Natural Language Semantics
https://doi.org/10.1007/s11050-023-09209-w

# Psycholinguistic evidence for restricted quantification

Tyler Knowlton[1] · Paul Pietroski[2] · Alexander Williams[3] · Justin Halberda[4] · Jeffrey Lidz[5]

Accepted: 31 July 2023

**Abstract**
Quantificational determiners are often said to be devices for expressing relations. For example, the meaning of *every* is standardly described as the inclusion relation, with a sentence like *every frog is green* meaning roughly that the green things include the frogs. Here, we consider an older, non-relational alternative: determiners are tools for creating restricted quantifiers. On this view, determiners specify how many elements of a restricted domain (e.g., the frogs) satisfy a given condition (e.g., being green). One important difference concerns how the determiner treats its two grammatical arguments. On the relational view, the arguments are on a logical par as independent terms that specify the two relata. But on the restricted view, the arguments play distinct logical roles: specifying the limited domain versus supplying an additional condition on domain entities. We present psycholinguistic evidence suggesting that the restricted view better describes what speakers know when they know the meaning of a determiner. In particular, we find that when asked to evaluate sentences of the form *every F is G*, participants mentally group the Fs but not the Gs. Moreover, participants forego representing the group defined by the intersection of F and G. This tells against the idea that speakers understand *every F is G* as implying that the Fs bear relation (e.g., inclusion) to a second group.

**Keywords** Quantification · Quantificational determiners · Psychosemantics · Verification

✉ T. Knowlton

[1] MindCORE, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Philosophy, Rutgers University, New Brunswick, NJ, USA

[3] Department of Linguistics and Department of Philosophy, University of Maryland, College Park, MD, USA

[4] Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

[5] Department of Linguistics, University of Maryland, College Park, MD, USA

Published online: 08 August 2023

## 1 Introduction

Quantificational determiners like *every*, *some*, and *most* are standardly described as devices for expressing second-order relations. In this sense, the meanings of natural language determiners are often said to be special cases of Generalized Quantifiers (Mostowski 1957; Barwise and Cooper 1981; see Westerståhl 2019 for a helpful review). For example, the meanings of (1a) and (2a) might be specified with (1b) and (2b), which are notational variants of (1c) and (2c).

(1) a. Every frog is green.
b. $\{x\text{: Frog}(x)\} \subseteq \{x\text{: Green}(x)\}$
c. Includes($\{x\text{: Green}(x)\}$, $\{x\text{: Frog}(x)\}$)

(2) a. Some frog is green.
b. $\{x\text{: Frog}(x)\} \cap \{x\text{: Green}(x)\} \neq 0$
c. Intersects($\{x\text{: Green}(x)\}$, $\{x\text{: Frog}(x)\}$)

On this view, the determiner combines with two grammatical arguments (e.g., *frog* and *is green*)—arguments that are themselves first-order *predicates*—to form a sentence according to which the extensions of the two arguments are related in the way specified by the determiner. In short, the determiner expresses a relation between two sets. As an analogy, suppose that the transitive verb *admires* in (3a) expresses a first-order relation, as suggested by (3b), where *admires* relates the extensions of *Kermit* and *Grover*.

(3) a. Kermit admires Grover.
b. Admires(Kermit, Grover)

Given this relational understanding of transitive verbs, quantificational determiners can seem like second-order transitive expressions that differ from verbs in that they relate sets of individuals instead of individuals.

Thinking about determiners in this way—as devices for expressing dyadic relations exhibited by sets of domain entities—has provided a useful framework for studying properties of natural language quantification and stating various generalizations; see, e.g., Barwise and Cooper (1981), Higginbotham and May (1981), Keenan and Stavi (1986), and Keenan (2002). One potential limitation of this approach is that prima facie, determiners that express the same relation can still differ in ways that matter for understanding (e.g., *each* versus *every* versus *all*; *most* versus *more than half*; *at most three* versus *fewer than four*). To deal with such differences, as opposed to simply denying that they are semantic, one might retain relational specifications of determiner meanings but supplement them with syntactic diacritics that trigger movement in ways that affect interpretation (e.g., Beghelli 1997; Beghelli and Stowell 1997; Szabolcsi 1997). Alternatively, one might propose that semantically distinct determiners can be truth-theoretically equivalent (across possible worlds) because their meanings reflect psychologically distinct *ways of specifying or thinking about* a common extension (e.g., Geurts and Nouwen 2007; Hackl 2009; Pietroski et al. 2009; Lidz et al. 2011; Knowlton et al. 2021a, 2022a).

A second potential limitation of the relational view is that natural language exploits only a small corner of the space of potential Generalized Quantifiers. This was known from the outset; Barwise and Cooper (1981) note, for example, that quantificational determiners respect logically contingent constraints (e.g., conservativity, discussed below). One might describe this situation by marking the occupied regions of the broad space permitted by the general and highly expressive system. Alternatively, one might begin with a more restrictive system, in which the unattested meanings cannot be expressed (e.g., Pietroski 2005, 2018; Ben-Yami 2009; though see also Ben-Yami 2012; Westerståhl 2012). The choice is similar to one in syntactic theory: either formulate grammars for natural languages using a powerful system, such as the class of Type 0 or Type 1 Grammars (in the sense of Chomsky 1956, 1959), thus allowing for grammars of unattested kinds; or begin with a more restrictive system—perhaps one that allows for "mildly context-sensitive grammars" (Joshi 1985; Joshi et al. 1990; Steedman 2000; Stabler 2001, 2013)—whose range more closely aligns with what is naturally observed. Similar choices arise in the study of phonology (Heinz and Idsardi 2011, 2013). Here, we press a similar point for semantics, and take the view that a more restrictive system provides a better explanation of how natural language determiners are understood (i.e., mentally represented) by language users (relatedly, see Icard and Moss 2022).

The view that determiners express relations has been dominant. Inspired by Frege (1879, 1884, 1892) and Russell (1905), it was brought to linguistics via Lewis (1970), Montague (1973), and others. But prior to Frege, it had seemed obvious that the classical quantifiers are devices for ascribing properties to some quantity of things in a *restricted* domain. As Hodges (2012, p. 247) puts it in his translation of an early sixth century commentary on Aristotle: "Determiners … combine with the subject terms and indicate how the predicate relates to the number of individuals under the subject; … *Every man is an animal* signifies that *animal* holds of all individuals falling under *man*." Higginbotham and May (1981, p. 54) put the same point as follows: the noun (phrase) with which a determiner combines can be thought of as "restrict[ing] the domain over which the variable bound by [the quantifier] ranges." And as Lepore and Ludwig (2007, p. 61) similarly say: the internal argument of *all* in the sentence *all men are mortal* "functions as if it were a variable restricted to taking on as values only men." A similar idea is pursued in Lasersohn (2021), which advocates interpreting nouns not as predicates, but as themselves restricted variables.

According to this non-relational "restricted quantification" view, *every frog* signals two important features about the proposition expressed with the sentence in (1a): the generalization is universal (as opposed to proportional or existential), and the generalization is about the frogs (thus excluding as irrelevant any other green things). In which case, *every* is not a device for relating two independently specified sets that correspond to the determiner's two syntactic arguments (*frog* and *is green*). Rather, *every* is device for applying a predicate (supplied by its external argument) to a restricted domain (defined by its internal argument).

Following Westerståhl (2019), the difference between relational and restricted quantification can be formalized as in (4), where '$X \upharpoonright$' is understood as "relativized to X".

(4) a. $\mathrm{EVERY}_{X}[\mathrm{FROG}_{X}, \mathrm{GREEN}_{X}]$
b. $\mathrm{FROG} \upharpoonright \mathrm{EVERY}_{X}[\mathrm{GREEN}_{X}]$

In the overtly relational (4a), the two syntactic arguments of the quantifier (*frog* and *is green*) are logically on a par. Both the internal and external argument supply terms in a relation ('FROG' and 'GREEN'), and the determiner specifies which relation ('EVERY'). On the restricted (4b), the arguments serve different logical roles. The internal argument restricts the domain of quantification, and the external argument supplies a further condition that some quantity of this restricted domain needs to meet. The determiner specifies the quantity.

The relational (4a) might be elaborated as in (1b-c) above, or in various other ways (e.g., by relating the cardinality of the frogs to the cardinality of the green things that are frogs; see Sect. 2). But however one cashes out 'EVERY' in (4a), the idea is that it relates the two pluralities supplied by its two arguments. On the other hand, the asymmetry inherent in the restricted (4b) can be further highlighted by cashing out 'EVERY' as in (5a), where the iota expression '$\iota$X:Frogs(X)' introduces the plural group "the frogs" and the quantifier specifies how many members of that group the predicate applies to in a first-order way.

(5) a. $\iota\mathrm{X{:}Frogs(X)} \rceil \forall\mathrm{x[Green(x)]}$

b. $\approx$ Relative to the frogs, every thing$_{\mathrm{x}}$ is such that it$_{\mathrm{x}}$ is green

Unlike the relational (1b), repeated below, the restricted (5a) has no part that represents the green things—at least not as such—though both (1b) and (5a) have parts that represent the frogs ('{x: Frog(x)}' and '$\iota$X:Frogs(X)').[1]

(1) b. $\{\mathrm{x{:}\ Frog(x)}\} \subseteq \{\mathrm{x{:}\ Green(x)}\}$

The particular unpacking of 'EVERY' in (5a) amounts to the additional claim that the restriction is the only part of the meaning that introduces a plurality. But although it highlights the asymmetry between the logical role of the determiner's two syntactic arguments, this additional claim is not entailed by the restricted view (see discussion of *most* in Sect. 4). In what follows, we concern ourselves primarily with the claim that quantifiers like *every* have restricted meanings, though the data presented below could also be used to argue for the more specific claim that *every*'s meaning introduces only a single group, along the lines of (5).

There are theoretical reasons for preferring the restricted view. Notably, it offers a simple account of the generalization that all determiners are "conservative" (e.g., Pietroski 2018; Westerståhl 2019; Knowlton et al. 2021b; Ludlow and Zivanović 2022). Essentially, the conservativity generalization is that all determiners have the following property: duplicating their internal argument in their external argument results in two sentences that are mutually-entailing (Barwise and Cooper 1981; Higginbotham and May 1981; Keenan and Stavi 1986; for recent debates surrounding the right characterization of the constraint, see e.g., Zuber and Keenan 2019; Pasternak

[1] The expression '$\iota$X:Frogs(X)' in (5a), glossed as *the frogs*, is shorthand for '$\iota\mathrm{X}[\forall\mathrm{x}[\mathrm{X(x)} \equiv \mathrm{Frog(x)}]]$': the things$_{\mathrm{X}}$ such that for each thing$_{\mathrm{x}}$, it$_{\mathrm{x}}$ is one of them$_{\mathrm{X}}$ iff it$_{\mathrm{x}}$ is a frog. In using the iota operator here, we do not commit ourselves to the view that sentences like *every frog is green* presuppose the existence of some contextually relevant frogs. The expression '$\iota\mathrm{X}[\forall\mathrm{x}[\mathrm{X(x)} \equiv \mathrm{Frog(x)}]]$' indicates the frogs; for present purposes, we remain agnostic about the nature of this plurality and any presuppositional commitments. In contrast, expressions like 'Frog(x)' and 'Green(x)' do not, on their own, imply the existence of a particular group (the frogs or the green things).

and Sauerland 2022). For instance, *every/some/no fish swim* is true if and only if *every/some/no fish are fish that swim* is true. Learnability results suggest that this is not a historical accident, but a universal that reflects a deep fact about the language faculty (Hunter and Lidz 2013; Steinert-Threlkeld and Szymanik 2019; Knowlton et al. 2022b; *cf*. Spenader and de Villiers 2019; Ramotowska 2022).

