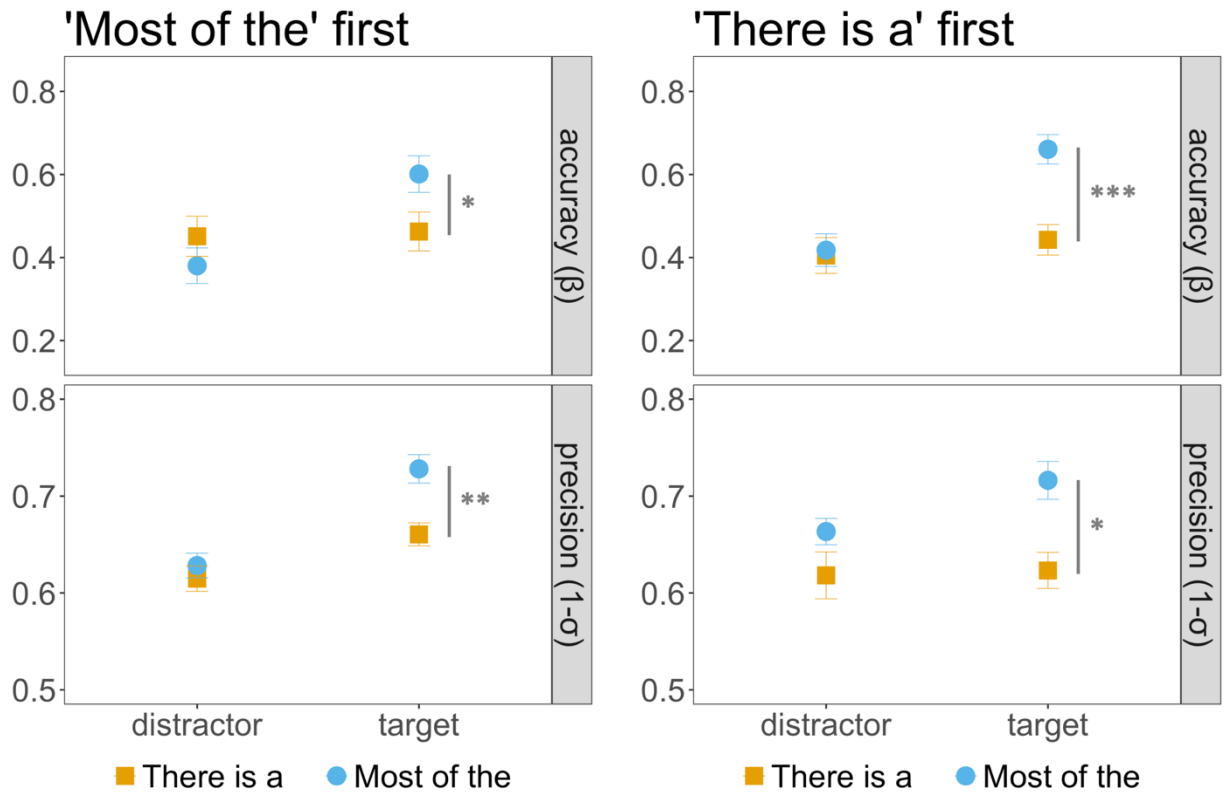

Figure 9: Average parameter fits from Experiment 2b, separated by which block participants completed first. Significance bars represent results of planned t-tests.

As mentioned, it's possible that these results indicate a bias to attend to groups or individuals on the grounds that *there is a*-statements can sometimes be successfully judged TRUE on the basis of just one dot. For this reason, we also considered accuracy on FALSE trials, for example, cases in which the statement was *there is a big dot that's blue* but each big dot was either red or yellow. In such cases, participants presumably would need to consider every single big dot to make the correct judgement. Even so, we find that their performance on the numerical estimation portion of the task remained the same (Figure 10). Participants were more accurate ($t_{23} = 3.22$, $p < .01$) and more precise ($t_{23} = 2.32$, $p < .05$) at estimating the cardinality of the set denoted by the restrictor following *most of the*-statements than following *there is a*-statements.

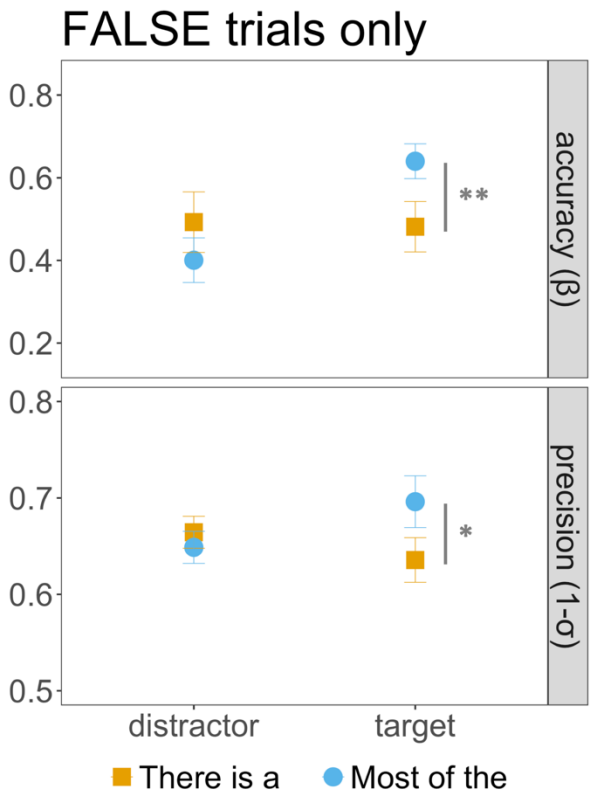

Figure 10: Average parameter fits from experiment 2b for trials in which the answer was FALSE. Significance bars represent results of planned t-tests.

4.3.3 Discussion

Experiment 2b replicates Experiment 2a in showing that participants knew the cardinality of the target set better following *most of the*-statements than following existential-statements. This result makes good on the intuition that *most of the* – but not *there is a* – directs attention to groups. There are three additional takeaways from Experiment 2b. First, giving participants extra time does not wash out the effect. Second, it doesn't seem to matter which condition participants start in, as we see the same pattern of performance in both groups. Third, there seems to be no difference between performance on trials that where the correct answer was TRUE and trials where the correct answer was FALSE.

