Beyond Truth Conditions: The Semantics of *most*

Tim Hunter[1], Justin Halberda[2], Jeffrey Lidz[1], Paul Pietroski[1]
[1]*University of Maryland,* [2]*Johns Hopkins University*

1. Introduction: What are Meanings?

The language faculty generates expressions that relate sounds and meanings, but it is not immediately obvious exactly what kind of formal object "meanings" are. A common assumption is that sentence meanings are truth conditions: unstructured functions from worlds to truth values.[1] By looking at the relationship between language understanding and verification procedures, this paper explores the possibility that a sentence's meaning could in fact be something strictly richer than a truth condition, which makes reference to certain kinds of algorithms or representations that may be used to determine the truth value of the sentence in a particular world.

If this possibility is correct, then a sentence's meaning will give some privileged status to a subset of the possible verification procedures. On the other hand, if a sentence's meaning is nothing more than a bare truth condition, then all verification procedures implementing the right function from worlds to truth values will have equal status. We provide evidence that meanings are not verification-agnostic truth conditions, by showing that varying the suitability of a scene to some types of verification procedures affects participants' accuracy in assigning truth values, whereas varying the suitability of a scene to some other types verification procedures does not.

To address these questions we investigated English-speakers' understanding of the word *most*. This was chosen as a starting point because the conventional truth-conditional semantics of *most* is relatively well-understood, and there exists a range of useful findings concerning the psychology of number and constraints on visual perception that we can bring to bear on experimental results. But there is no reason that the same questions we ask about *most* should not in principle be asked about any other expression of natural language.

The rest of this paper is organised as follows. In Section 2 we outline the space of possible verification procedures for sentences of the form *Most (of the) Xs (are) Y*, and review some previous work addressing the relationship between the meaning and verification of these sentences. We then present two experiments.

[1]Or perhaps structured abstracta composed of functions, as in Cresswell (1985), though there the structure serves only to ensure that complex expressions containing embedded propositions can have their semantics determined compositionally. It therefore seems reasonable to assume that this structure is "visible" only to the language faculty, for the purposes of determining the semantics of other linguistic expressions, and that the object delivered to other cognitive faculties is nonetheless an unstructured truth condition. But of course if evidence is found that sentence meanings have structure that is visible beyond the language faculty, one might ask if this is "the same structure" as Cresswell proposes.

The first experiment (in Section 3) indicates that the meaning of *most* makes reference to a kind of cardinality concept, contrary to an alternative based on one-to-one correspondence that otherwise appears likely given the outline in Section 2. After examining more closely the range of verification procedures consistent with these findings (Section 4), the second experiment (Section 5) attempts to distinguish between these more fine-grained possibilities, and asks how the meaning (provided by the language faculty) of a *most* sentence interacts with properties of the visual system which constrain participants' perception of the relevant cardinalities.

2. Verification Procedures for *most*

2.1. Hackl (in press): most *and* more than half

Hackl (in press) compares the verification procedures employed in determining the truth of sentences like those in (1) in order to provide evidence for a decision between which of the statements in (2) better expresses the meaning of (1a).

(1) a. Most of the dots are yellow.

b. More than half of the dots are yellow.

(2) a. $|\text{DOT} \cap \text{YELLOW}| > |\text{DOT} - \text{YELLOW}|$

b. $|\text{DOT} \cap \text{YELLOW}| > \frac{1}{2}|\text{DOT}|$

Hackl observes differences in behaviour between participants who are asked to determine the truth of a sentence like (1a), and those who are asked to determine the truth of a sentence like (1b). From this he concludes that, despite the truth-conditional equivalence of the two expressions in (2), (2a) better expresses the meaning of (1a): since the meaning of (1b) is clearly best expressed as (2b), and participants' understanding of the two sentences led them to different patterns of behaviour, the best expression of the meaning of (1a) must be something different from (2b).[2]

This argument obviously relies on the notion that there is something more to a meaning than just a truth condition, and that this "something more" is reflected in the verification procedures used by speakers to assign truth values to sentences, as described in Section 1. Under these assumptions, the argument that the meaning of *most* differs from that of *more than half* is convincing, but we can not easily conclude with any certainty from Hackl's results exactly which verification procedures are implicated in the meanings of *most* and of *more than half* — only that the two meanings differ in this respect. Even on the assumption that (2b) in some sense best expresses the verificational implications of (1b), it is not clear that (2a) best

[2]For Hackl, this finding is one piece of evidence in a larger argument that the meaning of *most* differs from that of *more than half* in that it is constructed compositionally as a superlative expression. However, the relevant point for our purposes is just that the different verification profiles of the two expressions in (1) provided evidence against expressing the meaning of (1a) as (2b).

expresses those of (1a), because there are other expressions of the relevant truth condition to consider as possibilities.