In this context, the crucial point in favor of the restricted view is that typologically unattested determiner meanings that are easily specified in terms of *non-conservative relations* cannot be specified as restricted quantifiers. For example, one can imagine a language with a determiner *equi* such that *equi frogs are green* means "the frogs and the green things are equinumerous". This unattested determiner meaning is easily stated in relational terms: $|\{x\text{: Frog}(x)\}| = |\{x\text{: Green}(x)\}|$. Some way of filtering out troublesome relations like 'A = B' is then needed if the relational view is to be retained (Keenan and Stavi 1986; Romoli 2015). But no filter is needed If quantificational determiners are understood as expressions that combine with their syntactic complement to create a restricted quantifier, as Westerståhl (2019) shows. Intuitively, the imagined meaning of *equi* cannot be specified by saying how the predicate *green* applies to the frogs. Conservativity is thus a spandrel of the restricted view. And since conservativity is perhaps the most robust and renowned semantic universal (see von Fintel and Matthewson 2008 for review), explaining it as a consequence of semantic theory is an important benefit.

Setting aside conservativity and other theoretical considerations, our aim here is to offer some initial psychological evidence for preferring the restricted view. In particular, the restricted view posits a difference in the logical role of the two syntactic arguments of the determiner, the internal (nominal) argument and the external (clausal) argument. The relational view does not. Taking this as a psychological hypothesis—a hypothesis about the representations that language users instantiate in understanding a sentence with a quantificational determiner—we expect some psychological reflection of this logical asymmetry on the restricted view, but not on the relational view. The studies reported below find exactly such an asymmetry. In short, when speakers of English are presented with a sentence of the form 'Every A B' in a situation where all the relevant correspondents of 'A' and 'B' are easily seen, these speakers mentally represent the correspondents of 'A' as a group *without* likewise representing the correspondents of 'B' as a group. Even more strikingly, speakers avoid representing the correspondents of the conjunction 'A & B' as a group. This tells against the idea that a sentence of the form 'Every A B' means that the relevant correspondents of 'A', taken together, are suitably related to the relevant correspondents of 'B', taken together. Likewise, it tells against an amended version of the relational view where the correspondents of 'A' are related to a group defined by intersecting both arguments, 'A ∩ B'.

## 2 Psycholinguistic predictions of formally distinct meanings

As discussed in Sect. 1, on standard views, the sentence in (6a) has a relational logical form like (6b), which can be glossed 'the big circles are among the blue things'.

(6) a. Every big circle is blue.

b. $\{x\text{: x is a big circle}\} \subseteq \{x\text{: x is blue}\}$
c. $\{x\text{: x is a big circle}\} = \{x\text{: x is a big circle}\} \cap \{x\text{: x is blue}\}$
d. $\iota X{:}\mathrm{BigCircles}(X) \rceil \forall x[\mathrm{Blue}(x)]$

A less common, though still reasonable, way to cash out the relational view is in (6c): 'the big circles are identical to the big circles that are blue'. This will be true if and only if every big circle is blue. Romoli (2015) offers some reasons for thinking the logical form in (6c) is preferable to the one in (6b) (having to do with how traces of movement are interpreted and accounting for the conservativity generalization discussed in Sect. 1). On this view, *every* still expresses a relation between two sets, just a different relation than is commonly thought, and a different second set than is commonly thought.

Our interest here is in comparing the two versions of the relational view in (6b) and (6c) to a restricted specification like (6d): 'the big circles are such that being blue is a feature of every one of them'. As noted, the important difference here is in how the two grammatical arguments are treated. The relational (6b) and (6c) treat both syntactic arguments—*is blue* and *big circle* in (6b) and *is big circle that is blue* and *big circle* in (6c)—as logically on a par. But those same syntactic arguments play different logical roles in (6d). Additionally, while (6d) represents the big circles as such, the blue things are not represented; 'Blue(x)' is a predicate that applies to every one of the big circles (though, to repeat, that the external argument is treated in a first-order way in (6d) is not required for the specification to be restricted).

Viewed as a theorists' way of specifying the truth-conditional content of (6a), the specifications in (6b-d) are equivalent. But they can instead be viewed as distinct hypotheses about the nature of the mental representation—the "psycho-logical form"—that serves as the meaning of (6a). Viewed this way, the formal distinctions between (6b-d) may have cognitive import. For example, a mind that lacked the mental analogue of the symbol '∩' but possessed a mental analogue of the symbol '⊆' could have (6b) as the meaning of (6a) but could not have (6c). Likewise, a mind that completely lacked any ability to relate one set to another could nonetheless represent (6d). Speakers with both minds would agree about the truth-conditions of (6a), but nonetheless be tokening different thoughts in understanding the sentence (cp. Church 1941 on the distinction between *functions in extension* and *functions in intension*).

To derive behavioral predictions from these formally distinct representations, we adopt the Interface Transparency Thesis of Lidz et al. (2011): verification procedures employed in understanding a declarative sentence are biased toward algorithms that directly compute the relations and operators expressed by the semantic representation that gives the meaning of that sentence, so that this representation is reflected transparently in the cognitive procedure of evaluation. This linking hypothesis does not imply that meanings *are* verification strategies, or that a given verification strategy will *always* be used to evaluate a given meaning. Instead, the idea is that the details of the representation—which relations and operations it explicitly encodes—will carry some detectable influence on which cognitive strategy is used to evaluate that representation. The precise strength of that influence will vary, but the thought is that it will be enough to explain otherwise puzzling behavior, like participants adopting a non-optimal strategy with respect to the task. On analogy, consider a linking hypothesis that is more frequently used to underwrite linguistic claims: acceptability is evidence of grammaticality. Obviously, grammaticality isn't the sole factor in determining acceptability. Sentences with multiple center-embeddings (*the dog the mouse the cat chased scared barked*) strike even trained linguists as unacceptable, though presumably for reasons other than ungrammaticality (as argued by Miller and Chomsky 1963). Still, grammaticality carries a detectable influence on acceptability. And in the same way, the Interface Transparency Thesis maintains that representational details of a meaning carry a detectable influence on the strategy used for verifying that meaning in a given situation.

In practice, this influence of the meaning can be detected in carefully controlled contexts; namely, when task factors are held equal and there is no obvious reason to use an alternative strategy. To take one example of this linking hypothesis in action, Pietroski et al. (2009) ask whether the meaning of *most* is specified in terms of one-to-one correspondence or comparison of cardinalities. In their task, participants were asked to evaluate *most of the dots are yellow* with respect to displays of blue and yellow dots that appeared on-screen for 200 milliseconds. On most display types tested, participants showed signs of using a cardinality-based strategy: their performance was characteristic of the Approximate Number System (e.g., Feigenson et al. 2004; Dehaene 2011). Importantly, participants in their experiment verified the sentences using cardinality estimations despite the fact that a strategy of one-to-one correspondence was not only available but was a superior alternative. That is, participants were more accurate and faster to respond if they were shown the same displays and asked whether there were leftover yellow dots. This correspondence-based "find the leftover" strategy would have given the same answer: most of the dots are yellow just in case the yellow dots correspond one-to-one with the blue dots with at least one remainder.[2] The fact that participants eschewed this "find the leftover" strategy in favor of an inferior cardinality-based alternative calls out for explanation. Given that there is no other reason to prefer the cardinality-based strategy, the meaning of the expression is likely to blame. So, Pietroski et al. reasoned, *most* is specified in terms of cardinality, not in terms of correspondence.

That said, this finding does not predict that a cardinality-based strategy will always be used to verify sentences with *most*. Sometimes, non-linguistic pressures in favor of one particular strategy outweigh the bias to use a strategy transparent to the meaning being evaluated (just as sometimes, cognitive pressures outweigh grammaticality in determining acceptability). Pietroski et al. (2009) present one such case. In their "column pairs sorted" condition, blue and yellow dots were lined up next to each other. Here, participants used line length—a computation far more accurate than cardinality comparison—as a proxy for whether the sentence *most of the dots are yellow* was true or false: if the yellow line is longer, respond "true". Participants were (wisely) resorting to a superior strategy given the task at hand. But since this change in strategy is explainable without appealing to properties of the meaning under evaluation, Pietroski et al. concluded nothing about the meaning of *most*.

[2] Unless one thinks that *most* means something like "significantly more than half" for some pragmatically determined threshold value; see Denić and Szymanik (2022) for discussion of this possibility. In which case, one might prefer to replace "with at least one remainder" with "with a significant remainder, given some pragmatically determined threshold".

Returning to the current case-study of *every*, the question at issue is not whether the determiner meaning is specified in terms of cardinality or correspondence, but whether it treats its syntactic arguments symmetrically (as on the relational view) or asymmetrically (as on the restricted view). Given Interface Transparency, the candidate psycho-logical forms (6b-d) give rise to different predictions about how participants will evaluate the sentence (6a) in suitably controlled settings. Given the relational specification in (6b), all else being equal, we should expect participants to evaluate (6a) by representing the big circles, independently representing the blue things, and relating the two groups. Given the relational (6c), all else being equal, we should expect participants to represent the big circles, independently represent the big blue circles (perhaps directly or perhaps by independently representing the big circles, the blue things, and computing the intersection of these sets), and relate the two groups. But given the restricted (6d), we should expect participants to treat the big circles and the blue things differently (perhaps opting to represent the big circles, and then deciding if they are distributively blue without bothering to independently represent the blue things or the big blue things as such). That is, on the restricted view, we should expect the logical asymmetry to give rise to a corresponding psychological asymmetry. The remainder of this section details the kind of evidence we seek in asking whether there is such an asymmetry.

Previous psycholinguistic work provides reason for thinking that phrases like *every big circle* or *all big circles* lead participants to mentally group the quantifier's internal argument (the big circles, in this case), even when combined with distributive predicates like *be blue* (Knowlton 2021; Knowlton et al. 2022a).[3] To determine whether there is evidence of an asymmetry, then, we need to establish whether participants likewise mentally group the quantifier's second argument (the blue things, in this case). Such a cognitive strategy is certainly available to participants. Previous work in psychophysics shows that adults and children are perfectly able to represent up to three psychological groups in parallel (Halberda et al. 2006; Zosh et al. 2011). Given that there is no reason to avoid representing two groups, then, the question is whether participants naturally will do so when evaluating sentences like *every big circle is blue*.

Fortunately, there are many behavioral signatures of participants having represented some individual objects as a psychological group, including the cardinality, the average hue, and the center of mass of visually-presented objects (e.g., Ariely 2001; Haberman and Whitney 2012; Whitney and Leib 2018). The experiments below use cardinality knowledge as a proxy for whether participants mentally represented the extension of a given argument as a group during sentence evaluation. In using this measure as a proxy, we do not mean to suggest that cardinality knowledge is a necessary prerequisite for group representation. In principle, any "summary statistic" properties could be used as evidence. But given that the prior work on how

[3] Given a manifestly collective predicate, like *all students gathered in the hall*, it might be less surprising for participants to group the students. After all, an individual student cannot gather. But at least as it is used in this experimental context, *be blue* is a distributive predicate in the sense that all the circles are blue if and only if each individual circle is blue. Even so, the studies cited above find that participants psychologically group the big circles upon encountering *every big circle* or *all big circles* to a greater extent than when encountering *each big circle* in the very same experimental context.

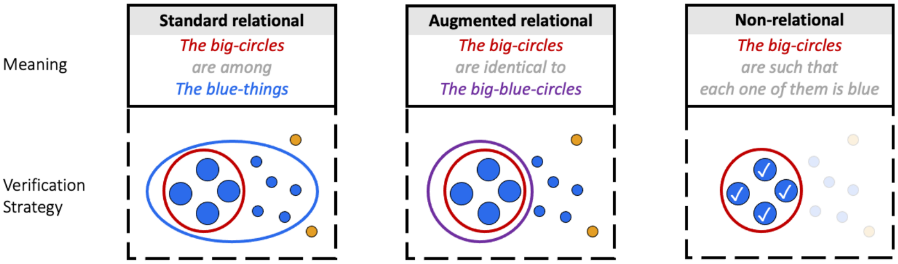

**Fig. 1** A schematic depiction of the predicted verification strategies associated with each hypothesized meaning of the sentence *every big circle is blue*. Both the standard relational view (i.e., (6b)) and the augmented relational view (i.e., (6c)) call for somehow relating two independent samples from the domain. The non-relational view (and in particular (6d)) predicts participants will only form a single grouping of domain entities (Color figure online)

many groups can be extracted in parallel relied on cardinality (Halberda et al. 2006; Zosh et al. 2011), we opt to also rely on cardinality here.