That we see that same result on the FALSE trials – where the statement was e.g., *there is a big dot that's red* but there were no big red dots – is important. It shows that the effect is not driven merely by looking at more dots in one condition and fewer in the other, as participants likely had to look at each big dot before rendering a judgement. This means that the difference in cardinality estimation ability stems either from (i) participants attending to and representing the dots *as a group* (following *most of the*) verses attending to and representing them *as individuals* (following *there is a*), or (ii) participants using different information during evaluation: for *most of the*-statements, the number of big dots is vitally important and is directly used in providing an answer; for existential-statements, this information is incidental

and there is thus no reason for participants to hold it in memory (other than to do well on our task).

Of course, in the case of the *there is a* FALSE trials, there is another strategy participants might have adopted: Attend to the mentioned color and check whether any of the dots of that color were the correct size. If the statement was *there is a big dot that's red*, for example, they might attend to the set of red dots and respond TRUE if any were big. This would be a reasonable strategy, especially given that in Experiment 1 we saw that color is potentially an easier search cue than size. However, this would predict that participants' estimates of the target color (e.g., red) should be better than their estimate of the distractor colors (yellow and blue, in this example), at least following *there is a*-statements. We did not find this to be the case. In fact, of all the different trial types asking about a color, participants' parameter fits were numerically *lowest* when asked about the target color following *there is a* trials.

4.4 Experiment 2c: *Most of the* vs. *There are*, one-second display time

There is still the worry that the better performance following *most*-statements has nothing to do with the proposed first-/second-order distinction, but with the fact that the restrictor was plural for *most*-statements (*most of the big dot**s***) but singular for existential-statements (*there is a big dot*). After all, we might expect mere mention of *the dots* to direct attention to the dots, taken together[11]

To test the contribution of a plural NP, Experiment 2c compared the now-familiar *most* block with plural existential-statements like *there are big dots that are blue*. As in Experiment 2a, participants only had one second to view the display.

4.4.1 Participants

12 University of Maryland undergraduates participated in exchange for course credit. Of these, 6 started in the *most of the* block, and 6 started in the *there are* block. Together, they correctly responded to 81.8% of the *most of the*-statements and 95.8% of the *there are*-statements.

4.4.2 Results

The average parameter values are shown in Figure 11. Planned t-tests confirm significant differences between target fits for *there are* and *most of the* along both dimensions (β: $t_{11} = 2.36$, $p < .05$; σ: $t_{11} = 6.92$, $p < .001$).

[11] One reason to expect this won't happen might be that attention to groups is triggered by a meaning represented in terms of a genuine relation between them, not their mere mention. By analogy, (1) and (2) from section 1 are still first-order specifications despite the appearance of set/predicate/plural (i.e., capital letter) variables.

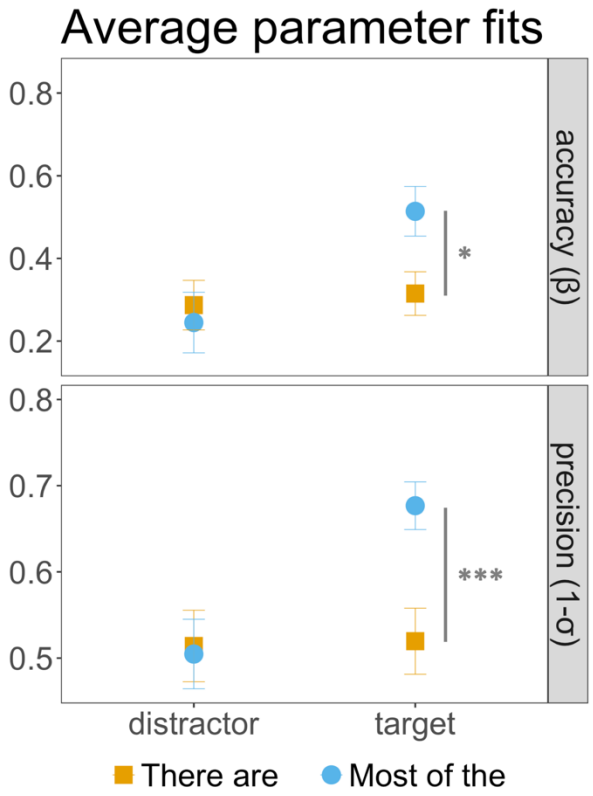

Figure 11: Average parameter fits from all participants in experiment 2c. Significance bars represent results of planned t-tests.

4.4.3 Discussion

As in Experiments 2a&b, we see the same pattern in Experiment 2c. This tells against the idea that the plurality of the NP accounts for *most*'s relative accuracy and precision boosts.

5. Experiment 3: *Each (of the)* vs. *Every* vs. *All (of the)*

Having established the difference between *most of the*- and singular and plural existential-statements, we now turn to a series of experiments pitting the universal quantifiers *each*, *every*, and *all* against each other. In these cases, not only are task, display, and participants held constant, truth-conditions are as well. The only change made between blocks is the choice of quantifier. Experiment 3a compares *each of the* and *all of the*; Experiment 3b compares *each* and *every*; finally, Experiment 3c compares *all of the* and *every*. As before, the general prediction is that evaluating a quantificational statement with a second-order format should yield higher accuracy and higher precision when estimating the cardinality of the restrictor set.

