Nat Lang Semantics (2011) 19:227–256
DOI 10.1007/s11050-010-9062-6

Interface transparency and the psychosemantics of *most*

Jeffrey Lidz · Paul Pietroski · Justin Halberda · Tim Hunter

Published online: 26 April 2011

Abstract This paper proposes an Interface Transparency Thesis concerning how linguistic meanings are related to the cognitive systems that are used to evaluate sentences for truth/falsity: a declarative sentence S is semantically associated with a canonical procedure for determining whether S is true; while this procedure need not be used as a verification strategy, competent speakers are biased towards strategies that directly reflect canonical specifications of truth conditions. Evidence in favor of this hypothesis comes from a psycholinguistic experiment examining adult judgments concerning 'Most of the dots are blue'. This sentence is true if and only if the number of blue dots exceeds the number of nonblue dots. But this leaves unsettled, e.g., how the second cardinality is specified for purposes of understanding and/or verification: via the *non*blue things, given a restriction to the dots, as in '|{x: Dot(x) & ~Blue(x)}|'; via the *blue* things, given the same restriction, and *subtraction* from the *number* of dots, as in '|{x: Dot(x)}| − |{x: Dot(x) & Blue(x)}|'; or in some other way. Psycholinguistic evidence and psychophysical modeling support the second hypothesis.

Keywords Analog magnitude · Approximate number system · Semantics–cognition interface · Number · Quantification · Mathematics · *Most* · Language processing · Language development

Guidelines for testing human research subjects were followed as certified by the Johns Hopkins University and the University of Maryland Institutional Review Boards. Subjects' rights were protected throughout.

J. Lidz (✉) · P. Pietroski
Department of Linguistics, University of Maryland, College Park, MD 20742, USA
e-mail: jlidz@umd.edu

P. Pietroski
Department of Philosophy, University of Maryland, College Park, MD 20742, USA

J. Halberda
Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218, USA

T. Hunter
Department of Linguistics, Yale University, New Haven, CT 06520, USA

1 Introduction: where does meaning make contact with the rest of cognition?

Theories of meaning aim to specify the semantic properties of expressions. It is not obvious what these properties are. But traditionally, theories have been responsive to two basic concerns. First, a semantic theory for a natural language L is often said to be ''empirically adequate'' to the extent that the theory associates declarative sentences of L with truth conditions in accord with speakers' intuitions.[1] Second, such a theory must be compositional, at least in the following sense: the theory assigns ''atomic'' semantic properties to finitely many expressions of L; and for every other expression, its semantic properties are somehow determined by its constituents and their arrangement. In short, the idea is that a semantic theory should compositionally associate sentences with truth conditions. Satisfying this requirement, even for a single language, remains an elusive goal. But there has been progress, with many insights gained.

The catch, as every semanticist knows, is that given *one* compositional specification of truth conditions—say, for sentences of the form 'Most Δs are βs'—it is often easy to construct others. Even given various assumptions about the relevant syntax and its semantic role, there may be many truth-conditionally equivalent representations of the semantic properties exhibited by the expressions of a given language. Among theorists, there is broad agreement that not all such representations are equally good as proposals about how competent speakers understand expressions.[2] Put another way, many theorists suspect that sentential meanings are individuated more finely than truth conditions, and that distinct specifications of truth conditions can suggest empirically distinguishable psychological hypotheses. But justifying specific proposals requires appeal to additional sources of evidence.

In this paper, we focus on one such source: the interface between linguistic expressions and the cognitive systems that provide the information used, in contexts, to evaluate (declarative) sentences for truth/falsity. In particular, one can gain insight into the meaning of the determiner *most* by examining how sentences like *Most of the dots are blue* interface with the visual system. We argue that the meaning of an expression constrains how the visual system can be used to evaluate the truth of that expression, even to the point of blocking computations native to the visual system that would allow for more accurate calculations.

[1] Cf. Davidson (1967) and Montague (1970), each of whom was inspired by Tarski's (1944) specification of a ''materially adequate'' notion of truth for certain invented languages. This leaves room for a pragmatics/semantics distinction, while allowing the use of model-theoretic techniques in describing entailments that competent speakers recognize. It also allows for views according to which truth conditions are unstructured abstracta (e.g., functions from possible worlds to truth-values). But at least to a first approximation: whatever ''meanings'' get assigned to sentences, they determine truth values given the nonlinguistic facts, and are determined by the constituent morphemes given the relevant syntax. For a second approximation that is less tied to truth per se, see Pietroski (2010).

[2] Evans (1981) suggested the potential relevance of many considerations, including aphasias; see also Davies (1987), Peacocke (1986), and Chomsky's (1986) E-language/I-language distinction, echoing Marr (1982) and Church (1941), who distinguished functions (in extension) from ways of computing them.

Extending other work, our conclusion is that competent speakers associate sentences with *canonical specifications* of truth conditions, and that these specifications provide *default verification procedures*. From this perspective, examining how a sentence constrains its verification can provide clues about how speakers specify the truth condition in question. More generally, our data support an *Interface Transparency Thesis (ITT)*, according to which speakers exhibit a bias towards the verification procedures provided by canonical specifications of truth conditions. In conjunction with specific hypotheses about canonical specifications, the ITT leads to substantive predictions, because given available information, the canonical procedure may have to rely on (noisy) input representations that lead to *less* accuracy in judgment, compared with an alternative strategy that is cognitively available to speakers. To foreshadow: if speakers verify *Most of the dots are blue* by comparing the number of blue dots to the result of subtracting this number from the number of dots—as suggested by the specification '|{x: Dot(x) & Blue(x)}|>|{x: Dot(x)}| – |{x: Dot(x) & Blue(x)}|'—this leads to predictable inaccuracies in judgment, thereby confirming the hypothesis that the operation of cardinality subtraction is invoked by the default verification strategy that speakers associate with *most*.

There is nothing new in the idea that grammars (as internalized procedures) generate objects that interface with other domains of perception, action, and cognition. From the earliest days of generative phonology, linguists have been concerned with the relation between phonological, articulatory, and acoustic properties of speech (Jacobson et al. 1952; Liberman et al. 1967; Stevens 1972)—asking about the degree to which phonological properties are constrained by independent systems of articulation and audition, both in the acquired grammar (Liberman and Mattingly 1985; Halle 2002; Poeppel et al. 2008) and in the acquisition process (Kuhl 1993; Werker 1995; Jusczyk 1997). Katz and Fodor (1963) suggested a parallel approach to the study of meaning. But until more recently, the tradition in natural language semantics has been to focus on relations that expressions bear to entities in an idealized model of the world that speakers talk about, as opposed to language-independent representations.

As a long notable exception, Jackendoff (1983, 1990, 2002) has usefully illustrated how theorists can draw conclusions about conceptual structures from linguistic data. Inferences in the opposite direction, however, have been harder to come by (but see Landau and Jackendoff 1993 for one attempt). This difficulty derives in part from the fact that one cannot be sure which conceptual systems interface with the language faculty, and in part from the fact that especially relevant cognitive subsystems have not been adequately described. In the current paper, we focus on the quantificational determiner *most* as a case study of the relation between cognitive and linguistic representations of quantification, comparison, and measurement. Quantificational expressions have been studied extensively and profitably within several disciplines—including linguistics, philosophy, and psychology—making it possible to formulate precise hypotheses about the interface between semantics and cognition in this domain.

Proportional quantifiers like *most* have long been of interest, in part because their contributions to sentential truth conditions cannot be specified in a standard first-order predicate logic of the sort characterized by Tarski (1944); see Rescher (1962).

To accommodate this expressive capacity of natural language, Barwise and Cooper (1981) adopted Generalized Quantifier Theory (Mostkowski 1957), treating quantifiers as expressing relations between sets as in Frege (1884, 1892); see also Higginbotham and May (1981). For example, *most* can be treated as expressing a comparative relation between the cardinalities of two sets—or equivalently, as a function that maps each ordered pair of sets (X, Y) to a truth value as in (1a).[3] Correlatively, (1b) is true iff the toys in the box outnumber the toys that are not in the box.

(1) a. MOST(X, Y) = TRUE iff $|Y \cap X| > |Y - X|$, otherwise FALSE
b. Most of the toys are in the box.

Likewise, 'every', 'some', and 'no' can be associated with the following relations, respectively: $|Y - X| = 0$; $|Y \cap X| > 0$; $|Y \cap X| = 0$. Generalized Quantifier Theory (GQT) thus provides a useful vocabulary for representing natural language quantifiers in a unified way. But for any given quantifier, the theory is silent with respect to the choice among truth-conditionally equivalent specifications of the corresponding second-order relation. Nonetheless, if GQT is correct, *most* indicates a relation **R** such that each competent speaker of English represents **R** in some way. Indeed, each speaker presumably represents **R** in a format that supports at least one evaluation procedure that can interface with cognitive systems which provide representations of the sort required to judge whether **R** is exhibited by the sets in question (e.g., the toys in the box and the other toys). This raises the question of whether all competent speakers represent **R** in a common way, and if so, what that common format is; cf. Hackl (2009).

2 Truth-conditionally equivalent alternatives

Pietroski et al. (2009) focus on the fact that the truth of (2) can be represented in either of the ways shown in (3), letting 'DOT' and 'BLUE' stand for {x: Dot(x)} and {x: Blue(x)}.[4]

(2) Most of the dots are blue.

(3) a. >(|DOT ∩ BLUE|, |DOT – BLUE|)
b. OneToOnePlus(DOT ∩ BLUE, DOT – BLUE)

The relation in (3a), expressed with '>', is exhibited by *cardinalities* of sets (natural numbers). In (3b), by contrast, 'OneToOnePlus' expresses a relation exhibited by

[3] While '>' signifies a relation between cardinalities, '−' does not signify cardinality-subtraction. In (1a), it signifies set-subtraction: 'Y − X' is equivalent to '{x: (x ∈ Y) & ~(x ∈ X)}'—though, following Boolos (1998), one can eschew the appeal to sets and speak of the 'Ys minus any Xs'.