The specific predictions, then, are as follows. If participants are asked to evaluate a sentence like *every big circle is blue*, the relational (6b) predicts that they will represent both *the big circles* and *the blue things*, and consequently encode the cardinality of both groups. The relational (6c) predicts that they will represent *the big circles* and *the big blue circles*, and consequently encode the cardinality of both groups. As noted above, there is good reason for thinking that a strategy involving encoding and comparing two groups' (and even three groups') cardinalities is cognitively available. But in contrast to those relational views, the restricted (6d) predicts that participants will treat the extensions of its two grammatical arguments differently. In principle, that could amount to grouping the external argument to the exclusion of the internal argument. But since past work finds that evaluating these sorts of sentences lead participants to mentally group the internal argument, finding here that they only represent—and thus encode the cardinality of—*the big circles* would provide support to the restricted view. Differences between the predicted verification strategies associated with relational and non-relational meanings is given schematically in Fig. 1.

## 3 Experiments

### 3.1 Experiment 1: *every big circle is blue*

The five experiments reported here follow the same basic structure. Participants first saw a sentence like *every big circle is blue* and were then shown displays consisting of different sized (big, medium, and small) and different colored (blue, red, and yellow) circles (see Fig. 2). This display remained on the screen for 1 second, after which participants pressed 'J' or 'F' to judge the sentence as true or false relative to the picture. Participants were subsequently asked to recall the cardinality of some group of circles. Some follow-up questions probed the determiner's internal argument (e.g., "how many *big* circles were there?"); others probed its external argument (e.g., "how

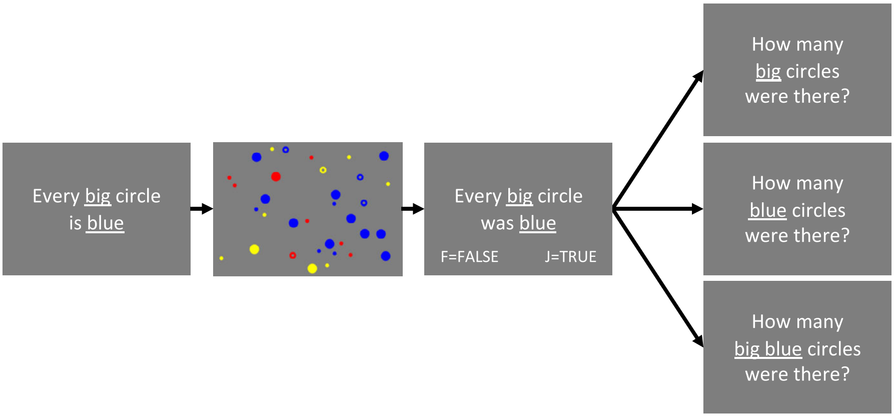
Every big circle is blue
Every big circle was blue
F=FALSE
J=TRUE
How many big circles were there?
How many blue circles were there?
How many big blue circles were there?

**Fig. 2** Trial structure of the experiments. Participants were initially asked to verify a quantificational sentence with respect to a dot-display. They were then asked to recall the cardinality of a particular group: the group defined by the internal argument (e.g., "how many big circles were there?"), the group defined by the external argument (e.g., "how many blue circles were there?"), or the group defined by the conjunction of both (e.g., "how many big blue circles were there?") (Color figure online)

many *blue* circles were there?"); others probed the conjunction of both arguments (e.g., "how many *big blue* circles were there?").

To control for any differences stemming from visual salience, participants' responses were compared against a baseline task. This task used the exact same stimuli but asked participants the cardinality question before displaying the image. For example, they might be asked "how many blue circles are there?", see the image for 1 second, then offer their response. On this baseline task, participants should perform as well as their visual systems will allow. Cardinality estimates following sentence verification can then be compared against this optimal-performance baseline. As discussed above, the relational view predicts two groups to be represented, and therefore predicts participants to do as well following sentence verification as on baseline for at least two out of the three cardinality questions. On the other hand, the restricted view predicts an asymmetry between the quantifier's two arguments. In particular, participants are expected to perform similarly to the optimal-performance baseline only when asked about the cardinality of the group defined by the quantifier's internal argument (i.e., size questions).

### 3.1.1 Method

**Participants** Fifty-three participants were recruited online using Amazon Mechanical Turk. All passed an English screener and gave informed consent prior to participating in the experiment. Two participants were excluded from further analysis for performing at chance or below on the true/false portion of the verification task. Three participants were excluded for taking longer than 5 seconds, on average, to respond to follow-up cardinality questions. This left 48 participants (which served as our target n for all subsequent experiments).

**Materials** Sentences in the verification task were of the form “Every {big/medium/small} circle is {blue/red/yellow}”. Half were true with respect to the display and half were false. The displays themselves consisted of a grey background with circles that were either blue, red, or yellow and were either big, medium, or small (see Fig. 2). Medium circles had holes in the middle to make them more easily distinguishable from the other two sizes (Chen 1982, 2005). Each display contained up to 48 circles from the nine possible size/color combinations. Each combination contained up to 10 circles, and in any given display there were six size/color combinations that contained at least one circle. False trials had between one and three disconfirming circles (e.g., if the sentence was “every big circle is blue” and the trial was false, then there would be between one and three big circles that were yellow or red).

Follow-up cardinality questions either probed the target size (big circles in our running example), the target color (e.g., blue circles), the target size/color combination (e.g., big blue circles). Filler trials probing a distractor group (sizes, colors, or size/color combinations not mentioned in the initial sentence) were also included.

**Procedure** Participants were first given a brief set of instructions, during which circle sizes and colors were labeled: “In this task, you’ll look at pictures of circles. There will be different types…”. After being introduced to the circles, participants were given three practice trials in which an initial question was displayed (e.g., “how many big red circles are there?”), pressing the spacebar revealed an image for 1.5 seconds, and then the question was displayed again (e.g., “how many big red circles were there?”). After completing the three practice trials, participants were told the real test would begin.

Participants then completed 15 trials of the baseline task in which they were presented with a question that probed a size, color, or size/color combination (e.g., “how many big circles are there?”). After pressing spacebar, the image was displayed for 1 second, followed by a reiteration of the question (e.g., “how many big circles were there?”). Participants typed their response and continued to the next trial. Following this task, participants were instructed that they performed well and in the next half of the task, the initial question would be replaced with a sentence for them to evaluate. They then completed 18 trials of the sentence verification task. They were presented with a sentence (e.g., “every big circle is blue”) and were then shown the corresponding display for 1 second followed by a reiteration of the sentence (e.g., “every big circle was blue”). Their first task was to indicate whether they thought the description was true or false relative to the picture (by pressing ‘J’ or ‘F’ on their keyboard). After answering this initial verification task, they were given a follow-up cardinality question (e.g., “how many big circles were there?”). They typed in a number and progressed to the next trial.

### 3.1.2 Results

Participants’ average accuracy on the initial true/false verification task was 74% (to be maximally conservative, we include all data in the statistical results reported below, but the results remain unchanged if only data from correctly-answered trials are included). The main dependent measure to consider is cardinality estimation accuracy. In particular, we are concerned with estimation accuracy following sentence

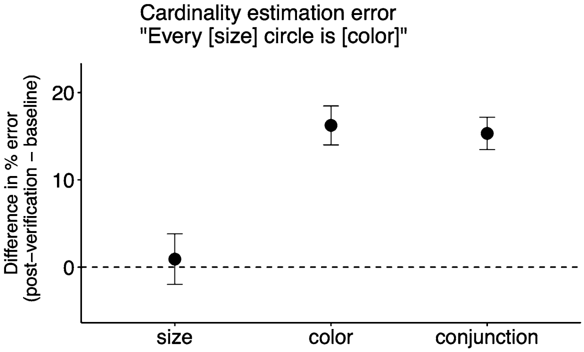

**Fig. 3** Difference in mean percent error between the baseline cardinality estimation task and following the sentence verification task in Experiment 1 ("Every {big/medium/small} circle is {blue/red/yellow}"). In general, higher percent error reflects poorer estimations. For example, if the actual number of dots shown was 10, a response of 8 or 12 would result in 20% error; a response of 6 or 14 would result in 40% error. A larger difference in mean percent error means that post-verification performance was worse than baseline performance for that question type

verification relativized to baseline cardinality estimation accuracy. To the extent that participants did represent a particular group during sentence verification, their cardinality estimation accuracy for that group should be no worse post-verification than it was in the baseline task (i.e., they should know the cardinality as well as their visual system will allow). But if they failed to represent a particular group, their baseline cardinality estimation accuracy should be better for that group than their post-verification accuracy (i.e., they should be expected to perform worse than their visual system will allow). By comparing cardinality estimation accuracy for each group (size; color; conjunction) separately, we control for any differences in visual salience (without relativizing to such baselines, it might be that participants perform better on one question type simply because that feature is an easier one for the visual system to extract).

A simple way to assess accuracy in general is to compute the percent error on each trial (e.g., if the actual number of circles shown was 10, a response of 8 or 12 would result in 20% error; a response of 6 or 14 would result in 40% error). In cardinality estimation tasks, error nearly always goes in one direction: participants routinely underestimate the actual number presented (e.g., Krueger 1984). So we computed a difference score for each trial type: the mean difference in percent error between the baseline cardinality estimation task and the sentence verification task (any trials on which error surpassed 250% were removed as these were likely to reflect typos). As seen in Fig. 3, the predictions of the restricted view were born out. Average percent error was similar post-verification and on the baseline task for cardinality questions that probed the group defined by the determiner's internal argument (size). But average percent error was much higher than baseline for questions that probed the external argument (color) and for questions that probed the conjunction of both (size and color).

Performing paired t-tests on our subjects' mean percent error in the two tasks confirms that accuracy when asked about the group defined by the internal argument (size) does not significantly differ from size-question performance on the baseline task (t(47)=0.32, p=.753). But accuracy when asked about the group defined by the external argument (color) does significantly differ from performance on the corresponding baseline color-questions (t(47)=7.00, p<.001). The same is true for conjunction-questions (t(47)=9.76, p<.001).

That said, arguing from this simple analysis relies on the non-significance of the size-question comparison. In an effort to avoid reasoning from non-significant results, we can adopt a more complex analytical approach. This approach relies on a widely-used model of cardinality estimation that allows for describing performance on a cardinality estimation task with two main parameters: a measure of accuracy ($\beta$) and a measure of variability ($\sigma$) (see Odic et al. 2016 for a helpful review). Responses (y) to being shown some number of things (x) is modeled as the Gaussian distribution in (7), where $\alpha$ is a scaling factor.

(7) $$y \sim \mathcal{N}\left(mean = \alpha x^{\beta}, sd = \sigma \times \alpha x^{\beta}\right)$$

Intuitively, the gist of this model is that a making a numerical estimate is like taking a sample from a normally-distributed pattern of activation on a 'mental number line'. If shown x things, this activation will generally be centered around x, with decreasing activation on either side of x. But our perceptual systems lead to a systematic underestimation of number, so oftentimes the activation will be centered around a value lower than x. Such routine underestimation leads to one source of 'noise' in the estimate of x and is captured by the accuracy parameter $\beta$. With an accuracy of .9, for example, encountering 10 things would lead to an activation pattern centered not around 10, but around $10^{.9} \approx 8$. In addition to underestimation, there is a second source of 'noise' in the estimate of x: internal variability in the activation pattern around $x^{\beta}$. This variability is reflected in the size of the standard deviation, which increases linearly with the mean (i.e., the representation of small numbers is more precise than the representation of large ones). The rate at which the standard deviation increases differs from person to person (and trial to trial), and this difference is captured by the parameter $\sigma$. The scaling factor $\alpha$ is added to both the mean and the standard deviation, to ensure that this model can apply similarly in various numerical ranges (e.g., 30 to 50 objects to enumerate, as in this experiment versus 300 to 500 objects, as in others). But the crucial parameters are those reflecting estimates of accuracy ($\beta$) and precision ($\sigma$).