2.2. *Verification without cardinalities*

In particular, while both the expressions in (2) make reference to cardinalities, there exist truth-conditionally equivalent expressions which do not. It is even tempting to suspect that such expressions might be more accurate representations of the meaning of (1a) than either of those in (2), given the intuition that it is easy to quickly determine the truth of (1a) in the scene shown in Figure 1 without determining any cardinalities at all — neither that of the set of all dots, nor the set of yellow dots, nor the set of blue (or non-yellow) dots. If we believe that the meanings of sentences inform verification procedures, then this might lead us to reject **both** of the expressions in (2), to the extent that they both imply that a comparison of cardinalities is required to verify (1a).

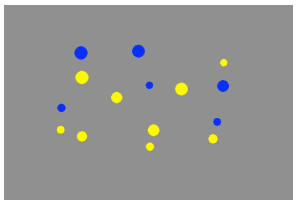


Figure 1: Intuitively, it seems to be possible to determine the truth of a sentence like *Most of the dots are yellow* without determining any cardinalities at all.

Besides the intuition about scenes like Figure 1, further evidence that *most* statements can be verified without determining cardinalities comes from research with young children. Halberda et al. (2008) tested three- and four-year-olds' understanding of *most* by asking them to determine the truth of sentences like *Most of the crayons are yellow* in scenes like those shown in Figure 2, while varying the number of crayons of each colour. Crucially, some children of this age ("non-counters") have not yet acquired the ability to represent and compare arbitrarily large integers: their understanding of cardinality does not extend beyond three or four. At some point in development, a child realises the recursive generalisation which permits representations of larger integers, and relatively suddenly gains the ability to represent all the remaining natural numbers (becoming a "full-counter").

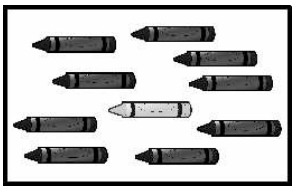
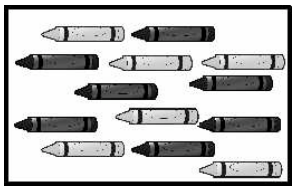
a.  b. 

Figure 2: Sample stimuli from Halberda et al. (2008), showing the easiest ratio of set cardinalities, 1:9 (a), and the hardest ratio, 6:7 (b).

The graph in Figure 3 shows the percentage of trials where participants correctly determined the truth of a *most* sentence as a function of the ratio of the size of the larger colour-set to the smaller one ("Weber ratio"), for children who were older than the estimated age of *most* comprehension in this task (Halberda et al. 2008). The significant point for current purposes is that even the non-counters performed significantly above chance for all but the hardest ratio. These are children that are unable to determine which of two given integers greater than three is the larger, so comparison of two cardinalities can not be a part of the verification procedure they are using to verify *Most of the crayons are yellow*. But they are performing **some** verification procedure for this sentence which results in above-chance accuracy.

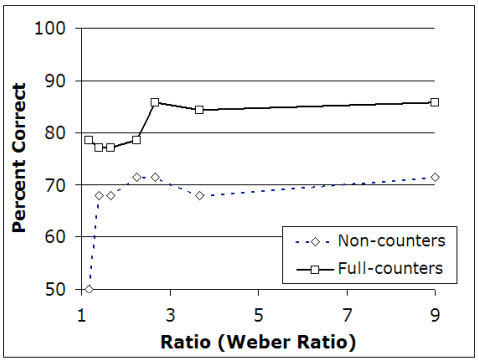


Figure 3: Percentage of responses correct for non-counters and full-counters at various ratios. Even non-counters perform significantly above chance at verifying *most* sentences.

2.3. Verification using one-to-one correspondence

What sort of verification procedure could be used in these cases where cardinality comparison is not possible, either because the participant is not a full-counter or because the relevant items are shown too quickly to count? If we consider the mathematical foundations of cardinality, then the possibility of a verification procedure based on the notion of **one-to-one correspondence** presents itself. A common way to define cardinality begins by stating that two sets A and B have the same cardinality if and only if the elements of A can be put in one-to-one correspondence[3] with the elements of B:

(3) $\quad |A| = |B| \iff \text{OneToOne}(A,B)$

Therefore it is possible to determine, for example, that the set of yellow dots has the same cardinality as the set of non-yellow dots, by determining that there is a one-to-one correspondence between the yellow dots and the non-yellow dots, without knowing what this shared cardinality is.