[4] Here, we ignore any procedural differences between conjunction/negation of predicates and intersection/subtraction of sets: {x: Dot(x) & Blue(x)} vs. {x: Dot(x)} ∩ {x: Blue(x)}, {x: Dot(x) & ~Blue(x)} vs. {x: Dot(x)} − {x: Blue(x)}.

sets themselves (or their elements). Two sets X and Y (e.g., X being the set of blue dots, and Y being the set of nonblue dots) exhibit this relation iff the elements of X and the elements of Y do not correspond one-to-one, but some *proper subset of* X is such that its elements do correspond one-to-one with the elements of Y. Thus, (3b) captures the idea that pairing each nonblue dot with exactly one blue dot would leave at least one blue dot unpaired with any nonblue dot.

On both analyses, *most* indicates the same relation. But only (3a) specifies this relation in terms of cardinalities. In terms of specifying truth conditions compositionally, (3a) and (3b) are equivalent. Yet they suggest different evaluation procedures. Both require, for the truth of (2), more blue dots than nonblue dots. But as procedures for determining if this requirement is met, (3a) calls for comparing numbers, while (3b) calls for pairing dots.

Hackl (2009) focuses on another kind of contrast, noting that there are alternative formulations of which sets and numbers are compared, even if one assumes that the *most*-relation is to be specified in terms of numbers. In particular, one might replace (3a) with (4), allowing for rational numbers.

(4) $>(|\text{DOT} \cap \text{BLUE}|, \frac{1}{2}|\text{DOT}|)$

Provably, (3a) is truth-conditionally equivalent to (4): the number of blue dots is more than half the number of dots iff there are more blue dots than nonblue dots. But as a procedure for determining if this truth condition is met, (3a) calls for subtracting the blue things from the dots, in a way that (4) does not; (4) calls for division by two, in a way that (3a) does not; and (4) also calls for computing the cardinality of all the dots, in a way that (3a) does not. This is the kind of contrast we want to consider. We return below to Hackl's reasons for not adopting the specification in (4). But we will suggest the specification in (5),

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

which does call for computing the cardinality of all the dots and subtracting from this number the cardinality of the blue dots.

Given the many truth-conditional equivalences, one wants to know if there is a fact of the matter about which, if any, are better than others. Are they mere notational variants, like the difference between measuring temperature in Fahrenheit or Celsius? Or can at least some of the contrasts be regarded as alternative psychological hypotheses about speakers? We pursue the latter option, taking the position that the meaning of a declarative sentence is not a mere compositionally determined truth condition, even if such conditions are functions from worlds to truth values; cf. Cresswell (1985). We argue that different representations of a truth condition often correspond to interestingly different proposals about how competent speakers specify that truth condition for purposes of canonical verification. But as noted above, and as Hackl (2009) discusses, finding evidence for or against any such proposals requires methods that go beyond the usual ones for eliciting competent speaker intuitions about the truth/falsity of sentences.

3 On verification procedures

The differences we have been talking about concern the operations represented in specifications of truth conditions. We will argue below that certain specifications are semantically privileged: competent speakers represent the truth-conditional contribution of *most* in terms of certain operations, thereby biasing speakers towards the use of algorithms that employ those operations in determining the truth/falsity of sentences like 'Most of the dots are blue'. But this is fully compatible with the fact that given any one specification of a truth condition, there can be *many* methods for determining whether that condition obtains. Indeed, the examples above illustrate this point. Our claim is not that speakers always, or even typically, use canonical specifications of truth conditions as algorithms for determining the truth/falsity of sentences in contexts. Verification obviously depends on the information available in the context at hand.

Consider the range of statements in (6a–d):

(6) a. Rabbits are furry.
b. Chicago has great architecture.
c. Most of the dots are blue.
d. La neige est blanche.

If you want to know whether (6a) is true, you might check some rabbits, or a website. If you want to know whether (6b) is true, you might go to Chicago and look around, or you might read a book. If you want to know whether (6c) is true, you might count if you have the time and opportunity, or you might estimate the relevant cardinalities. Or you might just ask someone else, especially if you are color blind. And of course, if you want to know whether (6d) is true, there are ways of finding out even without understanding the sentence: ask someone who speaks French. But when a speaker understands a sentence and judges it to be true or false in a given context, she presumably does at least two things: she compositionally determines the relevant truth condition, and she determines whether that condition obtains in the context. At least typically, the latter presupposes the former.

Now, you can reliably assess the truth of a sentence by asking your neighbor only if you can treat his response as a reliable indicator of whether that sentence's truth condition obtains. And in the general case, this requires that you know which truth condition this is. Suppose, for illustration, that you understand (2) as in (3a).

(2) Most of the dots are blue.

(3a) $>(|\text{DOT} \cap \text{BLUE}|, |\text{DOT} - \text{BLUE}|)$

If you defer to a neighbor, you are effectively relying on that neighbor to tell you whether the number of blue dots is greater than the number of nonblue dots. But then your *verification procedure*, for deciding whether (2) is true or false, does *not* proceed as follows: determine the number of blue dots, determine the number of

nonblue dots, and figure out if the first number is bigger. Your neighbor may or may not employ this procedure, but you don't. Understanding (2) as in (3a) does not commit you to following any particular procedure for evaluating (2). In this sense, actual verification procedures may be distinct from any procedures/algorithms that result from specifying truth conditions. But the question one seeks to answer, by whatever verification procedure one uses in the context, is determined by how one understands the sentence.

So in one perfectly fine sense, specifications of truth conditions are indeed verification procedures, even in contexts where these procedures cannot be employed; cf. Dummett (1973), Peacocke (1986), Horty (2007) and references there. If you understand (2) as in (3a), then you presumably know that one *could* determine the truth or falsity of (2) by determining and comparing the relevant cardinalities. In a given context, you might not be able to determine the truth or falsity of (2) in this way; and perhaps in practice, no real person could. The relevant dots might be too far away, or occluded. More generally, the verification procedure that is invoked by specifying a truth condition need not be practical in a given context. But when conditions are favorable, one can figure out if (2) is true by determining and comparing two cardinalities.

Indeed, the experiment presented in Sect. 5 below provides evidence for the following hypothesis:

> *Interface Transparency Thesis (ITT):*
> The verification procedures employed in understanding a declarative sentence are biased towards algorithms that directly compute the relations and operations expressed by the semantic representation of that sentence.

For example, suppose that speakers understand (2) along the lines shown in (5), repeated here.

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

The ITT implies that speakers who specify the truth of (2) in this way are biased towards verification procedures that involve representing the number of blue dots, the number of dots, and the result of subtracting the former from the latter. In this sense, sentence meanings are not verification independent. Rather, a sentence meaning determines an instruction to interfacing systems concerning what information to gather in order to verify the sentence.

In a particular context, this instruction may not be executable; in which case, if evaluation is required, speakers with other resources will try other methods. And for many lexical items, the canonical verification procedures may be atomic; cf. Fodor (2003). We suspect that even for 'dot' and 'blue', there is much to be said. But at least for ''logical'' vocabulary, an old and plausible idea is that lexical meanings provide default verification procedures that speakers use when they can.

As discussed below, we test the ITT by pitting the predicted bias for transparency against computations native to the visual system. We offer evidence that a sentence meaning can lead competent evaluators to ignore relevant information that their visual systems automatically provide—in favor of an alternative procedure that calls

for different information—as if the sentence meaning makes certain evaluation procedures preferable to others, as predicted by the ITT. Before turning to our experimentation, however, it is worth reviewing some earlier results that motivate this general conception of how meaning is related to verification.

4 *Most*: prior results

In Pietroski et al. (2009), we put people in a range of situations that differed in their amenability to a verification procedure for *most* characterized as in (3b), repeated below.

(3b) OneToOnePlus(DOT $\cap$ BLUE, DOT $-$ BLUE)

Consider the displays in Fig. 1. They each contain 10 yellow and eight blue dots, but differ in the degree to which they invite pairing the dots. In Fig. 1a, the dots are scattered randomly on the screen. In Fig. 1b, they are scattered, but in pairs such that the only singleton dots come from the larger of the two sets. In Fig. 1c, the dots are arranged in two columns and 10 rows, with each row consisting of either a mixed pair of yellow and blue dots, or only a singleton yellow/blue dot.

Across many trials, we flashed such displays on a computer screen for 200 ms each and asked people to determine on each trial whether (2) was true.

(2) Most of the dots are blue.