To the extent that participants' performance on both cardinality estimation tasks (baseline and post-verification) is identical, that performance should be fit equally well by this standard model. But using this model as a starting point, we can consider the contrasting model in (8), which allows the parameters to vary as a function of the task. For example $\beta_0$ is supplemented with $\beta_1$TASK. If we code the baseline task as 0 and the post-verification task as 1, then the value of $\beta_1$ represents the difference in accuracy between baseline cardinality questions and post-verification cardinality questions. If the baseline accuracy is .9, for example, and $\beta_1$ is -.1, that suggests that participants were less accurate at answering "how many?" questions following the sentence verification task.

(8) $$y \sim \mathcal{N}\begin{pmatrix} mean = (\alpha_0 + \alpha_1 TASK)\, x^{\beta_0 + \beta_1 TASK} \\ sd = (\sigma_0 + \sigma_1 TASK) \times mean \end{pmatrix}$$

**Table 1** Model comparisons for Experiment 1. Best model comparison values in bold

| Group probed | Model | AIC | BIC |
|---|---|---|---|
| Size | **Null** | **1763.836** | **1775.772** |
| | Effect of task | 1767.455 | 1791.328 |
| Color | Null | 1580.239 | 1592.043 |
| | **Effect of task** | **1474.125** | **1497.734** |
| Conjunction | Null | 1384.370 | 1396.175 |
| | **Effect of task** | **1248.957** | **1272.567** |

For each trial-type, these two models were fit using maximum likelihood estimation. Two model comparison measures were considered: the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, which reward models for capturing the data and penalize them for including greater numbers of parameters (Schwarz 1978; Stone 1979; Akaike 1998). Lower values are indicative of striking a better trade-off between fit and complexity.

As seen in Table 1, both measures of model comparison align with the predictions of the restricted view. That is, for size-questions, both the AIC and BIC support the null model, which does not differentiate baseline and post-verification cardinality estimation accuracy. This suggests that there is no difference between baseline and post-verification performance when asked about a group defined by size. But for color-questions and conjunction-questions, both measures support the augmented model that takes into account the task.

Examining the coefficients of the best-fitting models for color- and conjunction-questions in Table 2 provides further support for the restricted view. Namely, in the best-fitting models, accuracy decreases from baseline to post-verification (indicated by the negative $\beta_1$ estimate) whereas variability increases (indicated by the positive $\sigma_1$ estimate). This suggests that the effect is found along both of the relevant dimensions to consider when assessing cardinality estimation performance: for color and conjunction questions, participants performed significantly worse than baseline following sentence verification.

As a reviewer points out, we can also fit a single model that allows accuracy and precision to vary based on task type (baseline; sentence verification), question type (size; color; conjunction), and the interaction between the two, as in (9). The interpretation of the resulting coefficients is more complicated, but, given the results above, the prediction is clear: the interaction between task and question type should matter.

$$(9) \qquad y \sim \mathcal{N}\begin{pmatrix} mean = (\alpha_0 + \alpha_1 TASK + \alpha_2 QUESTION + \alpha_3 TASK \times QUESTION) \\ \times x^{\beta_0 + \beta_1 TASK + \beta_2 QUESTION + \beta_3 TASK \times \mathrm{QUESTION}} \\ sd = (\sigma_0 + \sigma_1 TASK + \sigma_2 QUESTION + \sigma_3 TASK \times \mathrm{QUESTION}) \times mean \end{pmatrix}$$

As expected, both model comparison measures favor the model in (9) over an alternative that includes terms for main effects of both task and question type but excludes the interaction between them (AIC: 4499.358 vs. 4540.873; BIC: 4559.939 vs. 4586.308). Moreover, we observe a significant effect of the interaction between task and question type both for accuracy ($\beta_3$=-.237 [95% CI: -.323 to -.152]; z=2.78; p<.01) and for variability ($\sigma_3$=.198 [95% CI: .169 to .227]; z=6.80; p<.001).

**Table 2** Model coefficients for color- and conjunction-questions in Experiment 1. Coefficients showing effect of task in bold

| Group probed | Coefficient | Estimate | SE | z value | P(z) |
|---|---|---|---|---|---|
| Color | $\alpha_0$ | 1.067 | 0.062 | 17.16 | <.001 *** |
| | $\boldsymbol{\alpha_1}$ | **0.587** | **0.204** | **2.88** | **<.01 **** |
| | $\beta_0$ | .9594 | 0.029 | 33.27 | <.001 *** |
| | $\boldsymbol{\beta_1}$ | **-.2199** | **0.064** | **-3.43** | **<.001 ***** |
| | $\sigma_0$ | .1574 | 0.008 | 18.88 | <.001 *** |
| | $\boldsymbol{\sigma_1}$ | **.1815** | **0.021** | **8.66** | **<.001 ***** |
| Conjunction | $\alpha_0$ | 1.032 | 0.044 | 23.61 | <.001 *** |
| | $\boldsymbol{\alpha_1}$ | **0.871** | **0.190** | **4.58** | **<.001 ***** |
| | $\beta_0$ | .9720 | 0.022 | 43.28 | <.001 *** |
| | $\boldsymbol{\beta_1}$ | **-.3401** | **0.055** | **-6.22** | **<.001 ***** |
| | $\sigma_0$ | .1328 | 0.007 | 18.96 | <.001 *** |
| | $\boldsymbol{\sigma_1}$ | **.1628** | **0.018** | **9.15** | **<.001 ***** |

To summarize the results, then, after being asked to evaluate a sentence like "every big circle is blue", participants recall the cardinality of the big circles as well as their visual system will allow. But they are less accurately and less precisely able to recall the cardinality of the blue circles and the cardinality of the big blue circles.

### 3.1.3 Discussion

As noted above, these results align with the restricted view: the group named by *every*'s internal argument (e.g., "the big circles") seems to be explicitly represented by participants during sentence verification, but the same cannot be said for the group named by *every*'s external argument (e.g., "the blue things").[4] This suggests that participants do not understand a sentence like "every big circle is blue" to express a relation between the big circles and the blue things, or between the big circles and the big blue circles.

The result that participants do not accurately recall the cardinality of the group defined by conjunction—big blue circles—is particularly surprising given that half of the trials in the experiment are true, meaning that the answer is the same for both "how many big circles were there?" and "how many big blue circles were there?" (if there were 10 big circles, and every one of them was blue, then there were 10

[4] A reviewer rightly notes that these results do not rule out the following possibility: maybe the things named by the external argument are mentally grouped at some stage of processing, but that group is quickly discarded before the subsequent cardinality question. While logically possible, this strikes us as unlikely, given what is known about the workings of the Approximate Number System (especially the result noted above that observers can represent the cardinality of three groups simultaneously with no apparent cost over and above representing just one; Halberda et al. 2006; Zosh et al. 2011). In any case, for the present argument, all that is needed is a difference in how both of the quantifier's arguments are treated, as the relational view predicts them to be treated on a par and the restricted view predicts them to be treated asymmetrically. Even if the present results merely reflect a difference in likelihood of the extension of either argument being retained in memory, the finding would still fit better with the restricted view.

big blue circles). Nonetheless, participants only encode and recall the cardinality of big circles. This further serves to support the idea that they only represent the group defined by the internal argument of *every* and it further confirms that the experimental design taps into strategies resulting from evaluating an expression's meaning, not knowledge obtained from building a mental model of the display or from downstream inferences. That is, the inference from "every big circle is blue" and "there were 10 big circles" to "there were 10 big blue circles" is not one that participants seem to draw in this task.

To be sure, the results of Experiment 1 only pertain to the meaning of *every*, whereas the restricted quantification view is a claim about quantificational determiner meanings in general. To begin to address the generalizability of the result, Experiment 2 replicates the effect using sentences in which universal quantification is indicated with *all* instead of *every*.

## 3.2 Experiment 2: *all big circles are blue*

If *all* is like *every* in having a restricted meaning, we should observe the same results as in Experiment 1: similar post-verification and baseline performance when asked to estimate the cardinality of a group defined by the relevant size, but worse post-verification performance when asked to estimate the cardinality of a group defined by the relevant color and when asked to estimate the cardinality of a group defined by the relevant conjunction of size and color. Such a result would help ensure that the findings of Experiment 1 do not merely reflect a quirk of *every*.

### 3.2.1 Method

**Participants** Fifty participants were recruited online using Amazon Mechanical Turk. All passed an English-screener and gave informed consent prior to participating in the experiment. One participant was excluded from further analysis for performing at chance or below on the true/false portion of the verification task. One participant was excluded for taking longer than 5 seconds, on average, to respond to follow-up cardinality questions. This left 48 participants.

**Materials** Materials were identical to Experiment 1 in all respects except that *every* was replaced by *all* and the necessary plural agreement was added. Sentences in the verification task were of the form "All {big/medium/small} circles are {blue/red/yellow}". Half were true with respect to the display and half were false.

**Procedure** The procedure was identical to that of Experiment 1. Participants completed 15 trials of the baseline task in which they answered a cardinality question that probed a size, a color, or a size/color combination (e.g., "how many big circles are there?"). They subsequently completed 18 trials of the sentence verification task, in which they first rendered their true/false judgment about sentences like "all big circles are blue", and then were asked the same sorts of cardinality questions as in Experiment 1.

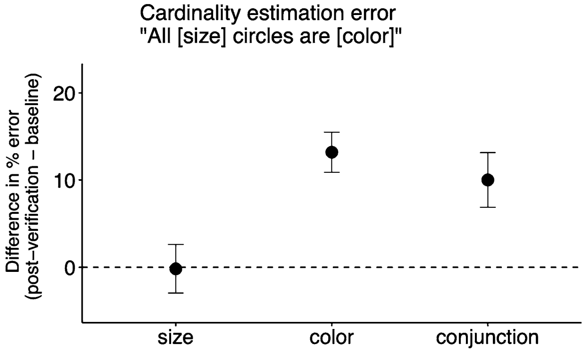

**Fig. 4** Difference in mean percent error between the baseline cardinality estimation task and following the sentence verification task in Experiment 2 ("All {big/medium/small} circles are {blue/red/yellow}")

**Table 3** Model comparisons for Experiment 2. Best model comparison values in bold

| Group probed | Model | AIC | BIC |
|---|---|---|---|
| Size | **Null** | **1695.114** | **1707.013** |
| | Effect of task | 1697.545 | 1721.341 |
| Color | Null | 1567.348 | 1579.207 |
| | **Effect of task** | **1491.407** | **1515.127** |
| Conjunction | Null | 1435.401 | 1447.150 |
| | **Effect of task** | **1412.135** | **1435.633** |

### 3.2.2 Results

Participants' average accuracy on the initial true/false verification task was 74.1%, nearly identical to accuracy in Experiment 1. Figure 4 shows that the predictions of the restricted view were again borne out. That is, we observe the same pattern of results in Experiment 2 with *all* as in Experiment 1 with *every*. In terms of simple comparisons, participants' mean percent error did not significantly differ from baseline on post-verification cardinality questions that probed the target size (t(47)=0.08, p=.935) but participants performed significantly worse than baseline when a cardinality question probed the target color (t(47)=6.84, p<.001) or the relevant conjunction of size and color (t(47)=6.20, p<.001).

Turning to model comparisons, we again find that both the AIC and BIC in Table 3 give reason to prefer the null model for size-questions and the augmented model for color-questions and conjunction-questions, as predicted by the restricted view. Likewise, Table 4 shows the same pattern of decreasing accuracy and increasing variability that we saw in Experiment 1. Namely, for color- and conjunction-questions there is a negative $\beta_1$ estimate and a positive $\sigma_1$ estimate.