[3]More formally, the elements of A can be put in one-to-one correspondence with the elements of B if and only if there exists a bijection (a surjective, injective function) with domain A and range B.

Extending this to the case where the cardinality of one set exceeds that of another, via the definition of the greater-than relation, it follows that (for finite sets) the cardinality of a set A is greater than that of a set B if and only if there exists some proper subset of A, call it A', such that there exists a one-to-one correspondence between the elements of A' and the elements of B:

(4) $|A| > |B| \iff \exists A'[\text{OneToOne}(A', B) \text{ and } A' \subset A]$

For convenience we define a new relation on sets OneToOnePlus as follows:

(5) $\text{OneToOnePlus}(A, B) \iff \exists A'[\text{OneToOne}(A', B) \text{ and } A' \subset A]$

and so we have:

(6) $|\text{DOT} \cap \text{YELLOW}| > |\text{DOT} - \text{YELLOW}|$
$\iff \exists A'[\text{OneToOne}(A', (\text{DOT} - \text{YELLOW})) \text{ and } A' \subset (\text{DOT} \cap \text{YELLOW})]$
$\iff \text{OneToOnePlus}(\text{DOT} \cap \text{YELLOW}, \text{DOT} - \text{YELLOW})$

Therefore it is possible to identify expressions which are truth-conditionally equivalent to those in (2), but which do not make reference to any cardinalities. This implies that it is possible to determine that the cardinality of the set of yellow dots is greater than that of the set of non-yellow dots, by recognising that there exists a one-to-one correspondence between the non-yellow dots and some proper subset of the yellow dots — again, without ever determining any cardinalities. This situation is illustrated in Figure 4. Studies of object-tracking competence in infants (Wynn 1992, Feigenson 2005) have revealed a cognitive system that can detect the required kind of one-to-one correspondences. It is therefore tempting to conclude that when a child without full numerical competence manages to correctly judge the truth of a *most* statement, or when a competent adult does so having glanced at Figure 1 so quickly that counting is not possible, a verification procedure based on one-to-one correspondence is being used.

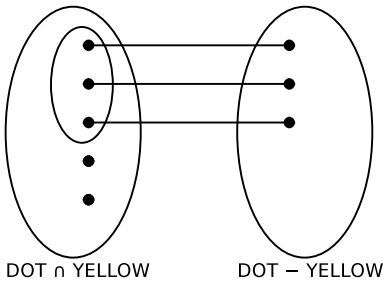


Figure 4: Recognising that a proper subset of the yellow dots can be put in one-to-one correspondence with (all of) the non-yellow dots might permit one to determine that *Most of the dots are yellow* is true.

2.4. Verification using the Approximate Number System

There is, however, another cognitive system known to psychologists that could permit verification procedures for *most* statements that neither compares cardinalities

nor relies on one-to-one correspondence. From birth, humans share with many non-verbal animals an **Approximate Number System** (ANS) that very quickly (within 150ms of visual stimulus onset (Nieder and Miller 2004)) generates representations of pluralities in ways that effectively order those pluralities according to cardinality — albeit stochastically, and within certain limits described by Weber's Law (Cordes et al. 2001, Feigenson et al. 2004, Dehaene 1997). Weber's Law states that discriminability (for our purposes, the ability to determine which of two ANS representations corresponds to the greater cardinality) depends only on the ratio of the two represented cardinalities.

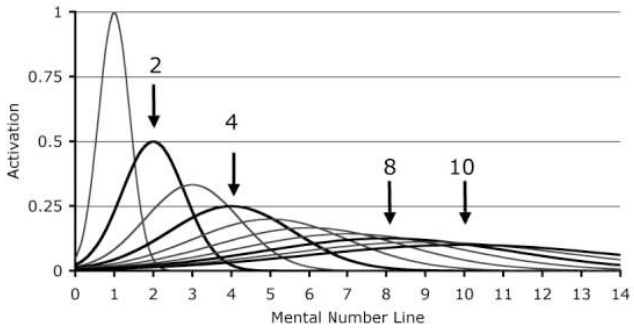


Figure 5: The representations of the ANS, modelled by a sequence of Gaussian curves with linearly increasing means and standard deviations.