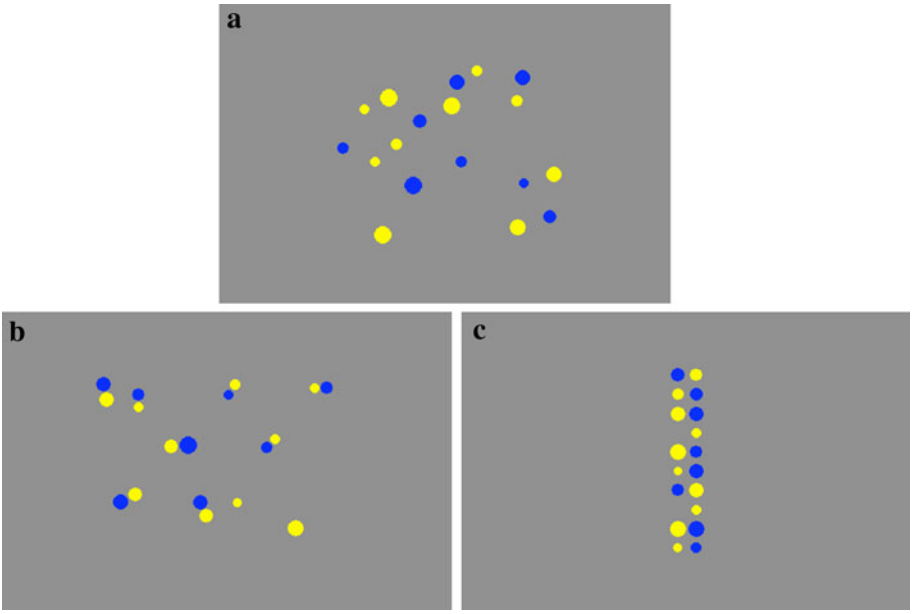


Fig. 1 Displays from Pietroski et al. (2009)

We found that subjects' accuracy was unaffected by manipulating suitability to a verification procedure stated in terms of one-to-one correspondence. And in separate studies, we confirmed that (in 200 ms) people can identify the color of the ''unpaired'' dots in scenes like Fig. 1b, with *better* performance than when asked to evaluate (2). This suggests that speakers do not understand (2) in terms of one-to-one correspondence. If they did, scenes that invite a OneToOnePlus verification procedure should have made verification easier. But performance across scenes revealed no such difference.

More positively, and more importantly, our data also provided evidence that subjects used approximate representations of numerosity to evaluate (2). Responses showed the behavioral signature of the *Approximate Number System (ANS)*, an evolutionarily ancient piece of cognitive machinery that is shared throughout the animal kingdom and does not require explicit training with number in order to develop (Dehaene 1997; Feigenson et al. 2004). The ANS generates an approximate representation of the number of items in a set, in accord with *Weber's law*: the discriminability of two quantities is a function of their ratio. We found that for the adult subjects in Pietroski et al. (2009), the probability of evaluating (2) correctly was a function of the ratio of the number of blue dots to the number of nonblue (yellow) dots. Moreover, not only did performance improve with easier ratios, the specific shape of this improvement fit an independently confirmed psychophysical model of ANS representations (Pica et al. 2004; Halberda et al. 2008), with R^2 values greater than .85 even for scenes like Fig. 1b. See Appendix 1 for details of the model.

This fit confirmed the hypothesis that ANS representations were involved in evaluating (2), and hence that at least in some conditions, the numerical content required to verify a claim like (2) is provided by the ANS. But while this system was implicated in verification, suggesting that subjects understood (2) as a claim to be evaluated by comparing cardinalities, nothing yet follows about the cardinalities compared. Likewise, nothing follows about how the cardinalities are represented, or where approximation is involved. But let us set aside the question of whether the cardinality comparisons for *most* concern precise cardinalities or ANS analogs, and recall the distinction highlighted in Hackl (2009), repeated in (7).

(7) a. $>(|\text{DOT} \cap \text{BLUE}|, |\text{DOT} - \text{BLUE}|)$
b. $>(|\text{DOT} \cap \text{BLUE}|, 1/2|\text{DOT}|)$

To distinguish these candidate meanings for *most*, Hackl asked people to evaluate sentences like (8a) and (8b) in an experimental paradigm he called ''self-paced counting.''

(8) a. Most of the dots are blue.
b. More than half of the dots are blue.

In this paradigm, inspired by studies of independent phenomena that used self-paced reading tasks, each participant sees a series of uncolored circles on a computer screen. Pressing the space bar causes some of the dots to become (or be revealed as) red or blue. Pressing the space bar again causes those dots to return to being

uncolored, and a subsequent subset becomes colored. This continues until the participant indicates his judgment, by pressing an appropriate button, as to whether the test sentence is true or false. Participants were told to respond as quickly and accurately as possible. The idea was to get a measure of how much information people need to make a confident judgment. Hackl found that while accuracy and overall response times for sentences like (8a) and (8b) were not significantly different, reaction times *between* successive space bar presses were significantly faster when the test sentence included *most* as opposed to *more than half*. He concluded that the verification procedures, and thus the specifications of the common truth condition, differ in some way.

Hackl went on to offer and defend some plausible speculations about why ''the strategy triggered by *most* is better suited for the way information is uncovered in these screens'' (p. 89). An adequate account of how *most* is related to *more* will need to accommodate his findings, along with his crosslinguistic data. But here we want to stress his use of an experimental technique designed to test for differing verification strategies corresponding to distinct representations of a common truth condition. In what follows, we will assume that the meaning of *most* is specified in terms of a relation between cardinalities (as opposed to one-to-one pairing of individuals in a set), but *not* in terms that invite comparison of the ''intersection'' cardinality (|DOT $\cap$ BLUE|) with *half of* the ''restricted domain'' cardinality (|DOT|). We will also assume that at least for purposes of verification when all the dots are presented rapidly and at once, the cardinalities to be compared are provided by the ANS.

While these are important steps forward, many questions remain. Evidence against the specifications (9b) and (9c) is not yet evidence in favor of (9a).

(9) a. >(|DOT $\cap$ BLUE|, |DOT $-$ BLUE|)
b. >(|DOT $\cap$ BLUE|, 1/2|DOT|)
c. OneToOnePlus(DOT $\cap$ BLUE, DOT $-$ BLUE)

There are other possibilities. In particular, while there may be no viable alternative to computing the intersection cardinality, the ''contrast'' cardinality (|DOT $-$ BLUE|) might be computed in various ways, depending on the context. Recall (5).

(5) >(|DOT $\cap$ BLUE|, |DOT| $-$ |DOT $\cap$ BLUE|)

One might think instead that given blue dots and yellow dots, and no others, subjects surely computed the number of yellow dots and took the result to be the number of nonblue dots. But if so, they computed and used (for verification) the cardinality of a set—the yellow dots—not represented by any word in the target sentence.

One can, of course, hypothesize that speakers understand *most* as indicated in (9a), and use information available in the two-color context to infer that (9a) is true iff there are more blue dots than yellow dots. Evidence that people do naturally evaluate sentences in this fashion, when not forced to do so, would be prima facie evidence against the Interface Transparency Thesis. For the ITT predicts a bias in favor of the canonical verification procedure, as opposed to any context-specific procedure, like using the number of yellow dots as the number of nonblue dots.

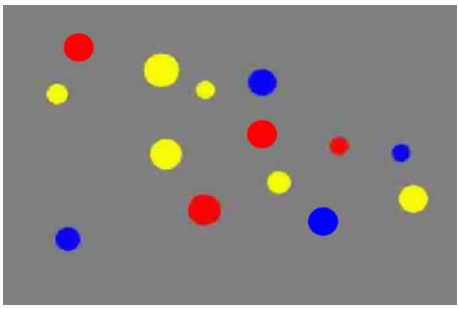
More generally, if representing the ~βs as such requires a context-specific inference—if people need to represent the ~βs in some ''more positive'' way and infer that they are the ~βs—then one might suspect that the meaning of *most* can be correctly specified without appeal to negation. Perhaps one could represent the dots minus the blues, or the dots minus the blue dots, without representing the not-blues as such. But then the difference between (9a) and (5), as hypothesized default verification strategies, is rather subtle. Are speakers biased towards (i) subtracting some dots from others, and computing two cardinalities, neither of which is the total number of dots; or (ii) using the intersection cardinality twice, and subtracting this number from the total number of dots? To repeat, the meaning of a sentence does not determine the verification procedure used in all contexts. But if the ITT is true, verification can provide a window into meaning, given independently confirmed claims about relevant aspects of cognition. Correlatively, having speakers evaluate sentences in controlled situations where the relevant aspects of non-linguistic cognition are relatively well understood can simultaneously test the ITT and specific hypotheses about how speakers specify the truth conditions of their sentences. In the case at hand, suppose that representing the number of yellow dots in an array is a rapid and automatic computation of the visual system, and likewise for the (total) number of dots in the array. And suppose that the number of yellow dots is *not* used to evaluate 'Most of the dots are blue', in an array with blue dots and yellow dots, while the number of dots *is* used. Evidence for such claims could be used to help confirm the ITT. For if the relevant interface system fails to use certain information that it automatically computes, when faced with an evaluative task, then it becomes plausible that the representation of the sentence must be responsible for such a failure (cf. Kahneman and Tversky 1973). It is precisely this argument that we undertook to pursue in detail in the experiment presented below.

5 The experiment

5.1 Background: Interface constraints imposed by the visual system

Imagine that a listener was shown a briefly flashed display of dots of many colors (Fig. 2) and was asked to assess whether (10) is true of the display.

Fig. 2



(10) Most of the dots are blue

(11) a. >(|DOT ∩ BLUE|, |DOT − BLUE|)
b. >(|DOT ∩ BLUE|, |DOT| − |DOT ∩ BLUE|)

The specification (11a) invites a verification procedure that attends to and enumerates the blue dots, likewise for the nonblue dots, and compares the two numbers. The specification (11b) invites a verification procedure that attends to and enumerates the dots, likewise for the blue dots, subtracts the latter from the former, and then compares the result to the number of blue dots. The difference between these two verification procedures lies in whether the nonblue dots are *selected*, with a subsequent step of estimating their cardinality, as in (11a), or whether this cardinality is *computed,* as in (11b). This leaves open just how selection is achieved. The important distinction will be whether the ANS is employed to estimate the nonblue dots or the dots—and correlatively, whether the second argument of '>' (i.e., the number compared with |DOT ∩ BLUE|) is an estimate of nonblue dots or a computation performed on two estimations.