### 3.2.3 Discussion

This replication suggests that the results from Experiment 1 do not reflect a quirk of *every*: good knowledge of the cardinality of the group named by the internal argument but not of the group named by the external argument or by the conjunction of both

**Table 4** Model coefficients for color- and conjunction-questions in Experiment 2. Coefficients showing effect of task in bold

| Group probed | Coefficient | Estimate | SE | z value | P(z) |
|---|---|---|---|---|---|
| Color | $\alpha_0$ | 1.110 | 0.064 | 17.21 | <.001 *** |
| | $\boldsymbol{\alpha_1}$ | **1.035** | **0.275** | **3.76** | **<.001 *****|
| | $\beta_0$ | .9484 | 0.030 | 32.10 | <.001 *** |
| | $\boldsymbol{\beta_1}$ | **-.3606** | **0.069** | **-5.26** | **<.001 *****|
| | $\sigma_0$ | .1828 | 0.010 | 19.02 | <.001 *** |
| | $\boldsymbol{\sigma_1}$ | **.1375** | **0.020** | **6.75** | **<.001 *****|
| Conjunction | $\alpha_0$ | 1.087 | 0.080 | 13.57 | <.001 *** |
| | $\boldsymbol{\alpha_1}$ | **0.673** | **0.177** | **3.81** | **<.001 *****|
| | $\beta_0$ | .9559 | 0.039 | 24.63 | <.001 *** |
| | $\boldsymbol{\beta_1}$ | **-.2622** | **0.061** | **-4.33** | **<.001 *****|
| | $\sigma_0$ | .2179 | 0.012 | 18.13 | <.001 *** |
| | $\boldsymbol{\sigma_1}$ | **.0633** | **0.020** | **3.23** | **<.01 ****|

arguments. Of course, this effect might still be unique to English universal quantifiers (as opposed to proportional quantifiers like *most*, negative quantifiers like *no*, and existential quantifiers like *some*); we return to this possibility in Sect. 4.

In the meantime, we consider other possible alternative explanations for the present results. One potential concern is that the surface-level asymmetry between the two predicates could explain the results. The internal argument is introduced as an NP (e.g., “big circle” or “big circles”) whereas the external argument is introduced as a VP (e.g., “is blue” or “are blue”). It could be that this low-level difference in how these predicates are introduced plays a role in driving attention to the group named by the internal argument to the exclusion of the group named by the external argument. Namely, it is possible that, regardless of the sentence meaning, participants will never mentally group (and thus enumerate) things named by a property introduced as “is X”. Likewise, it might be that, regardless of the sentence meaning, they always mentally group (and thus enumerate) things named by a property introduced as “X circle”. To be sure, if this sort of reasoning does explain our results, it is not obviously an argument in favor of relational quantification. After all, the standard view is that *every A is B* treats the As and the Bs on a par, despite the syntactic differences in how the predicates A and B are introduced. Nonetheless, Experiments 3 and 4 aim to address this possibility empirically, by equating how the two predicates are introduced.

### 3.3 Experiment 3: *every circle that is big is blue*

In this version of the task, the relevant part of the internal argument (e.g., “big”) is introduced in a relative clause. As a result, both predicates are introduced in the same way, at least on the surface (e.g., “is big” and “is blue”). Obviously there are still differences between *circle that is big* and *is blue*, but the purpose of this manipulation is to control for the surface-level difference present in Experiments 1 and 2. If that sort of surface-level difference between how the predicates were introduced (in a tensed VP versus in an NP) were for some reason responsible for the results of Experiments 1 and 2, owing to some psychological concomitant of the grammatical distinction, then the results should disappear in the present experiment. On the other hand, the restricted view predicts that this manipulation should make no difference and we should thus replicate the effects of Experiments 1 and 2: post-verification performance is expected to be similar to baseline performance for size-questions but worse for color- and conjunction-questions.

### 3.3.1 Method

**Participants** Fifty-three participants were recruited online using Amazon Mechanical Turk. All passed an English-screener and gave informed consent prior to participating in the experiment. Five participants were excluded from further analysis for performing at chance or below on the true/false portion of the verification task. This left 48 participants.

**Materials** Materials were identical to Experiment 1 except that the predicate in the internal argument was introduced with a relative clause. This led to sentences in the verification task being of the following form: "Every circle that is {big/medium/small} is {blue/red/yellow}". Half were true with respect to the display and half were false.

**Procedure** The procedure was identical to that of Experiments 1 and 2. Participants completed 15 trials of the baseline task in which they answered a cardinality question that probed a size, a color, or a size/color combination (e.g., "how many big circles are there?"). They subsequently completed 18 trials of the sentence verification task, in which they were first asked to offer a true/false judgment about sentences like "every circle that is big is blue" and subsequently asked the same sorts of cardinality questions as in Experiments 1 and 2.

### 3.3.2 Results

Participants' average accuracy on the initial true/false verification task was 74.8%. As seen in Fig. 5, this change in how the question was posed led to the same pattern of results as in Experiments 1 and 2. Participants' mean percent error did not significantly differ from baseline on post-verification cardinality questions that probed the target size (t(47)=0.98, p=.333), but participants performed significantly worse than baseline when a cardinality question probed the target color (t(47)=5.94, p<.001) or the relevant conjunction of size and color (t(47)=4.76, p<.001).

Turning to the model comparison analysis, the main results of Experiments 1 and 2 are replicated again: both the AIC and BIC in Table 5 give reason to prefer the null model for size-questions and the augmented model for color-questions. Conjunction-questions produced a less conclusive result, as the AIC and BIC disagree (likely owing to the fact that BIC penalizes additional parameters to a greater extent).

Looking at the model coefficients in Table 6, the significant positive $\sigma_1$ estimate suggests that if the effect is present for conjunction-questions, it is driven by an increase in variability.

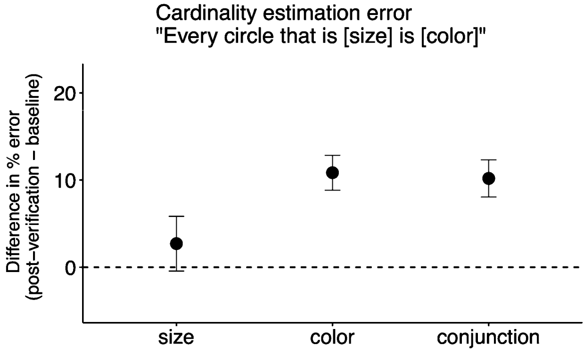

**Fig. 5** Difference in mean percent error between the baseline cardinality estimation task and following the sentence verification task in Experiment 3 (“Every circle that is {big/medium/small} is {blue/red/yellow}”)

**Table 5** Model comparisons for Experiment 3. Best model comparison values in bold

| Group probed | Model | AIC | BIC |
|---|---|---|---|
| Size | **Null** | **1714.583** | **1726.403** |
| | Effect of task | 1718.282 | 1741.923 |
| Color | Null | 1513.341 | 1525.208 |
| | **Effect of task** | **1483.49** | **1507.225** |
| Conjunction | Null | 1462.979 | **1474.824** |
| | Effect of task | **1455.743** | 1479.431 |

**Table 6** Model coefficients for color- and conjunction-questions in Experiment 3. Coefficients showing effect of task in bold

| Group probed | Coefficient | Estimate | SE | z value | P(z) |
|---|---|---|---|---|---|
| Color | $\alpha_0$ | 1.108 | 0.081 | 13.71 | <.001 *** |
| | $\boldsymbol{\alpha_1}$ | **0.448** | **0.161** | **2.78** | **<.01 **** |
| | $\beta_0$ | .9596 | 0.036 | 26.80 | <.001 *** |
| | $\boldsymbol{\beta_1}$ | **-.2060** | **0.057** | **-3.59** | **<.001 ***** |
| | $\sigma_0$ | .1908 | 0.010 | 19.02 | <.001 *** |
| | $\boldsymbol{\sigma_1}$ | **.0807** | **0.018** | **4.50** | **<.001 ***** |
| Conjunction | $\alpha_0$ | 1.189 | 0.084 | 14.21 | <.001 *** |
| | $\alpha_1$ | 0.116 | 0.141 | 0.83 | .408 |
| | $\beta_0$ | .9129 | 0.036 | 25.17 | <.001 *** |
| | $\beta_1$ | -.0751 | 0.057 | -1.32 | .187 |
| | $\sigma_0$ | .2130 | 0.011 | 18.76 | <.001 *** |
| | $\boldsymbol{\sigma_1}$ | **.0600** | **0.019** | **3.19** | **<.01 **** |

### 3.3.3 Discussion

Though the conjunction result is weaker, the difference between baseline and post-verification color-questions is clearly replicated. Experiment 3 thus offers at least a replication of the main result from Experiment 1 while controlling for surface-level differences by introducing both predicates within VPs. Expanding on this control, Experiment 4 offers another way to equate the introduction of the predicates; namely, by introducing both predicates within NPs.

## 3.4 Experiment 4: *every big one is a blue one*

As with Experiment 3, this version of the task attempts to equate how both predicates are introduced. Experiment 4 does so by introducing both predicates as modifiers of "one". Namely, sentences like "every big one is a blue one" were used. The use of pronominal *one* might introduce its own issues, since it requires anaphoric interpretation. But, while this may add additional noise, we assume that this added complication will cause no major problems, as the intended interpretation, with *one* anaphoric to *circles*, seems tolerably clear in the context of the experiment (answering questions about circles). So if surface-level differences between how the predicates were introduced were responsible for the results of Experiments 1-2, the results should disappear in the present experiment. On the other hand, if a restricted meaning is responsible for the observed asymmetry, then this manipulation should likewise make no difference and we should replicate the effects of Experiments 1 and 2: high cardinality estimation accuracy on size questions but comparatively low accuracy on color and conjunction questions.

### 3.4.1 Method

**Participants** Fifty-one participants were recruited online using Amazon Mechanical Turk. All passed an English-screener and gave informed consent prior to participating in the experiment. Two participants were excluded from further analysis for performing at chance or below on the true/false portion of the verification task. One participant was excluded for taking longer than 5 seconds, on average, to respond to follow-up cardinality questions. This left 48 participants.

**Materials** Materials were identical to Experiment 1 except that "circle" was not mentioned in sentences in the verification task. Instead, sentences had the following form: "Every {big/medium/small} one is a {blue/red/yellow} one", where context made it clear that "Every big one" referred to every big circle. Half of the sentences were true with respect to the display and half were false.

**Procedure** The procedure was identical to that of Experiments 1-3. Participants completed 15 trials of the baseline task in which they answered a cardinality question that probed a size, a color, or a size/color combination (e.g., "how many big circles are there?"). They subsequently completed 18 trials of the sentence verification task, in which they offered true/false judgments of sentences like "every big one is a blue one" before being asked the same sorts of cardinality questions as in Experiments 1-3.

**Fig. 6** Difference in mean percent error between the baseline cardinality estimation task and following the sentence verification task in Experiment 4 ("Every {big/medium/small} one is a {blue/red/yellow} one")

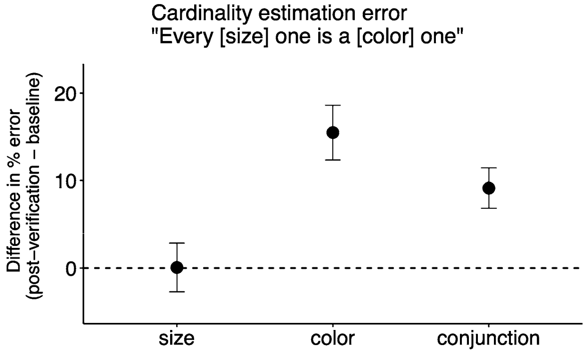

**Table 7** Model comparisons for Experiment 4. Best model comparison values in bold

| Group probed | Model | AIC | BIC |
|---|---|---|---|
| Size | Null | 1658.952 | **1670.749** |
| | Effect of task | **1656.721** | 1680.314 |
| Color | Null | 1708.582 | 1720.564 |
| | **Effect of task** | **1569.476** | **1593.440** |
| Conjunction | Null | 1393.118 | 1404.883 |
| | **Effect of task** | **1350.246** | **1373.776** |

### 3.4.2 Results

Participants' average accuracy on the initial true/false verification task was 73.1%. As seen in Fig. 6, this replication was successful. Participants did not perform significantly different from baseline when asked a cardinality question probing target size (t=.03, p=.979), but they were significantly worse than baseline when asked about the target color (t=5.04, p<.001) or the relevant conjunction of features (t=4.79, p<.001). And in terms of model comparison, Table 7 shows that the AIC and BIC both favor the augmented model for color-questions and conjunction-questions, as predicted.