5.1 General method

The setup of Experiment 3 mirrored that of Experiments 1 and 2. Again, participants completed two blocks of 136 trials in which they first read a quantificational statement, then viewed a dot display and evaluated whether the statement was TRUE ('J') or FALSE ('F') with respect to the display. As with Experiment 2b, we returned to free viewing time for the experiments in this section to allow participants more freedom over their choice of strategy. After responding to the TRUE/FALSE question, they were asked to give a cardinality estimate of one of the subsets present in the previous display. The same questions were used (e.g., "How many big dots were there?") and were given in the same distribution as in Experiment 2 (30 target size trials, 30 distractor size trials, 30 target color trials, 30 distractor color trials, 16 total dots trials). Displays, however, contained fewer dots on average, to reduce crowding and make the "how many" questions easier (the average cardinality of the restrictor set was 9 in Experiments 3a-c compared to 11.6 in Experiments 2a-c).[12]

5.2 Experiment 3a: *Each of the* vs. *All of the*

5.2.1 Participants

30 University of Maryland undergraduates participated for course credit. One was excluded from further analysis for scoring below 85% on the TRUE/FALSE portion, and five were excluded for being unable to complete the experiment in the allotted hour. This left us with 24 participants, 12 starting in the *all of the* block and 12 starting in the *each of the* block. These participants correctly evaluated 96.6% of the *all of the*-statements and 96.1% of the *each of the*-statements. On average, participants spent 3.8 seconds viewing the dot displays in the *all of the* block and 4.3 seconds viewing the displays in the *each of the* block, though this difference was not significant ($t_{23} = 1.18$, $p = .247$).

5.2.2 Results

Unlike in Experiment 2a-b, here, we find order effects (see Figure 12). Participants who started in the *each of the* block showed better performance on target questions following *all of the*-statements for both accuracy ($t_{11} = 3.75$, $p < .01$) and precision ($t_{11} = 5.34$, $p < .001$). The effect disappeared for those participants who started in the *all of the* block though. These participants were not more accurate or more precise when guessing the cardinality of the restrictor set after evaluating either statement ($\beta : t_{11} = 0.05$, $p = .957$; $\sigma : t_{11} = 1.11$, $p = .29$). Thus, we only get a within-subjects result when looking at the subjects who completed the *each of the* block before the *all of the* block.

Even for this group – where we find the predicted relative difference – participants' accuracy and precision following *each*-statements is relatively high compared to the trials in which they were asked about a distractor size. We conducted further analyses to test two possible explanations. First, it might be that some participants showed the predicted (worse) performance given *each*, whereas other participants realized from the start to use a group-

[12] Estimates for β in experiments 1 and 2 were somewhat lower than the "best" possible performance of .8. We suspected that this might be due to the displays being slightly crowded, which Im, Zhong, & Halberda (2016) have shown leads to increased underestimation.

based strategy (nothing about the proposed formatting distinction forbids participants from using whatever strategy comes to mind). This would result in a bimodal distribution of parameter fits, with some participants' target fits resembling their distractor fits and other participants' target fits looking more like their target fits following *all*-statements. Statistical tests for multimodality (Hartigan & Hartigan, 1985; Ameijeiras-Alonso et al., 2018) do not offer any evidence of this conclusion however (D = .07, p = .9; Excess mass = .14, p = .86).

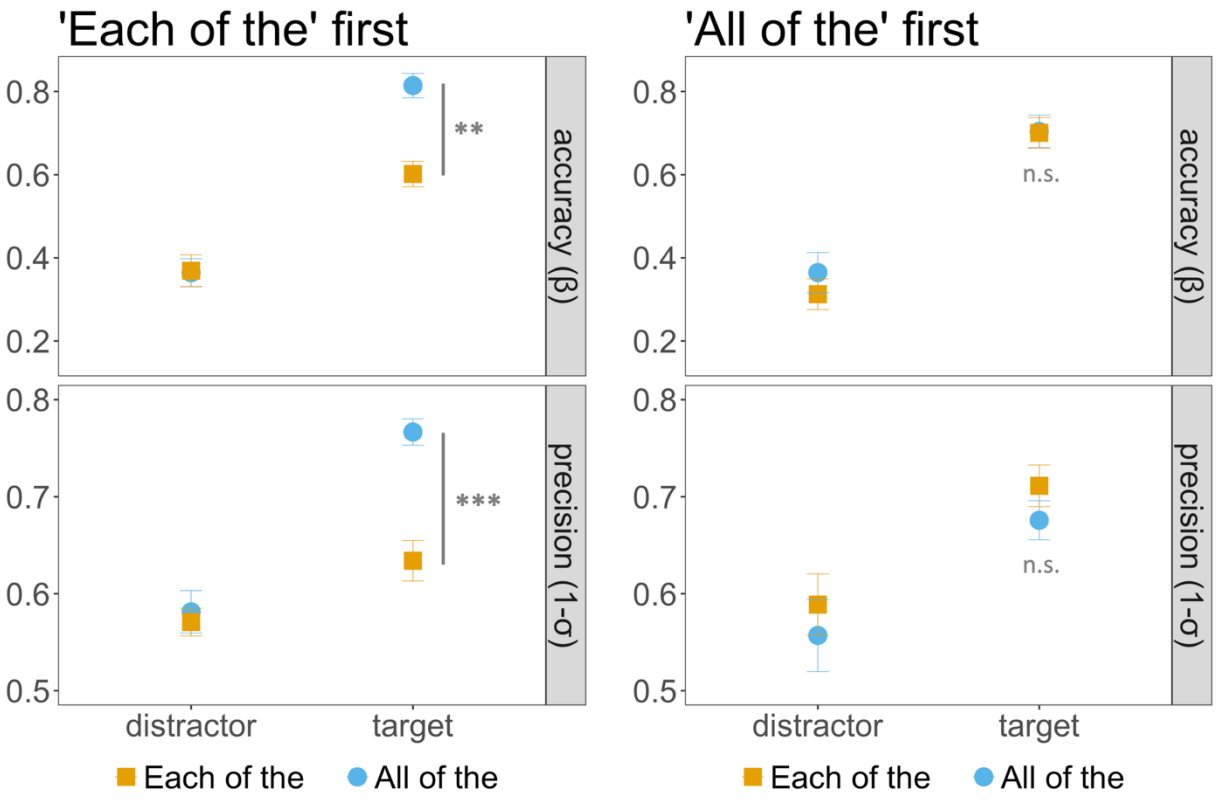

Figure 12: Average parameter fits from experiment 3a, separated by which block participants completed first. Significance bars represent results of planned t-tests.