We can therefore ask whether it is psychologically possible to directly select and enumerate both the blue dots (as both computations would require) and the nonblue dots (as required only for (11a)). Even without us briefly flashing the array, the reader can likely experience that selecting only the blue dots from among all of the dots is easy. Research on adults' ability to search for a colored item among colored distractors has shown this to be the case; *blue*, and all other categorizable colors, works as an early visual feature that can be found very quickly in a visual scene when the distractors are of saliently different colors, as they are in Fig. 2 (Wolfe 1998; Halberda et al. 2006). But similar research also reveals that a set defined by a negation of an early visual feature or by a disjunctive combination of early visual features (e.g. dots that are either yellow OR red) is *not* easily selectable. Adults are unable to rapidly search all items in an array in order to find all the items that are either yellow or red (Wolfe 1998; Treisman and Gormican 1988; Treisman and Souther 1985). This calls into question the viability of having the meaning in (11a) map directly onto a verification procedure which requires listeners to directly attend and enumerate both the blue and the nonblue dots for purposes of ordinal comparison. Because the nonblue dots are a heterogenous set, they cannot be attended directly. Moreover, building up the nonblue dots by constructing a disjunctive combination of all nonblue sets is also not a straightforward visual computation. Listeners simply would not be able to directly attend the heterogeneous set of nonblue dots.

But, looking at Fig. 2, it seems that we *can* assess whether most of the dots are blue, and so the question becomes (i) how we are accomplishing this and (ii) whether (11b) provides a more natural verification procedure. Additional evidence from the psychological literature is helpful in this regard.

Halberda et al. (2006) have demonstrated that adults can use the Approximate Number System to estimate the cardinality of up to three sets in parallel. On each trial in Halberda et al. (2006), participants were shown a brief flash that contained from 1 to 6 colors of dots randomly scattered on a black background, similar to

Fig. 2. Either before or after the flash, participants were asked to approximately enumerate only one of the sets (either the superset of all dots irrespective of color, or a particular color subset). On a ''Probe After'' trial, where subjects did not know which set to report until after the flash had gone, the most likely strategy is to enumerate as many sets as possible and hope that one of those sets would be the one asked. By comparing performance on ''Probe After'' to ''Probe Before'' trials, Halberda et al. (2006) were able to estimate how many sets adults could enumerate from a single flash. Results suggested that adults *always* attend and enumerate the superset of all dots. In addition to the superset, adults could also attend and enumerate some of the color subsets on multi-color trials. The typical adult appeared to enumerate the superset of all dots and two of the color subsets, but no more. For example, shown the flash depicted in Fig. 2, a typical adult would know that there had been approximately 14 total dots, and perhaps that there had been approximately 4 red dots and approximately 6 yellow dots, but nothing more.

That adults can enumerate multiple sets from a single flash using the Approximate Number System highlights the potential relevance of this system for verification procedures associated with natural language quantifiers like *most*. A meaning like (11a), translated directly into a verification procedure, is implausible because it involves selecting a heterogeneous set. However, this meaning invites the transformation in (12), wherein the set of nonblue dots is constructed by summing the cardinalities of each color subset comprising the nonblue dots.

(12) $>(|\text{DOT} \cap \text{BLUE}|, |\text{DOT} \cap \text{RED}| + |\text{DOT} \cap \text{YELLOW}|)$

However, such a transformation would be useful only when the display contains no more than three colors, given Halberda et al.'s observation of a 3-set limit on early visual attention and working memory. That is, to verify this meaning would require the visual system to attend the color subset of blue dots, the color subset of red dots and the color subset of yellow dots. If there are only these three colors present in the array, then an addition of yellow and red dots would provide the listener with the number of nonblue dots, which could then be compared to the number of blue dots to yield a truth value. But, because adult humans appear to be limited to enumerating only up to three sets at once, this verification procedure, and hence the meaning in (11a), becomes less plausible as the number of color subsets increases.

A meaning like (11b), however, is straightforwardly verified with these resources, since the sets required for its verification (one color plus the superset) are easily and automatically attended by the visual system. Moreover, this meaning does not become less plausible as the number of color subsets increases.[5] That is, to

[5] Halberda et al. (2006) found no reduction in enumeration accuracy for adults' ability to enumerate a color subset when the number of colors in the distractor subsets increased. Performance was the same for enumerating the blue dots if there were no other colors present, or if there were blue and yellow dots; blue, yellow, and red dots; or even blue, yellow, red, green, purple, and cyan dots. Also, Halberda et al. (2006) found no cost for estimating the cardinality of the superset of all dots as the number of colors in the stimulus increased. So, enumeration of the blue dots and the superset appear to be unaffected by increasing the number of color subsets, making the meaning expressed in (11b) plausible as the number of sets increases and the meaning in (11a)/(14) implausible.

verify a meaning like (11b) would require first enumerating the superset of all dots and the color subset of blue dots. The next step would involve a subtraction of these two values to calculate the number of nonblue dots. The final step would compare the number of blue dots to the number of nonblue dots to yield a truth value. Because only the superset and one color subset need be attended, the meaning in (11b), along with its associated verification procedure, is psychologically plausible, no matter how many color subsets there are, so long as it is possible to perform the subtraction and comparison computations.

In order to determine whether the canonical specification of the meaning associated with *most* is like (11a) or like (11b), we asked adult participants to verify whether most of the dots in an array were blue across many trials where we randomly varied the number of colors in the array. If participants verify *most* via the meaning expressed in (11a), then we expect accuracy to decline as the number of colors in the array increases. On the other hand, if participants verify *most* via the meaning expressed in (11b), then we expect the number of colors to have no impact on their responses; see Sect. 5.6.

5.2 Method

We used a common visual identification paradigm to evaluate the underlying meaning for *most*.

5.3 Participants

Twelve naive adults with normal vision each received $5 for participation.

5.4 Materials and apparatus

Each participant viewed 400 trials on an LCD screen (27.3 × 33.7 cm). Viewing distance was unconstrained, but averaged approximately 50 cm. The diameter of a typical dot subtended approximately 0.8° of visual angle from a viewing distance of 50 cm.

5.5 Design and procedure

On each trial, subjects saw a 150 ms display containing dots of at least two colors and at most five colors (blue, yellow, red, cyan, magenta). Blue dots were present on every trial. Subjects were asked to answer the question ''Are most of the dots blue?'' for each trial. The number of dots of each color varied between 5 and 17. Whether the blue set represented more than half of the total number of dots (that is, whether the correct answer to ''Are most of the dots blue?'' was Yes or No) was randomized. Subjects answered ''Yes'' or ''No'' by pressing buttons on a keyboard. Within each trial type (i.e., 2–5 colors), the ratio of blue to nonblue dots varied between five possible ratios (1:2, 2:3, 3:4, 5:6, and 7:8). Within each of these ratio bins the blue set was the larger set on half of the trials.

Half of the trials for each trial type (2–5 colors) for each ratio bin were ''dot size-controlled'' trials on which, while individual dot sizes varied, the size of the average blue dot was equal to the size of the average nonblue dot. On dot size-controlled trials the set with the larger number of dots would also have a larger total area on the screen (i.e., more total blue pixels when blue was the larger set). The other half of the trials were ''area-controlled'' trials in which individual dot sizes varied and the total number of blue and nonblue pixels on the screen was equated (i.e., smaller blue dots on average when blue was the larger set). On both dot size-controlled and area-controlled trials individual dot sizes were randomly varied by up to 35% of the set average. This discouraged the use of individual dot size as a proxy for number.

All trials were randomly shuffled such that number of colors (2–5), correct answer (Yes/No), ratio bin (1:2–7:8), and stimulus type (dot size-controlled, area-controlled) varied randomly during the experiment.

5.6 Predictions

If subjects rely on the imprecise cardinality representations of the ANS, then accuracy should decline as a function of ratio, and should be well fit by a psychophysical model of the ANS. With respect to the question of whether (11a) or (11b) underlies the meaning of *most*, we consider two hypotheses. First, if subjects determine the set of nonblue dots by determining the cardinality of each subset and then summing the nonblues together, in keeping with algorithm (12), we predict that subjects should succeed at the task when there are two and perhaps three colors on the screen but that performance should rapidly decline for higher numbers of colors (we will call this the ''selection hypothesis''). This prediction derives from the observation from Halberda et al. (2006) that at rapid presentation rates, the visual system can accurately track a maximum of three sets. The second hypothesis, which we will call the ''subtraction hypothesis,'' holds that the cardinality of the set of nonblue dots is determined by subtracting the cardinality of the focused set (the blue set) from the cardinality of the superset (the dots), as in (11b). Consequently, the computation determining the truth of a *most*-statement is predicted to be identical across all trial types. Since only two sets ever need to be selected by the visual system, the number of colors should have no impact on responses.

5.7 Results

Results were entirely consistent with the subtraction hypothesis, suggesting that algorithm (11b) reflects the canonical specification of the meaning of *most* and that this algorithm relies on the representations of the ANS. There were no differences across trial types as a function of the number of colors in the display (Fig. 3), and performance on every trial type was well fit by a psychophysics model of the ANS (Table 1; Fig. 4). See Appendix 1 for more details about the psychophysics modeling.