Interpreting the results are slightly more complicated when it comes to size-questions. The BIC prefers the null model, as predicted, but the AIC favors the augmented model. However, examining the model coefficients in Table 8 reveals that this difference is driven by an apparent *decrease* in variance (a negative $\sigma_1$ estimate) following sentence verification. In other words, participants performed slightly *better* on post-verification size-questions than on baseline size questions (to be precise, they were more precise following sentence verification). We suspect this is a spurious result, and it might be that the added complication of including pronominal *one* contributed to a noisier overall result in Experiment 4.

### 3.4.3 Discussion

As in the previous experiments, cardinality estimation performance on color-questions in Experiment 4 was worse than baseline along both of the relevant dimensions (accuracy and precision). Taken together with Experiment 3 (e.g., "every

**Table 8** Model coefficients for size-, color-, and conjunction-questions in Experiment 4. Coefficients showing effect of task in bold

| Group probed | Coefficient | Estimate | SE | z value | P(z) |
|---|---|---|---|---|---|
| Size | $\alpha_0$ | 1.695 | 0.181 | 9.34 | <.001 *** |
| | $\alpha_1$ | -0.217 | 0.222 | -0.98 | .330 |
| | $\beta_0$ | .7357 | 0.053 | 13.99 | <.001 *** |
| | $\beta_1$ | .0702 | 0.068 | 1.03 | .304 |
| | $\sigma_0$ | .3318 | 0.019 | 17.41 | <.001 *** |
| | **$\sigma_1$** | **-.0591** | **0.024** | **-2.44** | **<.05 *** |
| Color | $\alpha_0$ | 0.899 | 0.050 | 17.86 | <.001 *** |
| | **$\alpha_1$** | **1.374** | **0.249** | **5.52** | **<.001 ***** |
| | $\beta_0$ | 1.052 | 0.028 | 37.57 | <.001 *** |
| | **$\beta_1$** | **-.4603** | **0.060** | **-7.66** | **<.001 ***** |
| | $\sigma_0$ | .1715 | 0.009 | 19.87 | <.001 *** |
| | **$\sigma_1$** | **.1617** | **0.021** | **7.82** | **<.001 ***** |
| Conjunction | $\alpha_0$ | 1.076 | 0.068 | 15.71 | <.001 *** |
| | $\alpha_1$ | 0.135 | 0.129 | 1.04 | .297 |
| | $\beta_0$ | .9334 | 0.033 | 27.90 | <.001 *** |
| | $\beta_1$ | -.0500 | 0.058 | -0.87 | .385 |
| | $\sigma_0$ | .1809 | 0.010 | 18.43 | <.001 *** |
| | **$\sigma_1$** | **.1151** | **0.019** | **6.03** | **<.001 ***** |

circle that is big is blue"), these results militate against an explanation in terms of some low-level differences in how the two predicates are introduced. It doesn't seem to be true that participants never mentally represent groups named by predicates introduced in VPs, and it likewise doesn't seem to be true that participants always mentally represent groups named by predicates introduced within NPs.

Nonetheless, the analyses presented above are complicated. And one may feel the desire for a more straightforward dependent measure. Instead of comparing accuracy and precision of numerical estimates, perhaps the same effect could be seen by simply asking participants about their confidence in having attended the group in question. A final replication, Experiment 5, provides a simpler dependent measure along these lines: adding an 'opt-out' button that allows participants to declare they have no idea how many circles there were.

### 3.5 Experiment 5: adding an opt-out button

Instead of measuring participants' accuracy on cardinality questions, Experiment 5 measures their rate of choosing *not* to answer particular questions. The impetus for this manipulation is the general worry that, in the preceding tasks, the focus on providing a number lead to some idiosyncratic guessing patterns related to number. If idiosyncrasies of numerical responses exist, they might be responsible for making participants look as if they perform worse at questions probing the external argument and the conjunction of both arguments. As a solution, we adopt a response structure that has been fruitfully deployed elsewhere in psychophysics (e.g., Smith et al. 1995, 1997; Ferrigno et al. 2019): including an 'opt-out' button. The prediction is that if participants represent a particular group of circles during verification, they should opt-out no more often than baseline (which, as before, serves as a measure of how difficult it is to enumerate each type of group). If participants do not represent a particular group, they should opt out more often than baseline.

Given the results of Experiments 1-4 and the restricted quantification view, we expect participants to be more likely than baseline to opt out given a question that probes a group picked out by the relevant color (the external argument) or size/color combination. And we predict them to be no more likely than baseline to opt out of a question probing the relevant size (the internal argument).

### 3.5.1 Method

**Participants** Fifty-six participants were recruited online using Amazon Mechanical Turk. All passed an English-screener and gave informed consent prior to participating in the experiment. Four participants were excluded from further analysis for performing at chance or below on the true/false portion of the verification task. Four participants were excluded for taking longer than 5 seconds, on average, to respond to cardinality questions. This left 48 participants.

**Materials** Materials were identical to Experiment 1 except that both the baseline task and sentence verification task allowed participants the ability to opt out of answering the cardinality question. Underneath the question (e.g., "How many big circles were there?") there was a large red button labeled "I don't know!" that participants could press in lieu of making a guess.

**Procedure** The procedure was identical to that of Experiments 1-4 except that participants were instructed that they could press the red "I don't know!" button, without penalty, whenever they felt they would otherwise be making a complete guess. This only applied to the follow-up cardinality questions; participants were still required to answer the initial true/false questions, as in Experiments 1-4.

### 3.5.2 Results

Participants' average accuracy on the initial true/false verification task was 73.8%. Figure 7 shows that the predictions described above were borne out. We find a significant interaction between question type and task ($F_{2,94}$=7.33, p<.01; tests conducted on participants' rate of pressing the opt-out button). And as expected, participants were no more likely than baseline to opt-out of post-verification size-questions (t(47)=.47, p=.640) but they were significantly more likely than to opt-out of post-verification color-questions (t(47)=3.07, p<.01) and post-verification conjunction-questions (t(47)=2.91, p<.01).

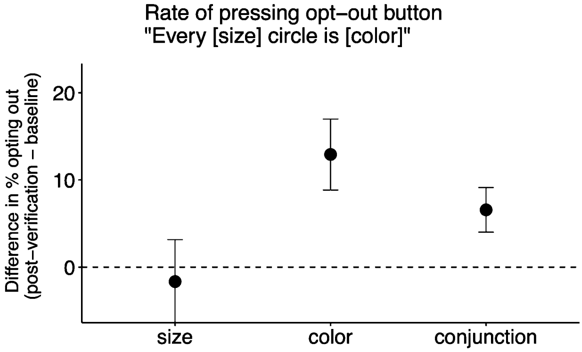

**Fig. 7** Difference in participants' average rate of pressing the 'I don't know!' opt-out button between the baseline cardinality estimation task and following the sentence verification task ("Every {big/medium/small} circle is {blue/red/yellow}"). Higher values reflect an increased rate of opting-out following sentence verification

### 3.5.3 Discussion

This result serves as additional confirmation of the original result. Namely, participants know the cardinality of the group defined by the internal argument (size) as well as their visual system allows, but show deficiencies in their knowledge of cardinalities of the groups defined by the external argument (color) or the conjunction of both arguments. As a result, they opt-out more often than their baseline rate for those latter two types of questions. This relatively straightforward signature of the effect does not require any psychophysical modeling and goes some way toward avoiding the worry that the effect reflects something idiosyncratic about number estimation.

## 4 General discussion

Taken together, the results of Experiments 1-5 suggest that given a universally quantified sentence to evaluate (e.g., *every big circle is blue*), participants explicitly group the satisfiers of the determiner's internal argument (e.g., *big circle*) but treat the satisfiers of its external argument (e.g., *is blue*) differently. Given the comparisons against baseline measures of performance, we can be confident that this difference stems from the sentence and not ancillary properties of the visual system.[5] This finding would be surprising if such determiners had genuinely relational meanings that call for treating both the internal and external arguments as logically on a par (e.g., "the big circles are among the blue things"). After all, when shown multi-colored arrays of dots and tested for their estimation abilities (roughly similar to our baseline condition here), humans can recall the cardinality of three groups of items without incurring any additional cost over and above representing only one group (Halberda et al. 2006; Zosh et al. 2011). So there is no reason to suspect a constraint from the visual system that precludes representing two groups simultaneously. Yet participants seem

[5] Post-verification cardinality questions probing each feature were compared against baseline questions probing that same feature, controlling for the possibility that one of the features is more visually salient than the other. To further control for this possibility, Knowlton et al. (2021b) conducted a "swapped argument" version of Experiment 1, in which participants were asked to evaluate sentences like *every blue circle is big* instead of *every big circle is blue*.

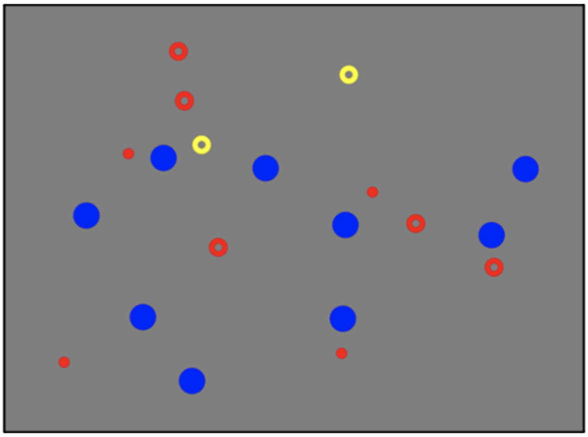

**Fig. 8** An example trial in which the sentence was *every big circle is blue* and it was true relative to the display. In such a case, the number of big circles and big blue circles is the same (Color figure online)

to avoid doing so in our task, despite knowing that they will be asked a "how many?" question after each trial of sentence verification.

Even more surprisingly, the results described above suggest that participants do not routinely represent the conjunction of the determiner's internal and external arguments (e.g., the big circles that are blue), at least not to the same degree that they group the things named by the internal argument. This result is especially striking because on half of the trials that participants completed the quantificational sentences used were true. For example, every big circle is blue in a case like Fig. 8. In such cases, the big circles are the big blue circles. And consequently, there are the same number of big circles as big blue ones. Nonetheless, we find that participants know the answer to a cardinality question that probes the internal argument (e.g., the big circles as such) perfectly well, but perform far worse when asked a question that describes the very same things as a conjunction of the internal and external arguments (e.g., the big blue circles).

That participants fail to routinely and rapidly make the inference from every big circle being blue to there being the same number of big circles and big blue circles is important. For one thing, it tells against an augmented version of the standard view on which the external argument of *every big circle is blue* is not *is blue* but *big circle that is blue* (Romoli 2015).[6] But perhaps more importantly, it bolsters the case for thinking that the experimental design deployed here taps into verification strategies that result from the meanings of the expressions themselves and not from downstream inferences.

[6] This is the hypothesis indicated by the representation in (6c), namely '{x: x is a big circle} = {x: x is a big circle} ∩ {x: x is blue}'. A reviewer rightly points out that the expectations about the behavioral repercussions of this representation differ based on how one interprets '∩'. For simplicity, we assumed '{x: x is a big circle} ∩ {x: x is blue}' and '{x: x is a big circle that is blue}' are identical. But one might instead suppose that '{x: x is a big circle} ∩ {x: x is blue}' calls for representation both of the big circles and of the blue circles and then arrives at representation of the big blue circles by means of a further computation: intersecting those two sets. In this case, one might expect (6c) to lead to representation of three sets: the big circles, the blue circles, and the big blue circles. In any case, the data do not bear out this prediction.