Alternatively, it might be that our participants attended to and represented the individual dots following *each*-statements but nonetheless got some information about the restrictor set during evaluation through other means. For instance, participants may have adopted a strategy of using viewing time as a proxy for how many relevant dots there were. In general, they may have learned to pay attention to the restrictor set as the experiment went on. Some evidence of this comes from the fact that grouping participants together and fitting only their first trial (on each question type) yields accuracies exactly as bad as we would expect for *each* given the results of our baseline experiment (see Figure 13).[13] Namely, we find an *each* β of around .4 for both target and distractor questions.

[13] We thank Darko Odic for suggesting this analysis.

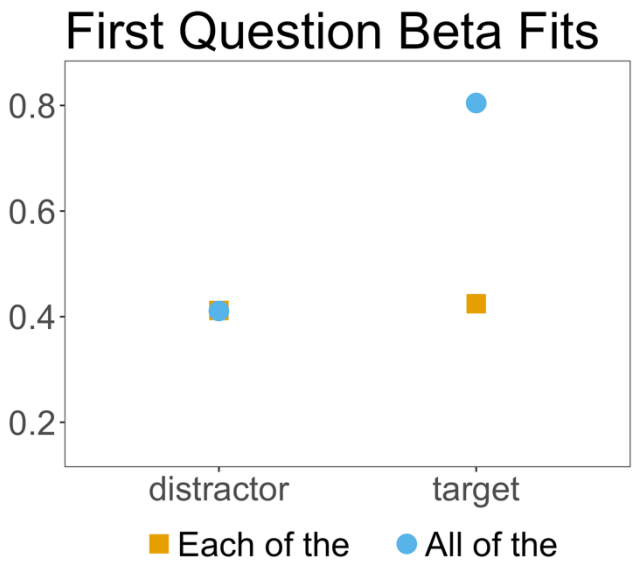

Figure 13: Group accuracy (β) fits on participants' first distractor and target questions of each block of Experiment 3a. Each fit is based on 24 estimates – one from each participant.

5.2.3 Discussion

Why did we only observe the predicted result among participants who started in the *each of the* block? One possibility is that participants who started with *all of the* used a group-based strategy as expected, then retained that same strategy for the *each of the* block. Perhaps participants who started with the *each of the* trials however didn't think to use the group-based strategy until they encountered *all of the* in the second block. Instead, they used a "sub-optimal" individual-based strategy throughout the first block before realizing that focusing on groups would give slightly better estimates (or even if they don't realize this *per se*, the group-based approach feels less effortful). Similar order effects were not found in Experiment 2, which contained statements that did not share truth-conditions (*most of the*… and *there is a*…). Plausibly then, their meanings differed too much for participants to pick up on the fact that some kind of group-based strategy could be retained upon switching from the *most of the* to the *there is a* block.

In any case, the *each*-first participants were better able to estimate the cardinality of the restrictor set following *all*-statements than following *each*-statements, despite their truth-conditional equivalence. The *all*-first participants performed identically on both conditions. This finding is well-explained given a first- / second-order formatting distinction and the ability to turn to a less transparent strategy when the task suggests it.

5.3 Experiment 3b: *Each* vs. *Every*

5.3.1 Participants

30 University of Maryland undergraduates participated in exchange for course credit. Two participants were excluded for mentioning during debriefing that they used an explicit

counting strategy and four were excluded for failing to finish both blocks in the allotted hour. This left us with the desired 24 participants, 12 of whom started in the *each* block, and 12 of whom started in the *every* block. They correctly evaluated 96.7% of the *each*-statements and 97% of the *every*-statements. On average they viewed the dot displays for 3.8 seconds before responding in the *each* block and for 3.4 seconds before responding in the *every* block, though this difference was not significant (t_{23} = 1.2, p = .24).

5.3.2 Results

As before, if *each*-statements invite attending to and representing individuals whereas *every*-statements encourage attending to and representing groups, we should see better accuracy and precision following *every*-statements. This is what we find for participants who started in the *each* block (Figure 14). We observe better accuracy for target questions following *every*-statements (t_{11} = 2.2, p < .05), and likewise better precision for target questions following *every*-statements (t_{11} = 3.36, p < .01).