Fig. 3

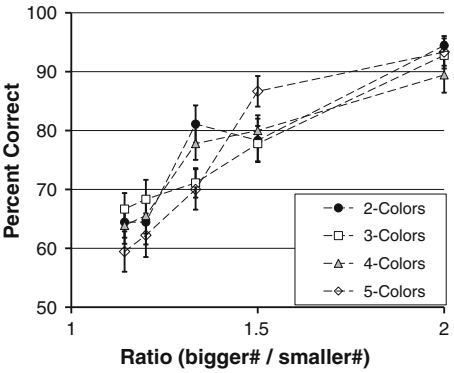
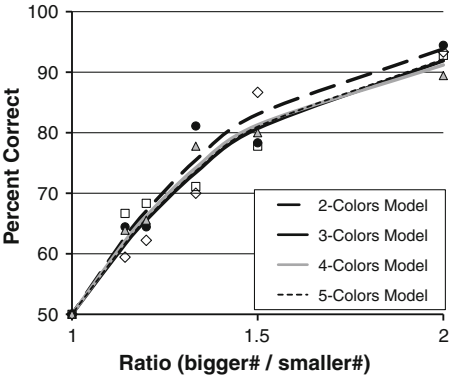


Table 1 Parameter estimates from psychophysical model

Trial type	R^2	Weber fraction	Nearest whole-number ratio
2-Color	.9480	.290	3:4
3-Color	.9586	.320	3:4
4-Color	.9813	.283	3:4
5-Color	.9625	.316	3:4

Fig. 4



A detailed description of the statistics and figures follows. Percent correct for each participant for each ratio was entered into a 4 Trial Type (2-, 3-, 4-, 5-color) × 2 Stimulus Type (dot size-controlled, area-controlled) × 5 Ratio Repeated Measures ANOVA. There was a significant effect of Ratio as subjects did better with easier ratios: $F(4, 44) = 109.092$, $p < .001$; a significant effect of Stimulus Type as subjects did slightly better on dot size-controlled than area-controlled trials: $F(1, 11) = 7.326$, $p < .05$; and most importantly, no effect of Trial Type as subjects did equally

well independent of the number of colors in the stimulus: $F(3, 33) = 0.276$, $p = .842$. Because the small but significant Stimulus Type effect does not bear on the inferences we make about the algorithms involved, we combined performance for each subject for each Ratio and each Trial Type for further analyses.

As can be seen in Fig. 3, while performance declines as a function of Ratio, performance is the same independent of the number of colors in the array.[6] This supports the predictions of the subtraction hypothesis that on every trial type, irrespective of the number of colors in the display, subjects attend the superset of all dots and the focused set (blue dots), enumerate each, and then perform a subtraction in order to calculate the number of nonblue dots, before comparing the number of blue dots to the number of nonblue dots. Obviously, we are not suggesting that this subtraction is a conscious subtraction, and we doubt that subjects are even aware of how they are figuring out what answer to give. The subtraction hypothesis, i.e., the meaning expressed in (11b), is meant to characterize the unconscious computations that underlie the meaning of *most* and allow it to interface with the rest of psychology.

The Approximate Number System is known to contain both the representational and the computational machinery necessary to represent imprecise cardinalities, perform subtractions of these cardinalities, and make ordinal comparisons of these cardinalities (Whalen et al. 1999; Dehaene 1997; Feigenson et al. 2004; Brannon et al. 2006). Thus, the ANS itself may be capable of implementing the entire algorithm expressed in (11b). A first step in evaluating whether this is the case is to see if performance on each Trial Type can be fit by a computational model of the ANS. See Appendix 1 for more details.

We rely on a classic psychophysical model that has been used by labs other than our own, indicating its acceptance in the literature (e.g., Pica et al. 2004). The average percent correct at each ratio across subjects is modeled for each trial type as a function of increasing *Weber ratio* (larger set/smaller set, or n2/n1). Each numerosity is represented as a Gaussian random variable (i.e., X2 & X1) with means n2 & n1 and standard deviations equal to the critical *Weber fraction* (w) * n. Subtracting the Gaussian for the smaller set from the larger returns a new Gaussian that has a mean of n2–n1 and a standard deviation of $w\sqrt{n1^2 + n2^2}$ (simply the difference of two Gaussian random variables). Percent correct is then equal to the area under the resulting Gaussian curve which is to the right of zero, computed as (13):

$$(13) \quad \frac{1}{2}\ \mathrm{erfc}\left(\frac{n_1 - n_2}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}}\right)$$

The one free parameter in this equation is the Weber fraction (w). This parameter determines percent correct for every Weber ratio (n2/n1). The mean of subject means for percent correct at each of the five ratio bins and the theoretically determined origin of the function (50% correct at Ratio = 1, where the number of

[6] Throughout the analyses, ratios will be displayed as the Weber ratio between the two sets (Weber ratio = bigger #/smaller #). This is important as it allows performance to be fit by a psychophysical model of the ANS.

blue dots and nonblue dots would in fact be identical) were fit using this psychophysical model. As can be seen in Fig. 4, the fits for all four trial types (2–5 colors) fell directly on top of one another. Table 1 summarizes the R^2 values, the estimated Weber fraction, and the nearest whole-number translation of this fraction for each fit.

These R^2 values suggest agreement between the psychophysical model of the ANS and subjects' performance in the experimental task (R^2 values > .9). The Weber fraction on these trial types confirms our earlier result that participants rely on the representations of the ANS to evaluate *most*.

The Weber fraction is expected to be approximately .11 to .14 for adults in number discrimination tasks, i.e. *more*-tasks, (Halberda and Feigenson 2008; Pica et al. 2004) and to range from .14 to .35 in adults when subjects are translating these representations into whole-number values (Halberda et al. 2006; Whalen et al. 1999). Our estimate of a Weber fraction of approximately .3 for all four trial types suggests that subjects may be translating the representations of the ANS into whole number values before evaluating *most* (see also Pietroski et al. 2009). That is, shown an array of 28 total dots, 16 of which are blue, these subjects may activate the ANS representations for 28 and 16, perform an ANS subtraction to represent the 12 nonblue dots, and translate the values 12 and 16 into whole-number estimates *twelve* and *sixteen* for purposes of evaluating *most*. Another possibility is that the entire computation is done within the ANS without ever translating into whole-number values. In such a model, the dual operations of subtraction and ordinal comparison may each contribute to determining the Weber fraction. Further work will be necessary to tease these two possibilities apart.

6 General discussion

We found no change in participants' ability to evaluate *most* as a function of the heterogeneity of a display. Rather, participants' performance at evaluating *most* for a wide range of ratios across all trial types was best fit by a model of the ANS whereby participants rely on a subtraction to compute the cardinalities of the sets to be compared. These results inform our understanding of how the meaning of *most* interfaces with the psychological mechanisms that provide numerical content, and lays the groundwork for further investigation of the interface between language and number.

More generally, our research addresses the relation between the units of meaning out of which truth conditions are built and the verification procedures that determine truth values. We have argued that semantic representations can be transparently mapped into verification procedures. When two equivalent semantic representations are being compared, as with the truth-conditionally equivalent (11a) and (11b), repeated here, examining the psychological processes implied by directly implementing these meanings as verification procedures can provide decisive evidence for distinguishing them.

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

Although these alternatives describe the same truth conditions, the psychological mechanisms required to implement them transparently are quite distinct. Whereas (11b) can be computed across all possible dot-flashing contexts using only the information provided directly by the visual system in concert with the ANS, (11a) is more psychologically brittle. Because it asks for information that cannot be directly provided by the visual system, it requires a context-driven transformation, identifying the set(s) in the context (e.g., (14)) that are appropriate redescriptions of the set of nonblue dots.

(14) $>(|\text{DOT} \cap \text{BLUE}|, |\text{DOT} \cap \text{RED}| + |\text{DOT} \cap \text{YELLOW}|)$

While such a transformation allows for accurate verification in contexts containing less than three colors of dots, this transformation would be less effective in contexts containing more than three colors of dots. However, as we have seen, the number of colors of dots played no role in explaining participants' *most* judgments, casting doubt on the hypothesis that they use a verification procedure based on (11a) in any context.

What may be surprising to consider, however, is that, in the context of only blue dots and dots of one other color (i.e., a 2-color trial), the expression in (11a)/(14) would lead to more accurate performance in evaluating *most* than the expression in (11b). Specifically, with only two colors present in the array, (11a)/(14) is a more accurate verification procedure within the ANS than (11b).

This last point requires elaboration. Various studies have demonstrated that adults can rely on the ANS to make ordinal judgments (more/less) between two sets whether they are presented serially or in parallel (Dehaene 1997). In all cases, the estimated Weber fraction for adults is considerably better than the 3:4 value we found here for all trial types. Typically, the Weber fraction for adults is closer to 7:8 (Dehaene 1997) and may be as high as 9:10 (Halberda and Feigenson 2008; Piazza et al. 2003), and children as young as 4 years have a Weber fraction of at least 3:4 (Halberda and Feigenson 2008). For this reason, if participants had simply selected the set serving as the first argument of the > relation (e.g., the blue dots) and the set serving as the second argument (e.g., yellow dots) directly on a 2-color trial and compared these using the ANS, as previous work has demonstrated they can, we would have observed a Weber fraction of at least 7:8. That participants' performance is far below this suggests that they are relying on a representation of *most* like the expression in (11b), even when there are more accurate, truth-conditionally equivalent, methods of verification available (i.e., (11a)/(14)).

This last observation provides the strongest evidence for the Interface Transparency Thesis introduced in Sect. 3 above. Even when there is a more precise algorithm that is native to the interface system, semantic judgments are driven by algorithms that transparently compute the relation expressed in the meaning. The semantic representation of *most* (i.e., its canonical specification) thus plays a determinative role in identifying the verification procedure for a sentence containing that word, at least when a transparent verification procedure is available. The fact that participants never employ the verification procedure most naturally associated with the canonical specification in (11a), even when that verification procedure is

positively invited by the context and would yield the most accurate estimate of the truth of the expression, provides compelling evidence against (11a) being the meaning of the expression.[7]

Of course, the fact that the verification procedure reflects precisely the structure of the meaning has a natural explanation in the set of circumstances in which *most* applies. The second argument of the > relation is not guaranteed by the world to have only one easily selectable property (e.g., yellowness), and, because of the 3-set limit on parallel enumeration (Halberda et al. 2006), the limitations of the psychological machinery would lead to drastically reduced performance as the heterogeneity of the remainder set increased. Thus, the most general verification procedure would be one that can apply independent of whether such a property exists in a particular circumstance. A verification procedure whose applicability varied as a function of contingent properties of the world would be less reliable than one which could apply across all circumstances.