To elaborate on this point, imagine that participants had performed well both on cardinality questions probing the internal argument and on cardinality questions probing conjunction. Given those results, one might have wondered whether there were two distinct reasons for good cardinality estimation performance on the task. It might have been that participants represented the group defined by the internal argument as a result of formal properties of the meaning but represented the group defined by the conjunction of both arguments as a result of inference (e.g., "there were about 10 big circles, every big circle was blue, therefore there must have been about 10 big blue circles"). But if the experimental design had permitted two routes to success in this way, one might wonder whether the same reasoning could explain the performance on the questions probing the internal argument. It is not immediately clear what such an inference would look like, but fortunately, the conjunction result removes this potential worry altogether. It shows that participants were not performing well on cardinality estimation questions thanks to explicit reasoning about the task.

Instead, the conjunction result provides more reason for thinking that meanings have particular formal characteristics that matter. Assuming that what speakers know when they know the meaning of *every* is an equivalence class of ways of specifying its truth-conditions leaves no easy way of explaining the present results, even if that equivalence class is somehow restricted to finite size. But if we instead suppose that meanings are mental representations specified in a particular format, then what the representation makes explicit can license some predictions about what participants are likely to represent in the course of sentence understanding and verification. This is not to say formal details of the representation in question will constitute an exhaustive theory of what goes on in a participants' mind when understanding a sentence, just that meanings play a causal role and that their contribution can be detected.

In particular, these results are well-explained if, contra the relational view, determiners like *every* and *all* have restricted meanings, which differentiate their two grammatical arguments in their logical role. Moreover, the present results support the further claim that these determiners have meanings that only call for grouping the restriction, provided by the internal argument. That is, suppose that instead of describing a relation exhibited by the set of frogs and the set of green things, a sentence like *every frog is green* has a meaning more like (5), repeated below, which introduces only a single group, the frogs.

(5) a. $\iota$X:Frogs(X) $\upharpoonright$ $\forall$x[Green(x)]
    b. $\approx$ Relative to the frogs, every thing$_{\mathrm{x}}$ is such that it$_{\mathrm{x}}$ is green

The green things, as such, are not explicitly encoded in this representation.[7] Likewise, the conjunction of the two grammatical arguments is not explicitly encoded in

[7] The question arises whether the difference in logical role itself is responsible for this difference in which groups are represented. Alternatively, this latter difference might be explained by a difference in whether the terms are themselves group-denoting as opposed to being an open formula including a singular variable. In this context, distinctions between *each*, *every*, and *all* might be relevant. The proposed meaning in (5) is for *every*. Knowlton (2021) argues that both *each* and *every* have restricted meanings, and both distributively apply the predicate supplied by the external argument to the members of the restricted domain, but only *every* calls for grouping that restricted domain; *each* calls for treating it as a series of independent individuals. On the other hand, *all* is restricted and calls for grouping the restricted domain (like *every*), but might not call for distributively applying the predicate supplied by the external argument.

(5). But the grouping of the internal argument is. If understanding an expression like *every frog is green* entails building a mental representation formally like (5) in this respect, then it should come as no surprise that participants mentally represent the frogs as such but not the green things or the green frogs. These latter two groups are not highlighted by the representation.

To be sure, our claim that determiners have restricted—as opposed to relational—meanings is not meant to be restricted to universal quantifiers like *every*. The same should equally-well apply to proportional quantifiers like *most* and existential quantifiers like *some*. Such non-universal quantificational content can be coded in non-relational terms analogous to the treatment given here for *every*. For example, the specification in (10) offers a restricted meaning of *most of the frogs are green* that also reflects the psycholinguistic work discussed above (Pietroski et al. 2009; Lidz et al. 2011; Knowlton et al. 2021a), which argues that the meaning of such a sentence implicates a mental grouping of the frogs as well as the green frogs. Despite introducing two groups, the specification for *most* in (10) is still restricted in the following sense: the determiner's meaning permits only a single selection from the domain (the frogs) and subsequently calls for further selection from that restricted domain (relative to the frogs, select the green ones).

(10) a. $\iota\mathrm{F{:}Frogs(F)} \rceil \exists \mathrm{Y{:}}\forall \mathrm{x[Y(x)} \equiv \mathrm{Green(x)]\{|Y| > (|F| - |Y|)\}}$
b. ≈ Relative to the frogs$_\mathrm{F}$ there are some things$_\mathrm{Y}$ (that are such that each thing$_\mathrm{x}$ is one of them$_\mathrm{Y}$ iff it$_\mathrm{x}$ is green) such that they$_\mathrm{Y}$ outnumber the difference between the frogs$_\mathrm{F}$ and themselves$_\mathrm{Y}$

This specification stands in contrast to a genuinely relational alternative: a meaning that calls for two separate selections from the domain (the frogs and the green things) and then intersects these selections to form the relevant conjunction ('{x: Frog(x)} ∩ {x: Green(x)}'). So while *most* likely does implicate representation of two groups, it could nonetheless be restricted as opposed to relational if the second group is not independently selected or formed on the basis of an independent selection. The relevant question is thus not whether *most* can be stated in restricted terms but whether evidence can be found that such a hypotheses about what speakers know when they know the meaning of *most* is preferred over a genuinely relational alternative.

In future work, we hope to find an empirical signature that reflects the difference between two independent selections from the domain that are intersected and a second selection made from an initially restricted domain (relatedly, see fn. 5). A potentially related question for future work is how the proposed representations connect to complex anaphora (e.g., Moxey and Sanford 1993; Sanford et al. 1996; Paterson et al. 1998). For example, in (11a), *they* is anaphoric to the fans who left, and in (11b), *they* is anaphoric to the fans who didn't attend.

(11) a. Most fans left the game early, they were tired.
b. Few fans went to the game, they watched at a bar.

As on traditional views (including Generalized Quantifier Theory and dynamic approaches), we do not assume that the proposed meanings of quantifiers obviate the need for a theory of anaphora (e.g., Kamp and Reyle 1993; Nouwen 2010). In particular, reference to the restricted domain as a group in the meaning representation does not mark that group as the only target of subsequent anaphora. And while the group(s) explicitly introduced in the meaning representation may exert some influence here (Knowlton and Schwarz Forthcoming), we still suspect many of these effects should instead be explained by additional principles of the discourse pragmatics (e.g., Kibble 1997; Hendriks and de Hoop 2001; Nouwen 2003; Zulaica-Hernández 2018). In any case, we do not mean to suggest that only the group(s) implicated by the quantifier meaning is represented at any point.

Setting aside these future directions, at least the main prediction of the restricted quantification view is clear: quantificational determiners should encourage participants to treat their two grammatical arguments asymmetrically. In addition to confirming this main prediction as it pertains to other kinds of quantifiers, it is necessary to broaden the empirical landscape to other languages. It is, of course, always possible that the present results reflect a quirk of English and will ultimately generalize no further than the cases thus far tested. To guard against this possibility, we are currently working to replicate these results in other languages, including Mandarin and Italian. In the meantime, we hope that by testing, and confirming, the predictions of the restricted view in one case, we have provided initial evidence for this hypothesis in a way that invites and encourages further cross-linguistic investigation.

## 5 Conclusion

A quantificational determiner like *every* in a sentence like *every dog slept* is often taken to express a second-order relation. On this view, the two syntactic arguments—*dog* and *slept*—supply two predicates, and *every* specifies how the extensions of those predicates are related. This relational conception of determiners stems, in part, from important developments in logic. In particular, Frege invented a logic that easily accommodates relations as part of a broader project of reducing arithmetic to logic. And within linguistics, this relational view has been useful for stating and discovering generalizations, such as the generalization that all determiners are conservative.

But despite often being modeled as expressing relations, quantificational determiners, in our view, are not understood (i.e., mentally represented by speakers) in relational terms. Instead, they combine with their syntactic complement to form a monadic quantifier over a restricted domain. The main difference between these views is how the determiner treats its two grammatical arguments: as serving identical logical roles (two terms in a relation) or as having distinct roles (restricting the domain versus providing an additional criterion that some number of domain entities must meet). This alternative, restricted conception of determiner meanings offers a simple explanation of the conservativity constraint and of the behavioral findings that participants treat the determiner's internal and external arguments in surprisingly different ways (at least in the cases so far tested). More generally, the restricted view of quantification follows in the linguistic tradition of searching for the most constrained system consistent with the data, with the aim of limiting overgeneration from the outset.

**Acknowledgements** We thank Amy Rose Deal and two anonymous *Natural Language Semantics* reviewers for their thoughtful comments on previous versions of this manuscript. For helpful discussion about this project, we thank Alexis Wellwood, Valentine Hacquard, Norbert Hornstein, and audiences at ELM 1, CUNY/HSP 33, and SALT 30. We also thank Simon Chervenak for help collecting the data. This work was supported by funding from the National Science Foundation (#BCS-2017525 and #NRT-1449815).

## Declarations

**Competing Interests** The authors declare no competing interests.

## References

Akaike, Hirotugu. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, eds. E. Parzen, K. Tanabe, and G. Kitagawa. New York: Springer. https://doi.org/10.1007/978-1-4612-1694-0_15.

Ariely, Dan. 2001. Seeing sets: Representation by statistical properties. *Psychological Science* 12(2): 157–162. https://doi.org/10.1111/1467-9280.00327.

Barwise, Jon, and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy* 4: 159–219. https://doi.org/10.1007/BF00350139.

Beghelli, Filippo. 1997. The syntax of distributivity and pair-list readings. In *Ways of scope taking*, ed. A. Szabolcsi, 349–408. Dordrecht: Springer. https://doi.org/10.1007/978-94-011-5814-5_10.

Beghelli, Filippo, and Tim Stowell. 1997. Distributivity and negation: The syntax of each and every. In *Ways of scope taking*, ed. A. Szabolcsi, 71–107. Dordrecht: Springer. https://doi.org/10.1007/978-94-011-5814-5_3.

Ben-Yami, Hanoch. 2009. Generalized quantifiers, and beyond. *Logique Et Analyse* 52(208): 309–326. https://www.jstor.org/stable/44084931.

Ben-Yami, Hanoch. 2012. Response to Westerståhl. *Logique Et Analyse* 55(217): 47–55. https://www.jstor.org/stable/44085100.

Chen, Lin. 1982. Topological structure in visual perception. *Science* 218(4573): 699–700. https://doi.org/10.1126/science.7134969.

Chen, Lin. 2005. The topological approach to perceptual organization. *Visual Cognition* 12(4): 553–637. https://doi.org/10.1080/13506280444000256.

Chomsky, Noam. 1956. Three models for the description of language. *I.R.E. Transactions on Information Theory* 2(3): 113–124. https://doi.org/10.1109/TIT.1956.1056813.

Chomsky, Noam. 1959. On certain formal properties of grammars. *Information and Control* 2(2): 137–167. https://doi.org/10.1016/S0019-9958(59)90362-6.

Church, Alonzo. 1941. *The calculi of lambda conversion*. Princeton: Princeton University Press.

Dehaene, Stanislas. 2011. *The number sense: How the mind creates mathematics*. Oxford: Oxford University Press.

Denić, Milica, and Jakub Szymanik. 2022. Are most and more than half truth-conditionally equivalent? *Journal of Semantics* 39(2): 261–294. https://doi.org/10.1093/jos/ffab024.

Feigenson, Lisa, Stanislas Dehaene, and Elizageth Spelke. 2004. Core systems of number. *Trends in Cognitive Sciences* 8: 307–314. https://doi.org/10.1016/j.tics.2004.05.002.

Ferrigno, Stephen, Gabrielle Bueno, and Jessica F. Cantlon. 2019. A similar basis for judging confidence in monkeys and humans. *Animal Behavior and Cognition* 6(4): 335–343. https://doi.org/10.26451/abc.06.04.12.2019.

von Fintel, Kai, and Lisa Matthewson. 2008. Universals in semantics. *The Linguistic Review* 25: 139–201. https://doi.org/10.1515/TLIR.2008.004.

Frege, Gottlob. 1879. Begriffsschrift. In *Frege to Gödel: A source book in mathematical logic*, ed. J. van Heijenoort, 1879–1931. Cambridge: Harvard University Press.

Frege, Gottlob. 1884. *Die Grundlagen der Arithmetik. Breslau: Wilhelm Koebner. English translation in The Foundations of Arithmetic*. Oxford: Basil Blackwell. Trans. J.L. Austin.