To model this pattern of discriminability we can consider a "mental number line" as illustrated in Figure 5. On this view, the "noise" in the representations increases with the number represented: the ANS representation triggered by the perception of a set with cardinality n is characterised by a Gaussian curve with mean n and standard deviation directly proportional to n. The overlap between any two curves therefore increases linearly with the ratio between the represented cardinalities, and thus by Weber's Law acts as a predictor of the difficulty of distinguishing two representations (and identifying which of the represented cardinalities is the larger). For example, it is clear from Figure 5 that the curves labelled 8 and 10 overlap to a much greater extent than those labelled 2 and 4, indicating that the ANS has more difficulty distinguishing a set of eight things from a set of ten things, than it does distinguishing a set of two things from a set of four things. Also, note that because the ratio of 2 to 4 is the same as the ratio of 4 to 8, the ANS's ability to distinguish a set of two things from a set of four things is as good as its ability to distinguish a set of four things from a set of eight things. The threshold of discriminability (the ratio by which two numbers must differ in order for the ANS to be able to reliably distinguish their representations) varies from individual to individual (Halberda et al. in press) and improves with age (Halberda and Feigenson in press); but wherever this threshold is, there will exist some cases where it is not possible to determine the precise cardinality of any sets of dots (because the relevant dots are perceived only very briefly), but where it is nevertheless possible to determine the truth of *Most of the dots are yellow* by constructing and comparing

ANS representations of the approximate cardinality of the set of yellow dots and that of the set of non-yellow dots.

2.5. Summary

To summarise, for the sentence *Most of the dots are yellow*, we have identified two ways to express the relevant truth condition (7) which make reference to different kinds of formal objects (sets and cardinalities in one case, and only sets in the other), and three classes of verification procedures (8) that could be used to compute the relevant truth value in a given scenario:

(7) a. $|\text{DOT} \cap \text{YELLOW}| > |\text{DOT} - \text{YELLOW}|$

b. $\text{OneToOnePlus}(\text{DOT} \cap \text{YELLOW}, \text{DOT} - \text{YELLOW})$

(8) a. Procedures involving computing and comparing precise cardinalities.

b. Procedures involving detecting one-to-one correspondences.

c. Procedures involving generating and comparing ANS representations.

The conventional expression of the relevant truth condition, (7a), suggests the class of verification procedures in (8a), but we have good reason to believe that some other verification procedures exist. The fact that we can express this same truth condition as (7b) alerts us to the existence of the class of verification procedures in (8b). Finally, having noted that the ANS provides a system of representations that support a stochastic version of the ordering relation that (7a) relies on, we can identify the class of procedures in (8c): stochastic versions of the comparison-based procedures in (8a).

In the next section we present evidence that when the possibility of using the cardinality-based procedures of (8a) is eliminated, speakers revert to the ANS-based procedures of (8c) rather than the correspondence-based procedures of (8b), even in situations which seem to be well-suited to correspondence-based procedures. We think the strength of this bias against correspondence-based verification procedures constitutes evidence against the claim that a competent speaker's understanding of a *most* statement is exhaustively characterised by a verification-agnostic truth condition.

3. Experiment 1

This experiment is also reported in Pietroski et al. (2008).

3.1. Design and Procedure

On each trial, participants saw a 200ms display containing dots of two colours, yellow and blue. Participants were asked to judge *Most of the dots are yellow* true

or false for each trial. The number of dots of each colour varied between five and seventeen. Whether the yellow set or the blue set was larger (and hence, whether the correct answer was "true" or "false") was randomised. Participants answered "true" or "false" by pressing buttons on a keyboard.

Each trial came from one of nine "bins", each characterised by a ratio. The first bin contained trials where the ratio of the smaller set to the larger set was close to 1:2; the second bin contained trials where the ratio was close to 2:3; and the remaining bins contained trials close to 3:4, 4:5, ..., 9:10. Each participants received ten trials in each bin for each of three conditions: Scattered Random, Scattered Pairs and Column Pairs. The total number of trials for each participant was therefore 9 ratios $\times$ 3 conditions $\times$ 10 trials = 270. These were presented in randomised order.

On Scattered Random trials, all the dots (yellow and blue) were scattered randomly throughout the display. See Figure 6a. In the other two conditions, dots were displayed in some way intuitively amenable to a one-to-one correspondence-based verification procedure, with yellow dots and blue dots occurring in pairs. On Scattered Pairs trials, every dot from the smaller set was displayed paired with (approximately four pixels away from) a dot from the larger set, and the remaining dots from the larger set were scattered randomly. See Figure 6b. On Column Pairs trials, dots were arranged in a grid with two columns and n rows, where n is the size of the larger set. Each row had either one dot from each set or a single dot from the larger set, with the position (left column or right column) of each dot chosen randomly for each row. See Figure 6c.