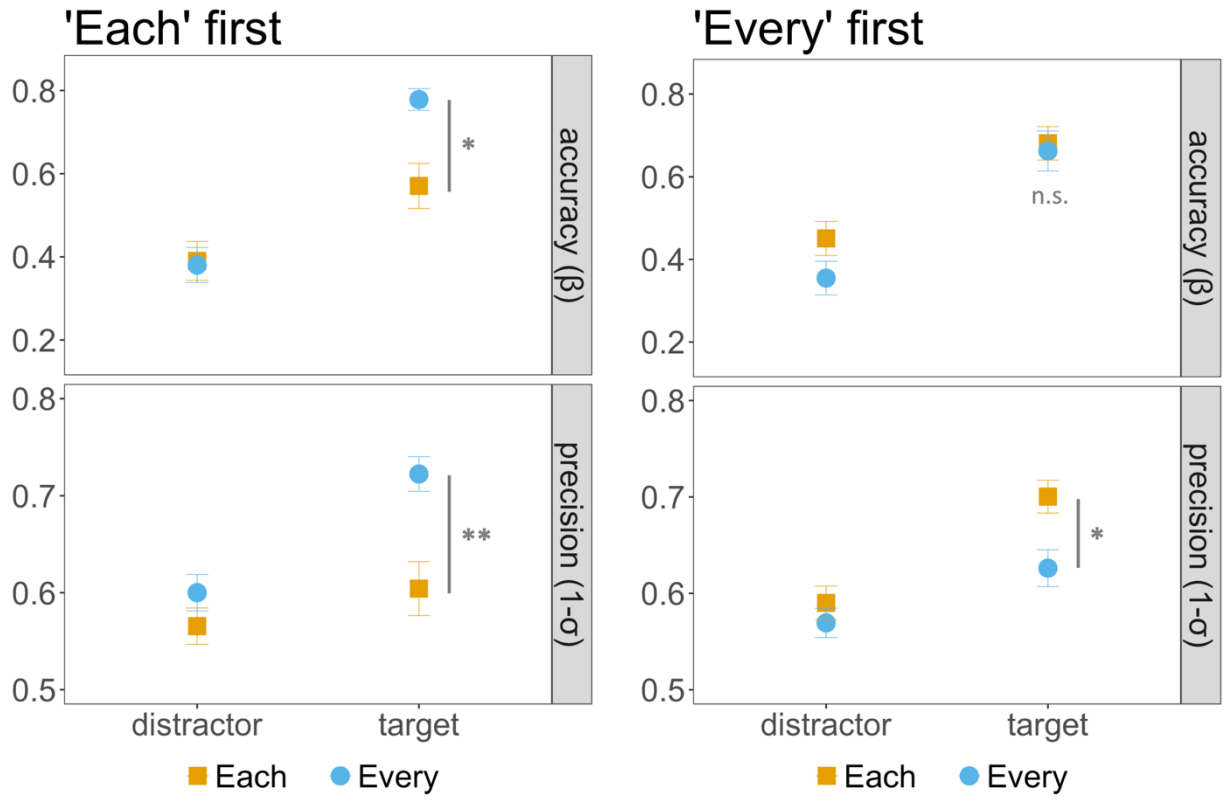

Figure 14: Average parameter fits from experiment 3b, separated by which block participants completed first. Significance bars represent results of planned t-tests.

For participants who started in the *every* block, we find no significant accuracy difference between *each* and *every* (t_{11} = 0.27, p = .795) but do find a significant difference between participants' precision on target questions following *each* and *every*-statements, with a precision boost following *each*-statements (t_{11} = 2.86, p < .05).

As in Experiment 3a, we can obtain group accuracy fits from participants' first trials of each type. Here too, we find that accuracy on the first *each* target trial is where we would expect it to be if participants didn't know the cardinality any better than they did in the baseline experiment (in this case β=.39 for the first *each*-target trial).

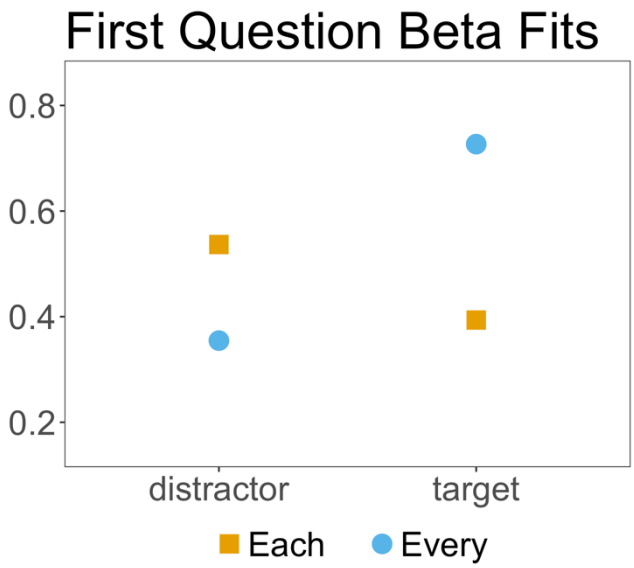

Figure 15: Group accuracy (β) fits on participants' first distractor and target questions of each block of Experiment 3b. Each fit is based on 24 estimates – one from each participant.

5.3.3 Discussion

Experiment 3b shows roughly the same pattern of performance following *each*-statements that we saw in Experiment 3a. In our view, this reflects *each*'s underlying first-order format. The significant precision boost following *each*-statements among the *every*-first participants is unexpected, but without a corresponding boost in accuracy, it is hard to interpret. Such a boost could reflect participants becoming more internally consistent over time or forming expectations about the distributions they were likely to be asked about over the course of the experiment. The first-question accuracy fits suggest that, as in Experiment 3a, participants' initial approach to *each*-statements differed from their approach to *every*-statements. Moreover, Experiment 3b shows that *every* patterns like *all of the* from Experiment 3a. This is consistent with *every* having a second-order format, like *all* and *most*.

5.4 Experiment 3c: *Every* vs. *All of the*

Given the results of Experiments 3a&b, *all* and *every* both invite attending to groups. If a formatting distinction is to blame for our results thus far, we should expect to see no difference in performance between statements with these two quantifiers when they are compared directly.

5.4.1 Participants

Another 28 University of Maryland undergraduates participated for course credit. After excluding four participants who attained less than 85% accuracy on the TRUE/FALSE portion, we were left with the desired 24 participants. Of these, 12 started in the *every* condition and 12 started in the *all of the* condition. They correctly evaluated 95.7% of the *all of the*-statements and 95.6% of the *every*-statements. They viewed displays following *all of the*-statements for an average of 3.9 seconds and displays following *every*-statements for an average of 4.3 seconds, though this difference was not significant (t_{23} = 1.02, p = .318).

5.4.2 Results

Consistent with *every* and *all* both being second-order and thus both biasing attention to groups, the two quantifiers behaved nearly identically (Figure 16). For the participants who started in the *every* block, we observed no significant difference following target questions along either dimension (β: t_{11} = 0.5, p = .963; σ: t_{11} = 1.64, p = .13). We likewise find no significant differences for participants who started in the *all of the* block (β: t_{11} = 0.43, p = .675; σ: t_{11} = 0.44, p = .671).