Finally, we wish to reiterate that treating semantic hypotheses as psychological hypotheses makes available certain kinds of evidence that are unavailable to semantic theories concerned only with compositionally determined truth conditions, and moreover, that such evidence enables us to distinguish otherwise equivalent hypotheses. We have argued that semantic hypotheses are best viewed as psychological hypotheses about the mental representations involved in defining the truth conditions for a sentence. These representations provide canonical specifications of meaning that can be mapped transparently to verification procedures involving the integration of linguistic information with information from adjacent cognitive systems. Knowing what information these systems can and cannot provide places constraints on the verification procedures. And these constraints can, in turn, be used to examine the semantic representations themselves, enabling us to distinguish semantic hypotheses that are otherwise equivalent. We believe that this approach has so far been fruitful for distinguishing hypotheses about the meaning of *most*, but we view the demonstration that such questions can be precisely asked and plausibly answered as the more significant contribution of this work, opening the door for progress in the field of psychosemantics.

Acknowledgements J.H. and J.L. devised the task; J.H., J.L., P.P., and T.H. defined the trial types of interest; T.H. implemented and ran the experiment; J.H. analyzed the data; J.L. wrote the manuscript with input from J.H., P.P., and T.H. J.H wrote the Appendix with input from J.L, P.P., and T.H.

Appendix 1: A tutorial on Weber fractions and the ANS

Here we describe the representations of the ANS, discrimination within the ANS, and the role of the Weber fraction in modeling performance in tasks that engage the ANS.

[7] Other results suggest that the present 2-color results are not due to subjects ''sticking with'' a verification procedure that will work for every trial type (e.g. 2–5 colors). Even when *only* 2-color trials are presented, adult performance is consistent with the meaning expressed in (11b) and not with (11a)/(14) (Pietroski et al. 2009).

In modeling performance on tasks that engage the ANS, it is necessary first to specify a model for the underlying approximate number representations. It is generally agreed that each numerosity is mentally represented by a distribution of activation on an internal ''number line.'' These distributions are inherently noisy and do not represent number exactly or discretely (e.g., Dehaene 1997; Gallistel and Gelman 2000). The representations of numerosity on the mental number line are often modeled as having linearly increasing means and linearly increasing standard deviation (Gallistel and Gelman 2000). In Fig. 5a we have drawn idealized curves which represent the ANS representations for numerosities 4–10 for an individual with a Weber fraction of .17 (nearest whole number ratio of 6:7), where the Weber fraction is a measure of the noisiness of the ANS representations (see below). Consider these curves to represent the amount of activity generated in the mind by a particular array of items in the world. An array of, e.g., six items will greatly activate the ANS numerosity representation of *six*, but because these representations are noisy this array will also activate representations of *five*, *seven*, etc. with the amount of activation centered on *six* and gradually decreasing to either side of *six*.[8] As the number of items in an array presented to a subject increases from 4 to 10, the standard deviation of the Gaussian curve that represents the corresponding numerosity increases, leading to a flattening and spreading of the activation.

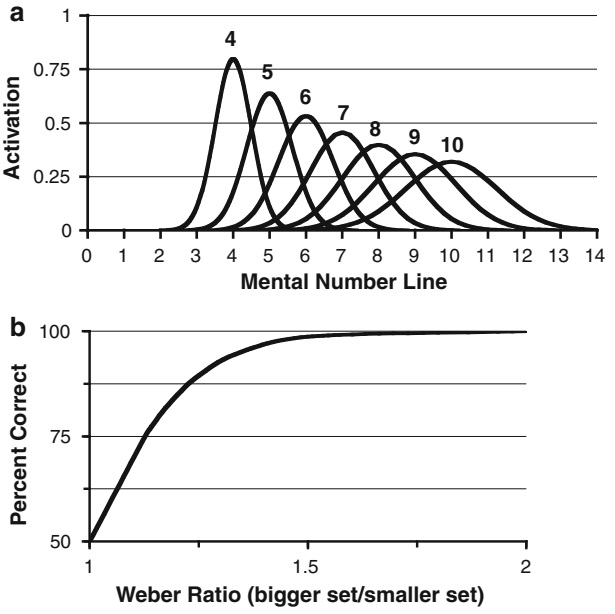


Fig. 5

[8] We describe the representation this way simply for ease of understanding. Strictly speaking, the number line of the ANS may be completely continuous and not have separate representations for, e.g., six items as distinct from, e.g., 5.76 items, and the entire curve centered on *six* might be considered the representation of *six-ness*.

To give a visual depiction of how number discrimination is possible in the ANS, consider the task of briefly presenting a subject with two arrays, e.g., five yellow dots and six blue dots, and asking the subject to determine which array is greater in number (Fig. 6a). The five yellow dots will activate the ANS curve representation of *five* and the six blue dots will activate the ANS curve representation of *six* (the subject uses attention to select which dots to send to the ANS for enumerating and then stores and compares those numerosity representations bound to their respective colors) (Fig. 6a,b).

An intuitive way to think about ordinal comparison within the ANS is to liken it to a subtraction (this will be mathematically equivalent to other ways of making an ordinal judgment within the ANS, and our use of subtraction here should be thought of as one illustration among several mathematically equivalent ones). Imagine that an operation within the ANS subtracts the smaller (i.e., yellow-*five*) representation from the larger (i.e., blue-*six*) representation (Fig. 6b). Because the *five* and *six*

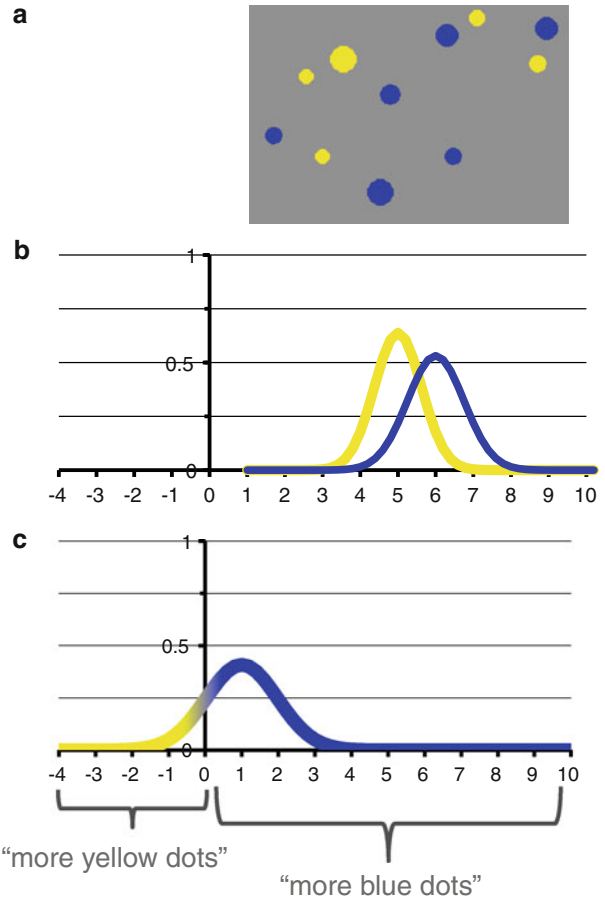


Fig. 6

representations are Gaussian curves, this subtraction results in a new Gaussian representation of the difference, which is a Gaussian curve on the mental number line that has a mean of *one* and a standard deviation of $\sqrt{(\sigma 5^2 + \sigma 6^2)}$; thus, when subtracting one Gaussian random variable from another (i.e., N6 – N5), the result is a new Gaussian random variable with the mean at the difference (i.e., $6 - 5 = 1$) and a variance that adds the variances of the original variables (i.e., $\sigma 5^2 + \sigma 6^2$). This results in a Gaussian curve that is centered on *one*, but that extends to both the left and right of *zero* (Fig. 6c). One can think of *zero* as the demarcation line separating evidence ''for'' and ''against,'' in that the area under the curve to the right of *zero* is the portion of the resulting representation that correctly indicates that *six* is greater than *five,* while the area under the curve to the left of *zero* is the portion of the resulting representation that incorrectly indicates that *five* is greater than *six*. This area to the left of *zero* results from the overlap between the original Gaussian representations, *five* and *six*, that were being discriminated in which some of the area of yellow-*five* is to the right (i.e., greater than) some of the area of blue-*six* (Fig. 6b).

From this resulting representation, there are multiple ways a subject might make a decision. Perhaps the simplest to consider is that, to give a discrete response, e.g., ''Yes, there are more blue dots than yellow dots,'' the subject could draw a single sample at random from this representation, and if it is to the left of *zero* respond, ''There are more yellow dots,'' and if it is to the right of zero respond, ''There are more blue dots'' (Fig. 6c). In this way, the probability of the subject getting this trial correct will depend on the relative area under the curve to the left and right of *zero,* which is itself determined by the amount of overlap between the original Gaussian representations for the numerosities being compared (i.e., *five* and *six*).