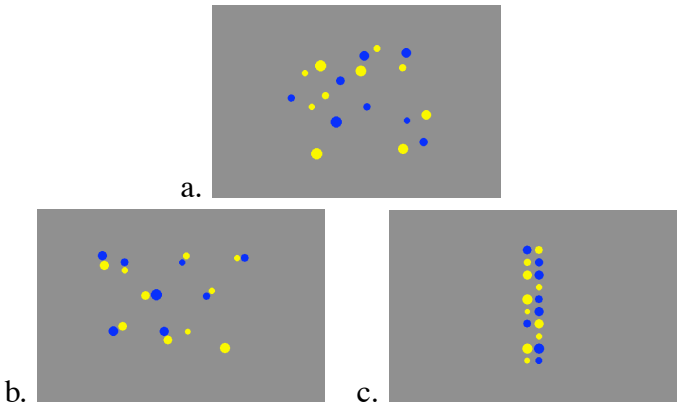


Figure 6: Sample stimuli from Experiment 1, from each condition: Scattered Random (a), Scattered Pairs (b) and Column Pairs (c).

Half of the trials for each condition were "area-controlled": individual dot sizes varied, but the number of yellow pixels was equal to the number of blue pixels (that is, the average yellow dot was smaller than the average blue dot whenever there were more yellow dots than blue dots). This prevented using the total area covered by a colour as a proxy for set cardinality. The other half of the trials for each condition for each ratio were "size-controlled": while individual dot sizes varied, the size of the average yellow dot was equal to the size of the average blue dot, so the set with more dots would also have a larger total area on the screen

(that is, more yellow pixels than blue pixels whenever there were more yellow dots than blue dots). This avoided confounding average dot size with set cardinality, because if all trials were area-controlled then one could determine the set with the larger cardinality by comparing dot sizes across colours. On both area-controlled and size-controlled trials, individual dot sizes varied randomly by up to 35% of the set average, such that dots of the same colour were not all of the same size (see Figure 6).

3.2. Predictions

We can identify three distinct hypotheses about the verification procedures used by participants. The 200ms display time does not permit verification procedures based on explicit counting, ruling out cardinality-based procedures (8a).

Firstly, participants might use one-to-one correspondence-based procedures (8b) on all trials. In this case we predict responses to be affected by dot layout (more accurate on Scattered Pairs and/or Column Pairs trials than on Scattered Random trials), but unaffected by ratio.

Secondly, participants might use ANS-based procedures (8c) on all trials. In this case we predict responses to be affected by ratio (more accurate on "easy" ratios like 1:2 and 2:3 than on "hard" ratios like 8:9 and 9:10), but unaffected by dot layout.

Thirdly, participants might adopt the most suitable verification procedure for each individual trial. In this case we predict responses to be affected by both ratio and dot layout. Broadly speaking, accuracy should be higher on trials that use the pairing layouts **or** use easy ratios; in either case, participants should be able to adopt a verification procedure which takes advantage of the display's properties.

Of course, these predictions rely on the assumption that the Scattered Pairs and Column Pairs trials do in fact permit the detection of the relevant one-to-one correspondences within the 200ms display time. Control experiments using identical stimuli have shown that the 200ms display time is sufficient to detect these one-to-one correspondences and identify the uniform colour of the remaining dots (Halberda et al. 2007). Nothing inherent to the stimuli, then, can be preventing participants from using correspondence-based verification procedures; if they do not do so, there must be another reason.

3.3. Results and Discussion

Percentage of correct responses for each participant was entered into a 3 condition (Scattered Random, Scattered Pairs, Column Pairs) $\times$ 2 trial type (size-controlled, area-controlled) $\times$ 9 ratio Repeated Measures ANOVA. There was a significant effect of ratio, as participants did better with easier ratios ($F(8,80) = 14.603, p < 0.001$), and no significant effect of condition ($F(2,20) = 0.215, p = 0.808$). This pattern of results can be seen in Figure 7. There was also no significant effect of trial type ($F(1,10) = 3.187, p = 0.105$), indicating that participants relied on the

number of dots and not other factors such as area that might be confounded with number, so performance has been collapsed across trial type in Figure 7.

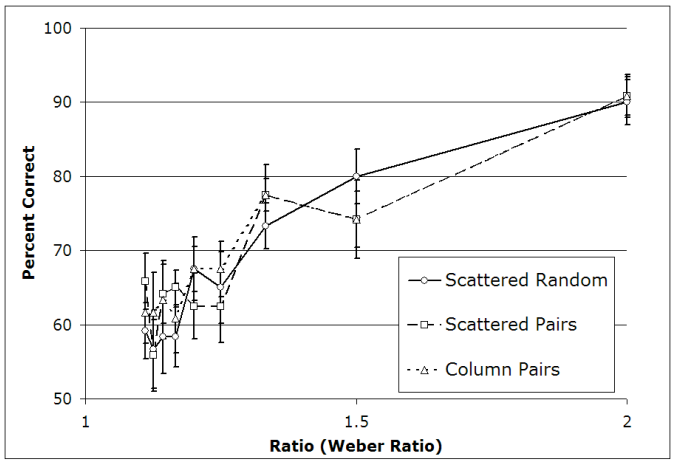


Figure 7: Percentage of responses correct for each trial type as a function of ratio. The ratio referred to as 1:2 in the main text appears at 2 on the x-axis; the ratio referred to as 9:10 appears at $\frac{10}{9} \approx 1.11$.