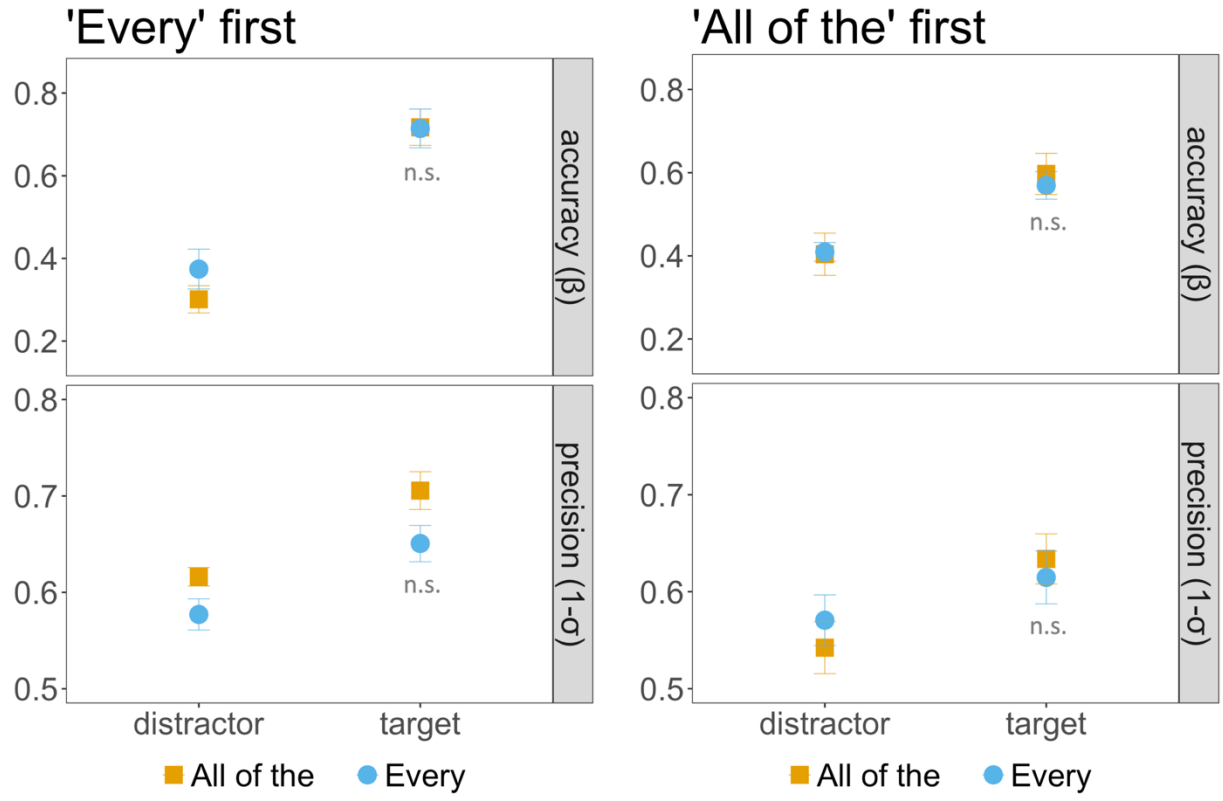

Figure 16: Average parameter fits from experiment 3c, separated by which block participants completed first. Significance bars represent results of planned t-tests.

5.4.3 Discussion

The fact that that participants, on average, performed virtually identically in both conditions not only confirms our predictions, but also helps to rule out a possible alternative explanation for the previous results in which participants shift strategies for other reasons (e.g., the pragmatic strangeness of being asked a statement in the second block of the experiment that is truth-conditionally equivalent to the statement probed in the first). Moreover, we see again that the contribution of the quantifier seems to matter more than the contribution of plurality (and partitivity), which was present in the *all of the* block but not in the *every* block.

6. General discussion

We began with the intuition that some quantifiers' lexical specifications highlight groups, while other quantifiers' specifications highlight individuals. We cashed out this intuition in terms of the distinction between first- and second-order representational formats. In order to look for evidence of this formatting distinction, we probed participants' memory for various sets by asking how well they could estimate those sets' cardinalities. The idea is that a second-order quantifier should bias participants to attend to and represent groups during evaluation, in turn resulting in more accurate and more precise cardinality estimates of whatever group the meaning highlighted (i.e., the restrictor set). A first-order quantifier, on the other hand, should invite attending individuals, in turn resulting in worse accuracy and precision when asked to estimate the cardinality of the set that those individuals constitute.

Our predictions were by and large borne out. Statements with *each*, *there is a*, and *there are* bias attention to individuals, whereas statements with *most of the*, *all (of the)*, and *every* pattern together in biasing attention to groups. This seems to be true irrespective of partitivity or plurality. We take this to be evidence that *each* has a first-order representational format, whereas *all* and *every* have second-order formats, like *most*. This amounts to a claim that the meanings of *each* and *every* tap into different conceptual resources, despite being truth-conditionally equivalent. This naturally raises an interesting acquisition question: what cues would lead a learner to associate a first-order meaning with one pronunciation and a truth-conditionally equivalent second-order meaning with another? We plan to explore this in future work.

For now, we turn to other possible explanations for these results in section 6.1. There is also the objection that our experiments rest on a tendentious characterization of the relationship between meaning and verification. We discuss this objection in section 6.2.

6.1 Alternative explanations

How can we be sure that it was a difference in representational format along the lines we've suggested and not something else that accounts for the observed difference between *each* on the one hand and *every*/*all* on the other? We tried to rule out potential experimental confounds by holding e.g., truth-conditions, participants, display, etc. constant and only changing the relevant quantifier between conditions. Still, a reasonable objection would be that some fact about these quantifiers' *usage* – as opposed to their meaning – could lead to such behavioral differences.

There is a challenge in giving usage-based explanations for these sorts of experiments though, as there is no clear theory of what kinds of usage facts should matter in what kinds of situations. To take one example, Solt (2016) reports (based on a corpus analysis) that English *most* is very rarely used when the percentage referenced is below 60%. Indeed, this apparent fact can have pragmatic effects on the felicity of *most*-claims in everyday discourse. By all accounts, it seems to be a robust and important detail that people know about *most*. But despite this, participants in experimental settings have no problem accepting displays in which 55% of dots are blue as perfectly fine instances of *most of the dots are blue* (see e.g., Pietroski et al., 2009).

For any given usage-based explanation then, the question is why that detail about a lexical item matters in experimental settings whereas other details seemingly don't. An explanation has bite to the extent that it comes with a reasonable linking hypothesis between the usage facts and the experimental data. To take another example, uses of *each* are by far less frequent than the other universals, but it's hard to see how that fact would lead one to attend to individual dots and subsequently not have as good of an estimate of a certain set's cardinality given a statement with *each*.