The more overlap there is between the two Gaussian representations being compared, the less accurately they can be discriminated. Consider comparing a subject's performance on a 5-dots versus 6-dots trial to a trial involving 9 versus 10 dots. Using the curves in Fig. 5a as a guide, we see that the overlapping area for the curves representing *five* and *six* is less than the overlapping area for the curves representing *nine* and *ten*, because the curves flatten and spread as numerosity increases. This means that it will be easier for the subject to tell the difference between *five* and *six* than between *nine* and *ten*, i.e., the resulting Gaussian for the subtraction will have more area to the right of *zero* for the subtraction of *five* from *six* than for the subtraction of *nine* from *ten*. Across multiple trials the subject would give more correct responses on the 5-versus 6-dots trials than the 9-versus 10-dot trials. More generally, the linear increase in the standard deviation of the curves representing the numerosities along the mental number line results in ratio-dependent performance whereby the discriminability of two numerosities increases as the ratio between them (e.g., bigger #/smaller #) increases. The spread of each numerosity representation in Fig. 5a from 4 to 10 is steadily wider than the numerosity representation before it. This means that the discriminability of any two numerosities is a smoothly varying function, dependent on the ratio between the two numerosities to be discriminated. In theory, such discrimination is never perfect because any two numerosities, no matter how distant from one another, will always share some overlap. At the same time, discrimination will never be entirely

impossible so long as the two numerosities are not identical. This is because any two numerosities, no matter how close (e.g., 67 and 68), will always have some non-overlapping area where the larger numerosity is detectably larger. Correct discrimination may occur on only a small percentage of trials if the two sets are very close in number, but it will never be impossible. This motivates the intuition that percent correct in a dot discrimination task should be a smoothly increasing function from a ratio of 1, where the number of yellow dots presented to the subject and the number of blue dots presented are identical and there is therefore no correct answer, to near-asymptotic performance (100% correct) when the ratio is large and therefore easier. How rapidly performance rises from chance (50%) to near-asymptotic performance (100%) is controlled by the subject's Weber fraction, which tracks the amount of spread in the subject's underlying number line Gaussian curve representations and therefore the overlap between any two numerosities as a function of ratio. In Fig. 5b we have drawn the expected percent correct for the task of determining which array, blue or yellow, has more dots. This curve, derived from the psychophysical model of the ANS, is the expected pattern for the subject depicted in Fig. 5a, whose Weber fraction (w) is .17. In Fig. 5b, we see the smooth increase in percent correct discrimination from a ratio of 1, where the yellow and blue dots have the same number, to near-asymptotic performance at a ratio of approximately 1.5 (e.g., a trial at this ratio might involve six yellow dots versus nine blue dots).

The precision of the ANS varies across individuals, with some people having a smaller Weber fraction (i.e., better performance and sharper Gaussian curves) and others having a larger Weber fraction (i.e., poorer performance owing to wider, noisier Gaussian curves) (Halberda et al. 2008). The Weber fraction indicates the amount of spread in the underlying mental number line representations (Halberda, in prep). In Fig. 7a we illustrate a subset of the idealized curves which represent the underlying ANS representations for a subject whose Weber fraction is .17 (better discrimination performance), and in Fig. 7b for a subject whose Weber fraction is .22 (poorer discrimination performance). Crucially, one can see that the subject in Fig. 7b has a greater degree of overlap between the Gaussian curves than the subject in Fig. 7a. It is this overlap that leads to difficulty in discriminating two stimuli that are close in numerosity. The hypothetical subject in Fig. 7b would have poorer discrimination in a dots discrimination task than the subject in Fig. 7a. In Fig. 7c we have drawn the ideal performance for these two subjects across many trials in a discrimination task.

Lastly, we consider the predicted curves for two of the algorithms discussed in this paper. Consider again the sentence 'Most of the dots are blue'. This statement invites one to compare the cardinality of the blue dots to the cardinality of the nonblue dots, as in (11). We specified two psychologically plausible ways of performing this comparison. In (11b), the numerosity of the nonblue dots is computed via a subtraction of the blue dots from the total number of dots. We noted that such an operation is feasible on psychological grounds as infants and adults are capable of retrieving the numerosity of the superset of all dots and the numerosity of the blue dots from a single flash (Halberda et al. 2006; Zosh et al. under review). In (12), the numerosity of the nonblue dots is computed via a context-dependent summation of the cardinalities for the color subsets that make up the nonblue dots.

We noted that such a computation is possible psychologically just so long as the individual color subsets can be enumerated (i.e., for 3 or fewer color subsets).

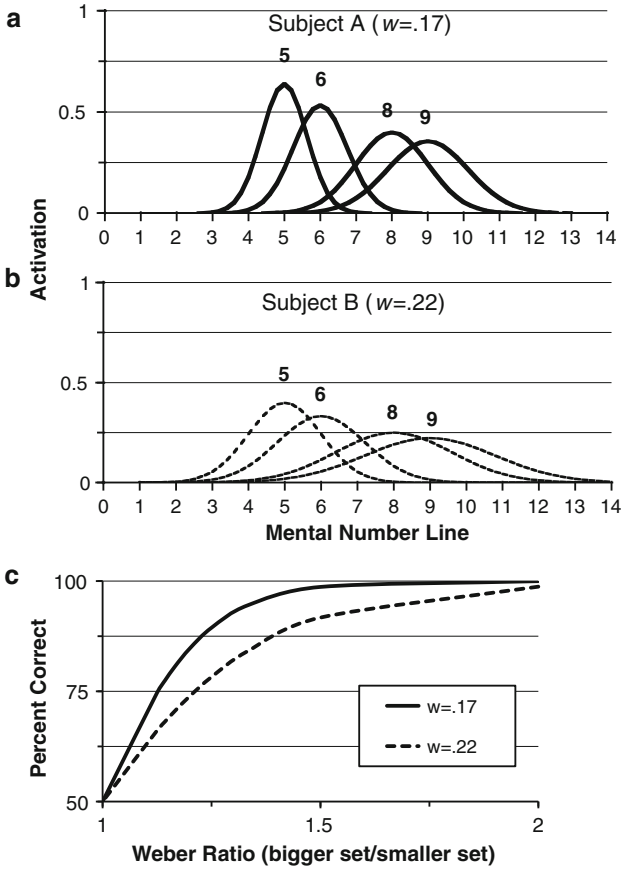


Fig. 7

(11) a. >(|DOT ∩ BLUE|, |DOT − BLUE|)
b. >(|DOT ∩ BLUE|, |DOT| − |DOT ∩ BLUE|)

(12) >(|DOT ∩ BLUE|, |DOT ∩ RED| + |DOT ∩ YELLOW|)

In (12), the observed Weber fraction will be lower than a simple comparison of two sets. The subtraction in (11b) will add noise from the variance of the Gaussian numerosity representations involved in the subtraction (e.g., note in the example above, shown in Fig. 6a–c, how variances add in Gaussian subtraction within the ANS). Both addition and subtraction lead to the addition of variance, but because the numerosities involved in the addition (12) are different from those involved in the subtraction (11b) these algorithms generate very specific and distinct predictions for how performance on the dot discrimination task should change as a function of

the number of color subsets in the display. In the experiment described in Sect. 5 we saw that the performance of subjects conformed to the predictions of (11b). Here, for the purposes of illustrating the value of the psychophysical model, we provide the predictions in greater detail for the performance of subjects engaged in (i) a simple comparison of two color subsets, (ii) the summation algorithm in (12) with increasing numbers of colors, and (iii) the subtraction algorithm in (11b) with increasing number of colors.

For all three computations, we will assume an ideal subject who has a Weber fraction of .17 (Fig. 5a). The first case to consider is a simple discrimination of blue versus yellow dots. As already described above, for a subject with an internal Weber fraction of .17, performance on such a task will be a smooth increase from chance (50%) to near-asymptotic performance (100%) as a function of increasing ratio (i.e., as the numbers of blue and yellow dots become more different). The steps underlying this process were displayed in Fig. 6. In Fig. 8a, we have reprinted the final performance curve from Fig. 5b and superimposed on it the average behavioral performance for each ratio from the actual subjects in the experiment (i.e., the ''most'' task) collapsed across all trials. In Fig. 8a we see that simple discrimination of, e.g., blue dots versus nonblue dots of a single color (e.g., yellow) from a subject with a Weber fraction of .17 (a reasonable estimate of adults' internal Weber fraction; see Halberda and Feigenson 2008; Ross and Burr 2010) predicts performance that far exceeds what we observed in our experiment. This suggests that adults are not directly enumerating the blue dots and the nonblue dots and performing a simple discrimination on these two sets.

Second, for the case of a ''sum the nonblue'' algorithm as in (12), each added addition in the second argument on the right side of the '>' sign will, perhaps counterintuitively, *reduce* the variance of the ''nonblue'' representation relative to the simple discrimination in Fig. 8a. This is because in Gaussian random variable addition and subtraction it is the variances that add, but for the ANS it is the standard deviation of the representations that increases linearly with the mean. One therefore arrives at a Gaussian representation with less error if one builds it through addition than if one enumerates it directly in the ANS (e.g., the standard deviation of the resulting Gaussian representation from an addition N7 + N5 is less than the standard deviation of N12 enumerated directly by the ANS). As the number of color subsets to be summed in the nonblue dots increases performance should *improve* slightly from simple discrimination. For a 2-color trial (e.g., blue and yellow dots only), the predicted performance from this algorithm is identical to a simple discrimination of, e.g., blue and nonblue dots (Fig. 8b). For 3-color, 4-color, and 5-color trials (e.g., blue versus red + yellow + green + cyan), the variance will be reduced by each addition, leading to a gradual increase in performance as a function of the number of colors, so long as subjects can enumerate the color subsets (Fig. 8b).

Another possible pattern that we discussed in Sect. 5.1 informed by research on visual attention and working memory for multiple color subsets (Halberda et al. 2006), is that subjects would simply fail to enumerate all of the color subsets on 4-color and 5-color trials which would result in performance curves in Fig. 8b that are radically different from those seen on 2-color and 3-color trials (e.g., good

performance on 2-color trials and near-chance performance on 5-color trials). Neither this pattern nor the improved performance displayed in Fig. 8b relative to simple discrimination was observed in the behavior of subjects in the ''most'' task.