These results favour the second hypothesis presented in Section 3.2, on which participants use ANS-based procedures on all trials. Their performance improved as the ratio of yellow dots to non-yellow dots became less "even" (further from 1:1), as predicted by the ratio-dependence of ANS comparisons, but did not improve (or change at all) when dots were presented in obvious pairs.

Furthermore, the particular pattern of improvement as the ratios became less "even" — that is, the particular **shape** of the upward trend in Figure 7 — matched the function predicted by the standard model of the psychophysics of the ANS extremely closely. If we suppose that a participant's representation of the cardinality of a set of n_1 dots is a Gaussian curve with mean n_1 and standard deviation wn_1, and thus that the representation for n_2 dots is a Gaussian curve with mean n_2 and standard deviation wn_2, then the standard model predicts that the probability of the participant judging n_2 to be greater than n_1 is determined by the curve representing the difference between these two random variables: the Gaussian curve with mean $(n_2 - n_1)$ and standard deviation $w\sqrt{n_1^2 + n_2^2}$ (Pica et al. 2004). (Here w is a constant characterising the acuity of this particular individual's ANS, called the "internal Weber fraction".) In particular, the probability of n_2 being judged greater than n_1 is the proportion of the area under this curve to the right of zero (because this is the probability of $(n_2 - n_1)$ being judged greater than zero), as given by the following formula:

$$(9) \qquad \Pr(n_2 \text{ judged greater than } n_1) = \tfrac{1}{2}\operatorname{erfc}\left(\frac{n_1 - n_2}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}}\right)$$

Therefore for each value of w, the model determines a function mapping Weber ratio

to percentage correct responses.[4] The crucial point is that there exists a value of w for each condition such that this function matches the results extremely closely.[5] See Figure 8 and Table 1.[6] This constitutes strong evidence that participants used ANS-based verification procedures, and not any other verification procedure which would show improved accuracy with less evenly-matched ratios.[7]

Condition	Correlation (R^2)	Internal Weber Fraction (w)
Scattered Random	0.9677	0.32
Scattered Pairs	0.8642	0.33
Column Pairs	0.9364	0.30

Table 1: The high values of R^2 (close to 1) indicate a high correlation between the predictions of the ANS model, for the given value of the internal Weber fraction (w), and the pattern of results for each condition.

At least as importantly from a linguistic point of view, however, as telling us which particular verification procedure was used, these results suggest that the choice of which verification procedure to use is not as unconstrained as one might have thought. Participants did **not** adopt the most suitable verification procedure

[4]While the formula in (9) is expressed in terms of n_1 and n_2 (and w), the result is uniquely determined by the ratio of n_1 to n_2 (and w):

$$\frac{n_1 - n_2}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}} = \frac{\frac{1}{n_2}(n_1 - n_2)}{\sqrt{2}w\frac{1}{n_2}\sqrt{n_1^2 + n_2^2}} = \frac{\frac{n_1}{n_2} - 1}{\sqrt{2}w\sqrt{\left(\frac{n_1}{n_2}\right)^2 + 1}}$$

[5]Typical values for w in a task where participants are asked to directly judge the truth of a sentence like *There are more yellow dots than blue dots* are around 0.14 (Pica et al. 2004), so our participants' accuracy in verifying *Most of the dots are yellow* is poorer — their performance showed the signature of a system with $w \approx 0.3$, meaning larger standard deviations and thus noisier representations. Further research is required to determine exactly why this is, although Experiment 2 will suggest one possibility.

[6]The curve for the Scattered Random condition falls slightly below that of Scattered Pairs because of a slight tendency for participants to guess randomly on some trials ($\approx$6% in each condition). Given the size of the standard errors for each condition (see Figure 7), the estimated value of w for each condition — and therefore the predicted curve for each condition — should be considered statistically indistinguishable from the others.