Frege, Gottlob. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100: 25–50.

Geurts, Bart, and Rick Nouwen. 2007. 'At least' et al.: The semantics of scalar modifiers. *Language* 83(3): 533–559.

Haberman, Jason, and David Whitney. 2012. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In *From perception to consciousness: Searching with Anne Treisman*, eds. J. Wolfe and L. Robertson, 339–349. Oxford: Oxford University Press.

Hackl, Martin. 2009. On the grammar and processing of proportional quantifiers: *Most* versus *more* than half. *Natural Language Semantics* 17: 63–98. https://doi.org/10.1007/s11050-008-9039-x.

Halberda, Justin, Sean F. Sires, and Lisa Feigenson. 2006. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science* 17(7): 572–576. https://doi.org/10.1111/j.1467-9280.2006.01746.x.

Heinz, Jeffrey, and William Idsardi. 2011. Sentence and word complexity. *Science* 333(6040): 295–297. https://doi.org/10.1126/science.1210358.

Heinz, Jeffrey, and William Idsardi. 2013. What complexity differences reveal about domains in language. *Topics in Cognitive Science* 5(1): 111–131. https://doi.org/10.1111/tops.12000.

Hendriks, Petra, and Helen de Hoop. 2001. Optimality theoretic semantics. *Linguistics and Philosophy* 24: 1–32. https://doi.org/10.1023/A:1005607111810.

Higginbotham, James, and Robert May. 1981. Questions, quantifiers and crossing. *The Linguistic Review* 1: 41–80. https://doi.org/10.1515/tlir.1981.1.1.41.

Hodges, Wilfred. 2012. Formalizing the relationship between meaning and syntax. In *The Oxford handbook of compositionality*, eds. M. Werning, W. Hinzen, and E. Machery, 245–261. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199541072.013.0011.

Hunter, Tim, and Jeffrey Lidz. 2013. Conservativity and learnability of determiners. *Journal of Semantics* 30(3): 315–334. https://doi.org/10.1093/jos/ffs014.

Icard, Thomas F., and Lawrence S. Moss. 2022. A simple logic of concepts. *Journal of Philosophical Logic* 52: 1–26. https://doi.org/10.1007/s10992-022-09685-1.

Joshi, Aravind K. 1985. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In *Natural language parsing: Psychological, computational, and theoretical perspectives*, eds. D. Dowty, L. Karttunen, and A. Zwicky, 206–250. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.007.

Joshi, Aravind K., K. Vijay Shanker, and David Weir. 1990. The convergence of mildly context-sensitive grammar formalisms. In *Foundational issues in natural language processing*, eds. P. Sells, S. Shieber, and T. Wasow, 31–81. Cambridge: MIT Press.

Kamp, Hans, and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-1616-1.

Keenan, Edward L. 2002. Some properties of natural language quantifiers: Generalized quantifier theory. *Linguistics and Philosophy* 25(5/6): 627–654. https://www.jstor.org/stable/25001867.

Keenan, Edward L., and Jonathan Stavi. 1986. A semantic characterization of natural language determiners. *Linguistics and Philosophy* 9(3): 253–326. https://doi.org/10.1007/BF00630273.

Kibble, Rodger. 1997. Complement anaphora and dynamic binding. *Semantics and Linguistic Theory* 7: 258–275. https://doi.org/10.3765/salt.v7i0.2783.

Knowlton, Tyler. 2021. *The psycho-logic of universal quantifiers*, PhD dissertation, University of Maryland. https://doi.org/10.13016/fdr8-3qqh.

Knowlton, Tyler, Tim Hunter, Darko Odic, Alexis Wellwood, Justin Halberda, Paul Pietroski, and Jeffrey Lidz. 2021a. Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences* 1: 134–144. https://doi.org/10.1111/nyas.14618.

Knowlton, Tyler, Paul Pietroski, Alexander Williams, Justin Halberda, and Jeffrey Lidz. 2021b. Determiners are "conservative" because their meanings are not relations: Evidence from verification. *Semantics and Linguistic Theory* 30: 206–226. https://doi.org/10.3765/salt.v30i0.4815.

Knowlton, Tyler, Paul Pietroski, Justin Halberda, and Jeffrey Lidz. 2022a. The mental representation of universal quantifiers. *Linguistics and Philosophy* 45: 911–941. https://doi.org/10.1007/s10988-021-09337-8.

Knowlton, Tyler, John Trueswell, and Anna Papafragou. 2022b. New evidence for the unlearnability of non-conservative quantifiers. In *Proceedings of the 23rd Amsterdam Colloquium*, 367–374.

Knowlton, Tyler, and Florian Schwarz. Forthcoming. "Every" provides an implicit comparison class when "each" does not. In *Proceedings of the 47th Annual Penn Linguistics Conference*.

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

Lasersohn, Peter. 2021. Common nouns as modally non-rigid restricted variables. *Linguistics and Philosophy* 44(2): 363–424. https://doi.org/10.1007/s10988-019-09293-4.

Lepore, Ernest, and Kirk Ludwig. 2007. *Donald Davidson's truth-theoretic semantics*. Oxford: Oxford University Press.

Lidz, Jeffrey, Paul Pietroski, Justin Halberda, and Tim Hunter. 2011. Interface transparency and the psychosemantics of most. *Natural Language Semantics* 19(3): 227–256. https://doi.org/10.1007/s11050-010-9062-6.

Ludlow, Peter, and Sašo Zivanović. 2022. Language, form, and logic. In *Pursuit of natural logic's holy grail*, Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199591534.001.0001.

Lewis, David. 1970. General semantics. *Synthese* 22: 18–67. https://doi.org/10.1007/BF00413598.

Miller, George A., and Noam Chomsky. 1963. Finitary models of language users. In *Handbook of mathematical psychology*, ed. D. Luce, 2–419. New York: Wiley.

Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In *Approaches to natural language*, eds. K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes. Vol. 49 of *Synthese library*. Dordrecht: Springer. https://doi.org/10.1007/978-94-010-2506-5_10.

Mostowski, Andrzej. 1957. On a generalization of quantifiers. *Fundamenta Mathematicae* 44(2): 12–36. https://doi.org/10.4064/fm-44-1-12-36.

Moxey, Linda M., and Anthony J. Sanford. 1993. *Communicating quantities: A psychological perspective*. London: Lawrence Erlbaum Associates.

Nouwen, Rick. 2003. Complement anaphora and interpretation. *Journal of Semantics* 20(1): 73–113. https://doi.org/10.1093/jos/20.1.73.

Nouwen, Rick. 2010. What's in a quantifier? In *The linguistics enterprise: From knowledge of language to knowledge in linguistics*, eds. M. Everaert, T. Lentz, H. De Mulder, Ø. Nilsen, and A. Zondervan, 235–256. Amsterdam: John Benjamins.

Odic, Darko, Hee Yeon Im, Robert Eisinger, Ryan Ly, and Justin Halberda. 2016. PsiMLE: A maximum-likelihood estimation approach to estimating psychophysical scaling and variability more reliably, efficiently, and flexibly. *Behavioral Research Methods* 48: 445–462. https://doi.org/10.3758/s13428-015-0600-5.

Pasternak, Robert, and Uli Sauerland. 2022. German measurement structures: Case-marking and non-conservativity. *Journal of Comparative Germanic Linguistics* 25(2): 221–272. https://doi.org/10.1007/s10828-022-09134-y.

Paterson, Kevin B., Anthony J. Sanford, Linda M. Moxey, and Eugene Dawydiak. 1998. Quantifier polarity and referential focus during reading. *Journal of Memory and Language* 39(2): 290–306. https://doi.org/10.1006/jmla.1998.2561.

Pietroski, Paul. 2005. Meaning before truth. In *Contextualism in philosophy: Knowledge, meaning, and truth*, eds. G. Peter and G. Preyer, 255–302. Oxford: Oxford University Press.

Pietroski, Paul. 2018. *Conjoining meanings: Semantics without truth values*. Oxford: Oxford University Press.

Pietroski, Paul, Jeffrey Lidz, Tim Hunter, and Jeffrey Halberda. 2009. The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language* 24(5): 554–585. https://doi.org/10.1111/j.1468-0017.2009.01374.x.

Ramotowska, Sonia. 2022. *Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences*, PhD dissertation, University of Amsterdam dissertation.

Romoli, Jacopo. 2015. A structural account of conservativity. *Semantics-Syntax Interface* 2(1): 28–57. https://pure.ulster.ac.uk/en/publications/a-structural-account-of-conservativity-3.

Russell, Bertrand. 1905. On denoting. *Mind* 14(56): 479–493. https://doi.org/10.1093/mind/XIV.4.479.

Sanford, Anthony J., Linda M. Moxey, and Kevin B. Paterson. 1996. Attentional focusing with quantifiers in production and comprehension. *Memory & Cognition* 24: 144–155. https://doi.org/10.3758/BF03200877.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2): 461–464. https://doi.org/10.1214/aos/1176344136.

Smith, J. David, Jonathan Schull, Jared Strote, Kelli McGee, Roian Egnor, and Linda Erb. 1995. The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *Journal of Experimental Psychology. General* 124: 391–408. https://doi.org/10.1037/0096-3445.124.4.391.

Smith, J. David, Wendy E. Shields, Jonathan Schull, and David A. Washburn. 1997. The uncertain response in humans and animals. *Cognition* 62(1): 75–97. https://doi.org/10.1016/S0010-0277(96)00726-3.

Spenader, Jennifer, and Jill de Villiers. 2019. Are conservative quantifiers easier to learn? Evidence from novel quantifier experiments. In *Proceedings of the 22nd Amsterdam Colloquium*, ed. Julian J. Schlöder Dean, *McHugh, and Floris Roelofsen*, 504–512.

Stabler, Edward P. 2001. Minimalist grammars and recognition. In *Linguistic form and its computation*, eds. C. Rohrer, A. Roßdeutscher, and H. Kamp, 327–352. Stanford: Stanford Center for the Study of Language and Information.

Stabler, Edward P. 2013. The epicenter of linguistic behavior. In *Language down the garden path: The cognitive and biological basis of linguistic structures*, eds. M. Sanz, I. Laka, and M. Tenenhaus, 316–323. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199677139.003.0018.

Steedman, Mark. 2000. *The syntactic process*. Cambridge: MIT Press.

Steinert-Threlkeld, Shane, and Jakub Szymanik. 2019. Learnability and semantic universals. *Semantics and Pragmatics* 12(4): 1–39. https://doi.org/10.3765/sp.12.4.

Stone, M. 1979. Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B, Methodological* 41(2): 276–278. https://doi.org/10.1111/j.2517-6161.1979.tb01084.x.

Szabolcsi, Anna. 1997. Strategies for scope taking. In *Ways of scope taking*, ed. A. Szabolcsi, 109–154. Dordrecht: Springer. https://doi.org/10.1007/978-94-011-5814-5_4.

Westerståhl, Dag. 2012. Explaining quantifier restriction: Reply to Ben-Yami. *Logique Et Analyse* 55(217): 109–120. https://www.jstor.org/stable/44085104.

Westerståhl, Dag. 2019. Generalized quantifiers. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta. Metaphysics Research Lab, Stanford University (winter 2019 ed.). https://plato.stanford.edu/archives/win2019/entries/generalized-quantifiers.

Whitney, David, and Allison Yamanashi Leib. 2018. Ensemble perception. *Annual Review of Psychology* 69: 105–129. https://doi.org/10.1146/annurev-psych-010416-044232.

Zosh, Jennifer M., Justin Halberda, and Lisa Feigenson. 2011. Memory for multiple visual ensembles in infancy. *Journal of Experimental Psychology. General* 140(2): 141–158. https://doi.org/10.1037/a0022925.

Zuber, Richard, and Edward L. Keenan. 2019. A note on conservativity. *Journal of Semantics* 36(4): 573–582.

Zulaica-Hernández, Iker. 2018. Complement anaphora in Spanish: Proportional references and discourse relations. *Journal of Psycholinguistic Research* 47(2): 449–466. https://doi.org/10.1007/s10936-017-9527-6.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.