Third, for the case of a ''subtraction algorithm'' as in (11b), the superset of all dots and the focused set of, e.g., blue dots will be selected on each trial and enumerated irrespective of the number of colors involved in the display (for evidence that the number of colors does not increase the variance of the numerosity representations see Halberda et al. 2006). The subtraction involved in (11b) will add a constant variance irrespective of the number of colors. This predicts poorer performance than a simple discrimination of, e.g., blue and yellow dots (Fig. 8a

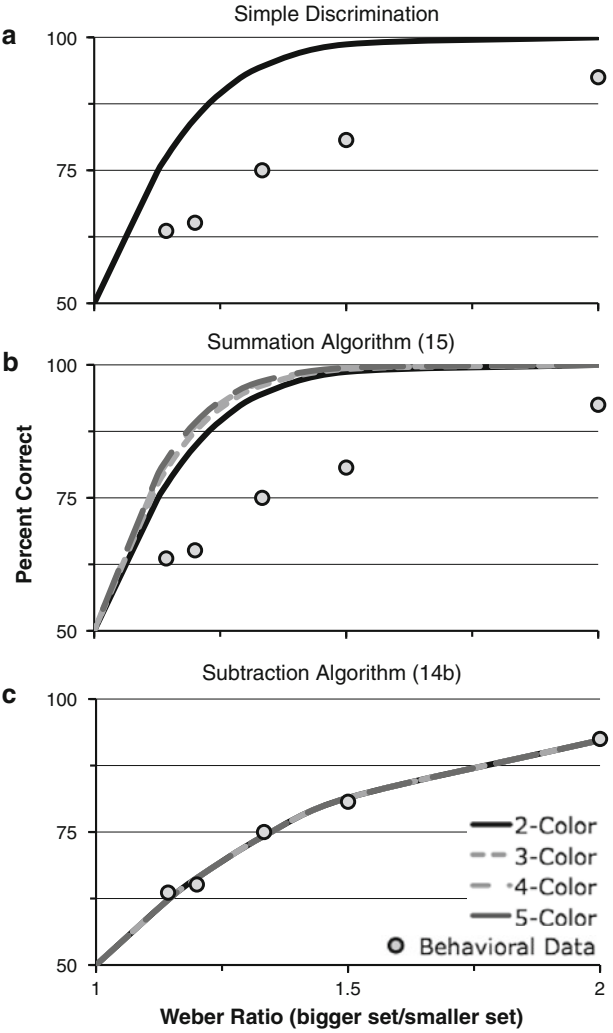


Fig. 8

compared to Fig. 8c), and it predicts that performance will remain at this suppressed level on 3-color, 4-color, and 5-color trials with no change in performance as a function of the number of colors in the display (Fig. 8c). This is what we observed for the performance of subjects in our experiment (Fig. 8c). Note that the predicted curves in Figure 8c, all falling directly atop each other, are for the subtraction algorithm for a subject whose *internal* Weber fraction is .17. In the experimental analysis presented in Sect. 5.7. we used the standard simple discrimination algorithm (Fig. 8a) to fit performance and found an *observed* Weber fraction of approximately .3. By assuming a subtraction algorithm (11b), we've fit performance well with a predicted internal Weber fraction of .17 (Fig. 8c). That independent methods suggest adults have an internal Weber fraction in the neighborhood of .11–.2 (Cordes et al. 2007; Halberda et al. 2008; Izard and Dehaene 2008; Ross and Burr 2010) lends further support for the modeled curves in Fig. 8c, and we believe the subtraction algorithm that generates these predictions, as stated in (11b), is the most likely source for the behavior we observed from subjects in the ''most'' task.

The psychophysical model of the ANS generates specific predictions for the performance of subjects on a variety of dot discrimination trials. These predictions allow a test of which algorithm most accurately describes the representations engaged as adults assess the truth of a statement like 'Most of the dots are blue'.

References

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

Boolos, G. 1998. *Logic, logic and logic*. Cambridge, MA: Harvard University Press.

Brannon, E.M., D. Lutz, and S. Cordes. 2006. The development of area discrimination and its implications for number representation in infancy. *Developmental Science* 9(6): F59–F64.

Chomsky, N. 1986. *Knowledge of language: Its nature, origins and use*. New York: Praeger.

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

Cordes, S., C. Gallistel, R. Gelman, and P. Lathan. 2007. Nonverbal arithmetic in humans: light from noise. *Perception & Psychophysics* 69: 1185–1203.

Cresswell, M. 1985. *Structured meanings*. Cambridge, Mass.: MIT Press.

Davidson, D. 1967. The logical form of action sentences. In *Essays on actions and events*, 105–148. Oxford: Clarendon Press.

Davies, M. 1987. Tacit knowledge and semantic theory: Can a five per cent difference matter? *Mind* 96: 441–462.

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

Dummett, M. 1973. *Frege: Philosophy of Language*. Cambridge, Mass.: Harvard University Press.

Evans, G. 1981. Semantic theory and tacit knowledge. In *Wittgenstein: To Follow a Rule*, ed. S. Holtzman and C. Leich, 118–137. London: Routledge and Kegan Paul.

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

Fodor, J. 2003. *Hume variations*. Oxford: Oxford University Press.

Frege, G. 1884. *Die Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner. J.L. Austin, English trans., *The foundations of arithmetic* (Oxford: Basil Blackwell, 1974).

Frege, G. 1892. *Über* Funktion *und Begriff* [English translation as 'Function and concept']. P. Geach and M. Black, trans., *Translations from the philosophical writings of Gottlob Frege*, 30–32 (Oxford: Blackwell, 1980).

Gallistel, C., and R. Gelman. 2000. Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences* 4(2): 59–65.

Hackl, M. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics* 17: 63–98.
Halberda, J. in prep. What is a Weber fraction? Ms., Johns Hopkins University.
Halberda, J., and L. Feigenson. 2008. Developmental change in the acuity of the ''Number Sense'': The approximate number system in 3-, 4-, 5-, 6-year-olds and adults. *Developmental Psychology* 44(5): 1457–1465.
Halberda, J., M.M.M. Mazzocco, and L. Feigenson. 2008. Differences in primitive math intuitions predict math achievement. *Nature* 455: 665–668.
Halberda, J., S.F. Sires, and L. Feigenson. 2006. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science* 17: 572–576.
Halle, M. 2002. *From memory to speech and back*. The Hague: Mouton de Gruyter.
Higginbotham, J., and R. May. 1981. Questions, quantifiers, and crossing. *The Linguistic Review* 1: 41–80.
Horty, J. 2007. *Frege on definitions*. Oxford: Oxford University Press.
Izard, V., and S. Dehaene. 2008. Calibrating the mental number line. *Cognition* 106: 1221–1247.
Jacobson, R., G. Fant, and M. Halle 1952. *Preliminaries to speech analysis: the distinctive features and their correlates*. Technical Report 13. Acoustics Laboratory, MIT.
Jackendoff, R. 1983. *Semantics and cognition*. Cambridge, Mass.: MIT Press.
Jackendoff, R. 1990. *Semantic structures*. Cambridge, Mass.: MIT Press.
Jackendoff, R. 2002. *Foundations of language*. Oxford: Oxford University Press.
Jusczyk, P. 1997. *The discovery of spoken language*. Cambridge, Mass.: MIT Press.
Kahneman, D., and A. Tversky. 1973. On the psychology of prediction. *Psychological Review* 80: 237–251.
Katz, J.J., and J.A. Fodor. 1963. The structure of a semantic theory. *Language* 39: 170–210.
Kuhl, P. 1993. Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics* 21: 125–139.
Landau, B., and R. Jackendoff. 1993. ''What'' and ''where'' in spatial language and spatial cognition. *Behavioral and Brain Sciences* 16(2): 217–238.
Liberman, A.M., F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review* 74: 431–461.
Liberman, A.M., and I.G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition* 21: 1–36.
Marr, D. 1982. *Vision*. Cambridge, Mass.: MIT Press.
Montague, R. 1970. Universal grammar. *Theoria* 36: 373–398.
Mostkowski, A. 1957. On a generalization of quantifiers. *Fundamenta Mathematicae* 44: 12–36.
Peacocke, C. 1986. Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language* 1: 101–123.
Pica, P., C. Lemer, V. Izard, and S. Dehaene. 2004. Exact and approximate arithmetic in an Amazonian indigene group. *Science* 306: 499–503.
Pietroski, P. 2010. Concepts, meaning and truth: First nature, second nature and hard work. *Mind & Language* 25(3): 247–278.
Pietroski, P., J. Lidz, T. Hunter, and J. Halberda. 2009. The meaning of *most*: Semantics, numerosity and psychology. *Mind & Language* 24(5): 554–585.
Poeppel, D., W. Idsardi, and V. van Wassenhove. 2008. Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London* B 363: 1071–1086.
Rescher, N. 1962. Plurality quantification. *Journal of Symolic Logic*, 27: 373–347.
Ross, J., and D. Burr. 2010. Vision senses number directly. *Journal of Vision* 10(2): 1–8.
Stevens, K. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In *Human communication: A unified view*, ed. E.E. David Jr. and P.B. Denes, 51–56. New York: McGraw-Hill.
Tarski, A. 1944. The semantical concept of truth and the foundations of semantics. *Philosophy and Phenomenological Research* 4: 341–375.
Treisman, A., and S. Gormican. 1988. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* 95(1): 15–48.
Treisman, A., and J. Souther. 1985. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: Human Perception & Performance* 16(3): 459–478.
Werker, J.F. 1995. Exploring developmental changes in cross-language speech perception. In *An invitation to cognitive science, Part I: Language*, ed. D. Osherson (series), L. Gleitman and M. Liberman (vol. eds), 87–106. Cambridge, Mass.: MIT Press.

Whalen, J., C.R. Gallistel, and R. Gelman. 1999. Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science* 10: 130–137.

Wolfe, J.M. 1998. Visual search. In *Attention*, ed. H. Pashler, 13–73. London: University College London Press.

Zosh, J.M., L. Feigenson, and J.P. Halberda (submitted). Working memory capacity for multiple collections in infancy.