[7]In particular, this tells against the hypothesis that participants' understanding of *most* required them to verify a stricter truth condition than that assumed in this paper, something along the lines of "significantly more than half" of the dots being yellow. Since the graph in Figure 7 shows the percentage of responses which agreed with the condition $|\text{DOT} \cap \text{YELLOW}| > |\text{DOT} - \text{YELLOW}|$, a verification procedure for a stricter "significantly more" truth condition would also show "poor accuracy" for ratios close to 1:1, because this is where (the truth values determined by) the two truth conditions diverge. But the responses closely matched the function predicted by the model on the assumption that participants were attempting to verify the truth of $|\text{DOT} \cap \text{YELLOW}| > |\text{DOT} - \text{YELLOW}|$, as opposed to any other truth condition; note that the curves in Figure 8 predict better than chance accuracy for Weber ratios even slightly above 1. For further elaboration of this point, and other evidence against the "significantly more than half" hypothesis, see Pietroski et al. (2008).

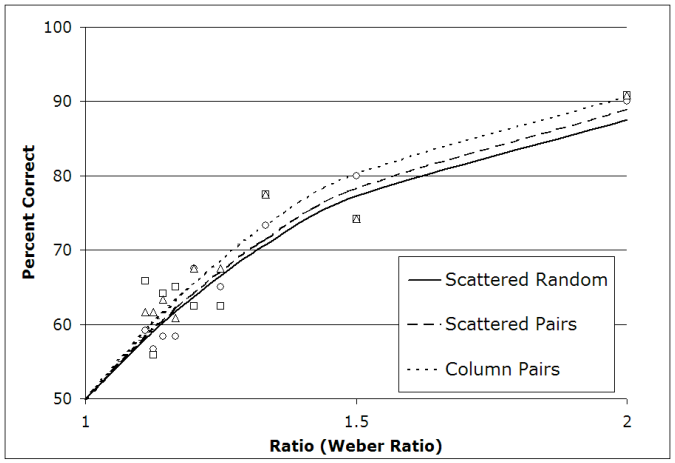


Figure 8: The data from Figure 7 displayed (as points) for comparison with the predictions of the standard model of ANS performance (as curves).

for each individual trial, as the third hypothesis in Section 3.2 suggests: in the Scattered Pairs and Column Pairs conditions, where the fact that the yellow dots and the non-yellow dots were in the OneToOnePlus relation was made obvious (and perceptible by participants in our control studies (Halberda et al. 2007)), participants showed exactly the same ANS-based pattern of responses as in the Scattered Random condition. Speakers' reluctance to take advantage of the pairing of dots in the display is unexplained if the meaning of the sentence *Most of the dots are yellow* is a verification-agnostic truth condition. It is more consistent with the view that the meaning of a sentence comes with some bias towards certain kinds of verification procedures.

In beginning to describe the precise nature of this bias, we can tread only very carefully. Clearly we do not want to deny that a sentence like *Most of the dots are yellow* **can** be verified using procedures which are not ANS-based. If presented with a Scattered Random display with a "hard" ratio (say, ten yellow dots and nine non-yellow dots) for an unlimited amount of time, a speaker would probably determine the truth of the sentence by counting dots and comparing precise cardinalities (8a), never making use of the ANS at all. Even if the only information provided about a scene is that the yellow dots and the non-yellow dots are in the OneToOnePlus relation, a speaker may well be able to determine that the sentence is true (8b). But the results of this experiment suggest that there is at least some asymmetry between the range of possible procedures; that the ANS-based procedures, despite being less accurate, are in some sense more directly available for verification of a *most* sentence than correspondence-based procedures. For more detailed discussion of the nature of this asymmetry, see Pietroski et al. (2008).

4. Finer-grained Distinctions Among ANS-Based Verification Procedures

The results from Experiment 1 indicate that participants used ANS representations to perform some stochastic version of the comparison indicated in this expression of the relevant truth condition:

(10) $|\text{DOT} \cap \text{YELLOW}| > |\text{DOT} - \text{YELLOW}|$

We can now ask more detailed questions about exactly how this comparison is carried out. In particular, we focus on the (approximate) representation of the cardinality $|\text{DOT} - \text{YELLOW}|$.

To investigate more closely how a representation of this cardinality is constructed, we need to turn to displays containing dots of more than two colours. We can identify two distinct procedures which could in principle be used in this scenario to construct a representation of the numerosity of the non-yellow dots. The first, the **Subtraction Procedure**, involves attending to the "superset" containing all dots, attending to the set of yellow dots, and performing a subtraction on the two generated ANS representations. The second, the **Selection Procedure**, involves attending to each of the non-yellow colour sets individually, and summing the representations of the cardinalities of these sets in cases where there is more than one non-yellow colour present.