The distributive / collective asymmetry mentioned earlier is perhaps an example of a linguistic fact which suggests a reasonable linking hypothesis. Without committing to any analysis of why, it can be observed that sentences with *each* are always given distributive interpretations such that the predicate of an *each*-statement applies to the individuals in the domain. On the other hand, *all* is less often – or at least, not always – used in this way. The judgement generally reported in the literature (e.g., Beghelli & Stowell, 1997) is that the ease of accepting the distributive interpretation (in which there were as many singings as there were boys) is *each* > *every* > *all*, as in (31).

(31) a. Each boy sang happy birthday (by himself / # in perfect harmony).
b. Every boy sang happy birthday (by himself / ? in perfect harmony).
c. All the boys sang happy birthday (by themselves / in perfect harmony).

The linking hypothesis between this apparent fact and our experimental task might look as follows: because *each* is always used with distributive predicates, and because it is generally peoples' first choice of universal when expressing a distributive thought, people are naturally biased to think about individuals after hearing *each*.

A problem with this proposal though, is that *every* also largely patterns like *each* with respect to giving rise to distributive interpretations. The collective interpretation of (31b) is not obviously available, so one has to look to infrequent examples like *it took every villager to raise the child* or differences between the *each* and *every* variants of whether-island sentences like (32) in order to begin to see clear non-distributive readings of *every* (Szabolcsi, 2015).

(32) Determine whether {each/every} dragon is dangerous.

The result from our experiments is that *every* and *all* pattern together to the exclusion of *each*. But in terms of being used distributively, *each* and *every* pattern largely together to the exclusion of *all*.

Also problematic for this particular account is the fact that quantifiers like *most* are often used distributively as well (e.g., in (33) the distributive reading is relatively easier).

(33) Most boys sang happy birthday (by themselves > in perfect harmony).

As mentioned, *most* is a manifestly second-order quantifier. And in our experiments, *most* patterns more like *every* and *all* than like *each*. So the fact that a quantifier is often used in distributive contexts likely can't be what gives rise to our results, even if it ends up being true that collective predicates themselves trigger the same sort of attention to groups that we achieved by using certain quantifiers.

In fact, the causation might well go the other way around: *each* can only give rise to distributive interpretations precisely because it's first-order (meaning, roughly, that the predicate in question needs to apply to each individual, not to the collection as a whole). In fact, our tentative way of formalizing *each* as restricted quantification in first-order logic in (1) is the semantics Szabolcsi (2010) gives for the distributive operator responsible for enforcing distributivity in sentences without an overt *each* (see also LaTerza, 2014 who treats *each* as a pronunciation of the distributive operator).

6.2 Revisiting meaning and verification

Though the evidence we presented here to motivate our view involves processing and verification, our claim is not about how sentences with *each, every,* or *all* are processed. It is also importantly not the claim that statements with *each*, *every*, or *all* are always evaluated in some particular way. Rather, it's a claim about the format in which those quantifiers are represented in the minds of speakers. The hypothesis is that the meaning of e.g., *each* is not representationally neutral, and that its actual format eschews representations of sets or collections in favor of representing individuals.

Our evidence in this case rests on the assumption that the format of the meaning has some influence in determining which verification strategy will be used (Lidz et al., 2011). In this sense, we think that meanings provide "default" verification procedures. But such procedures might not be the easiest or most frequent ones. "Default procedure" here means "the strategy that would apply if there were no other factors or influences present".

For example, if a speaker understands *each dog is brown* as an instruction for building a first-order thought, then this understanding provides a natural verification strategy that involves attending to individuals. Of course, speakers are free to resort to a different strategy, depending on meaning-irrelevant factors like what information is most readily available. If they are told that there are five dogs and five brown dogs, for example, speakers will surely not use an individual-based strategy to give a judgement about whether each dog is brown. Having a "default" procedure doesn't preclude participants from reasoning.

Likewise, one might choose to use a different strategy because it is vastly superior or vastly easier to deploy. Suppose an experiment were set up such that every time the correct answer was TRUE, the dots appeared on the top half of the screen and every time the correct answer was FALSE, the dots appeared on the bottom half. This setup makes available a simple and reliable strategy, which participants would likely pick up on quickly: respond "true" if the

dots are on the top and "false" otherwise. If participants used this strategy, our result about knowing the cardinality of the restrictor set better following one quantifier than the other would disappear. But we wouldn't conclude that this change in strategy reflected anything at all about the meanings of the sentences participants were evaluating. Rather, we would conclude it reflects the fact that we invited participants to make use of a different strategy (which in this case happens to be only circumstantially related to the meaning of the expression presented).

7. Summary and General Conclusion

One tradition in Semantics treats expressions as names for things in the world: *rabbit* is the name for a set and *every* is the name for a mind-independent relation. On this view, meanings are representationally neutral in that the formalisms deployed by theorists aren't meant to be related to whatever mental vocabulary humans use to represent the semantic properties of linguistic expressions (see Williams, 2015 for discussion). And for some purposes – e.g., exploring the compositional properties of meanings – it makes sense for theorists to avoid making any assumptions about representational formats.

But while it's true that logically equivalent descriptions of meanings are empirically equivalent for purposes of capturing truth conditions, it does not follow that meanings themselves are representationally neutral. This is a substantive claim about what meanings are. Here we've argued that meanings are not representationally neutral, and specifically that (1) better captures the mental representation of *each rabbit is furry* than (3) (and that the opposite is true for statements with *all* or *every*).

$\lambda R.\lambda F.$
(1) $\forall x{:}x \in R(x \in F)$
(3) $R \subseteq F$

We argued for these claims on the grounds that *each* differs from *every* and *all* (and *most*) with respect to whether it invites attention to and representation of groups. If this is on the right track, it implies that the first-/second-order distinction is psychologically real. And consequently, it offers evidence for the claim that meanings have specific representational formats with detectable psychological consequences (Pietroski et al., 2009; Lidz et al., 2011; Pietroski et al., 2011).

Acknowledgements

This work was supported by the National Science Foundation (Grant #1449815). We thank Alexander Williams for valuable feedback on previous drafts, Ellen Lau and Darko Odic for helpful discussion, and audiences at MACSIM 2017, BUCLD 2018, and the University of Maryland's Cognitive Neuroscience of Language Lab.

Declarations of interest: none

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. Trends in cognitive sciences, 15(3), 122–131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. Psychological science, 19(4), 392–398.

Ameijeiras-Alonso, J., Crujeiras, R. M., & Rodríguez-Casal, A. (2018). Multimode: An R Package for Mode Assessment. arXiv preprint arXiv:1803.00472.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. Psychological science, 12(2), 157–162.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. In Philosophy, language, and artificial intelligence (pp. 241–301). Springer.

Beghelli, F., & Stowell, T. (1997). Distributivity and negation: The syntax of each and every. In Ways of scope taking (pp. 71–107). Springer.

Boolos, G. (1984). To be is to be a value of a variable (or to be some values of some variables). The Journal of Philosophy, 81(8), 430–449.

Burr, D., & Ross, J. (2008). A visual sense of number. Current Biology, 18(6), 425–428. Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. Trends in cognitive sciences, 13(2), 83–91.

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in cognitive sciences*, 13(2), 83-91.

Chen, L. (1982). Topological Structure in Visual Perception. *Science, 218*(4573), 699-700.

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition*, *12*(4), 553-637.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. Vision research, 43(4), 393–404.

Church, A. (1941). The Calculi of Lambda Conversion. Princeton: Princeton University Press.

Dehaene, S. (2011). The number sense: How the mind creates mathematics. OUP USA. Dowty, D. (1987). Collective predicates, distributive predicates and all. In Proceedings of the 3rd escol (pp. 97–115).

Dowty, D. (1987). Collective predicates, distributive predicates and all. In Proceedings of the 3rd ESCOL (pp. 97-115). Ohio: (Eastern States Conference on Linguistics), Ohio State University.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. Trends in cognitive sciences, 8(7), 307–314.

Feiman, R., & Snedeker, J. (2016). The logic in language: How all quantifiers are alike, but each quantifier is different. Cognitive psychology, 87, 29–52.

Gaydos, H. F. (1958). Sensitivity in the judgment of size by finger-span. American Journal of Psychology, 71, 557–562.

Gil, D. (1992). Scopal quantifiers: some universals of lexical effability. Meaning and Grammar: Cross-linguistic Perspectives, 10, 303.

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., & 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., & 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.

Halberda, J., & Odic, D. (2014). The precision and internal confidence of our approximate number thoughts. Evolutionary origins and early development of number processing, 305.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. Psychological science, 17(7), 572–576.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. The annals of Statistics, 13(1), 70-84.

Horty, J. (2007). Frege on Definitions: A Case Study of Semantic Content. Oxford: Oxford University Press.

Im, H. Y., Zhong, S., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research*, *126,* 291-307.

Ioup, G. (1975). Some universals for quantifier scope. Syntax and semantics, 4, 37–58.
Kurtzman, H. S., & MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. Cognition, 48(3), 243–279.

Izard, V. & Dehaene, S. (2008). Calibrating the mental numberline. *Cognition 106*, 1221–1247.

Knowlton, T., Yang, Y. Wong, A., Langfus, J. Pietroski, P., Lidz, J., & Halberda, J. (in prep). Two majority quantifiers in Cantonese bias distinct verification strategies.

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics*, 35, 536–542.

Kurtzman, H. S., & MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, 48(3), 243-279.

Laming, D. R. J. (1997). The measurement of sensation (No. 30). Oxford University Press.

LaTerza, C. (2014). Distributivity and plural anaphora (Doctoral dissertation).

Libertus, M. E., Odic, D., & Halberda, J. (2012). Intuitive sense of number correlates with math scores on college-entrance examination. Acta psychologica, 141(3), 373–379.

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

Link, G. (1983) The Logical Analysis of Plurals and Mass Terms. In Meaning, Use, and Interpretation of Language, ed. R. Bäuerle, C. Schwarze, and A. von Stechow, 303–323. De Gruyter.

Lu, Z.-L., & Dosher, B. (2014). Visual psychophysics: From laboratory to theory. MIT Press.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. MIT Press. Cambridge, Massachusetts.

Odic, D., Im, H. Y., Eisinger, R., Ly, R., & Halberda, J. (2016). Psimle: A maximum-likelihood estimation approach to estimating psychophysical scaling and variability more reliably, efficiently, and flexibly. Behavior research methods, 48(2), 445–462.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 1-9.

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

Pietroski, P., Lidz, J., Hunter, T., Odic, D., & Halberda, J. (2011). Seeing what you mean, mostly. Experiments at the Interfaces, 37, 181.

Pietroski, P. M. (2005). Events and semantic architecture. Oxford University Press.

Pietroski, P. M. (2018). Conjoining meanings: Semantics without truth values. Oxford University Press.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rescher, N. (1962). Plurality Quantification. Journal of Symbolic Logic 27: 373–4.

Scha, Remko. 1984. Distributive, Collective, and Cumulative Quantification. In *Truth, Interpretation, and Information*, ed. Jeroen Groenendijk, Martin Stokhof, and Theo Janssen. Dordrect: Foris.

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

Stevens, S. S. (1964). Concerning the psychophysical power law. Quarterly Journal of Experimental Psychology, 16(4), 383–385.

Stevens, S. S., & Stone, G. (1959). Finger span: Ratio scale, category scale, and JND scale. Journal of Experimental Psychology, 57, 91– 95.

Szabolcsi, A. (2010). Quantification. Cambridge University Press.

Szabolcsi, A. (2015). Varieties of quantification. In N. Riemer (Ed.), The routledge handbook of semantics (chap. 18). Routledge.

Tichý, P. (1969). Intension in terms of Turing machines. *Studia logica*, *24*(1), 7-21.

Tomaszewicz, B. M. (2011). Verification strategies for two majority quantifiers in polish. In Proceedings of sinn und bedeutung (Vol. 15).

Vendler, Z. (1962). Each and every, any and all. Mind, 71(282), 145–160.

Wiggins, D. (1980). 'Most' and 'All': Some Comments on a Familiar Programme, and on the Logical Form of Quantified Sentences. In M. Platts (ed.) Reference, Truth and Reality, Essays on the Philosophy of Language. London: Routledge & Kegan Paul.

Williams, A. (2015). Arguments in syntax and semantics. Cambridge University Press.