However, some constraints on the parallel visual perception of sets have been identified that bear on the psychological plausibility of these two procedures. Halberda et al. (2006) found that when adults are briefly presented with a scene containing dots of between two and six colours, they can use the ANS to estimate the cardinality of up to three sets in parallel. One of these is necessarily the "superset" containing all the dots in the display, so this amounts to a constraint that at most two subsets can be selected. The property of having a particular colour is a salient "early visual feature" (Treisman and Gormican 1988), so one can, for example, attend to all the yellow dots and construct an ANS representation of the cardinality of this set (Halberda et al. 2006). However, it is **not** possible to select dots on the basis of a disjunction or negation of such properties (Wolfe 1998). So one can not, in a scene with, say, yellow dots as well as dots of four other colours, atomically attend to the set of all non-yellow dots — this would require selecting dots on the basis of a disjunction like "red or blue or ...", or the negation "not yellow". In summary, when briefly presented with an array of dots of a number of different colours, humans can generate (at most) three ANS representations: firstly, one for the set of all dots present; secondly, one for the set of dots of a particular colour, say, yellow; and thirdly, one for the set of dots of another colour, say, blue.

We therefore know that the Subtraction Procedure is psychologically plausible, no matter how many non-yellow colours are present, because it only requires two ANS representations to be generated from the visual stimulus: one for the "superset" of all dots, and one for the set of yellow dots. The Selection Procedure is plausible in cases where there is only one non-yellow colour present, say, blue, because it is sufficient to attend to the set of blue dots and the set of yellow dots (in addition to the "superset"). In cases with more than one non-yellow colour

present, however, the Selection Procedure becomes impossible, because it would require attending to more than two colour subsets of the display (in addition to the "superset"): the set of yellow dots, and more than one non-yellow colour set.

There is another significant distinction between these two verification procedures. In cases where it is possible to use the Selection Procedure (that is, when there are only two colours of dots present), it will give more accurate results than the Subtraction Procedure. This is because in these cases the representation of $|\text{DOT} - \text{YELLOW}|$ is atomically detected by the Selection Procedure, but is computed indirectly by the Subtraction Procedure. Therefore the noise inherent to ANS representations is magnified by the use of the Subtraction Procedure.

The combined implications of varying the number of non-yellow colours present in a display for the two verification procedures considered here are summarised in Table 2.

Number of colours present	2	3	4	5
Subtraction Procedure	good	good	good	good
Selection Procedure	better	impossible	impossible	impossible

Table 2: The effects of varying the number of non-yellow colours present in a scene on two possible procedures for verifying the statement *Most of the dots are yellow*.

5. Experiment 2

This experiment is also reported in Lidz et al. (2008).

5.1. Design and Procedure

On each trial, participants saw a 150ms display containing dots of at least two colours and at most five colours (chosen from yellow, blue, red, green, cyan, magenta). Yellow dots were present on every trial.[8] Participants were asked to judge *Most of the dots are yellow* true or false for each trial. The number of yellow dots and the number of non-yellow dots varied between five and seventeen. Whether the yellow set or the non-yellow set was larger (and hence, whether the correct answer was "true" or "false") was randomised. Participants answered "true" or "false" by pressing buttons on a keyboard.

Within each of the four conditions (two to five colours), the ratio of the cardinality of the smaller set (yellow or non-yellow) to that of the larger set ranged over 1:2, 2:3, 3:4, 5:6 and 7:8 (a subset of the ratios used in Experiment 1). Each participant received fifteen trials in each ratio bin for each of the four conditions. The total number of trials for each participant was therefore 5 ratios $\times$ 4 conditions $\times$ 15 trials = 300. These were presented in randomised order.

[8]For irrelevant technical reasons, the target colour was actually blue in this experiment rather than yellow as in Experiment 1. We abstract away from this change for ease of exposition.

Half of the trials for each condition were "area-controlled": individual dot sizes varied, but the number of yellow pixels was equal to the number of non-yellow, non-background pixels (that is, the average yellow dot was smaller than the average non-yellow dot whenever there were more yellow dots than non-yellow dots). This prevented using the total area covered by a colour as a proxy for set cardinality. The other half of the trials for each condition for each ratio were "size-controlled": while individual dot sizes varied, the size of the average yellow dot was equal to the size of the average non-yellow dot, so the set with more dots would also have a larger total area on the screen (that is, more yellow pixels than non-yellow, non-background pixels whenever there were more yellow dots than non-yellow dots). This avoided confounding average dot size with set cardinality, because if all trials were area-controlled then one could determine the set with the larger cardinality by comparing dot sizes across colours. On both area-controlled and size-controlled trials, individual dot sizes varied randomly by up to 35% of the set average, such that dots from the same set were not all of the same size.

5.2. Predictions

We can identify three distinct hypotheses about the verification procedures used by participants. Note that we can expect responses on two-colour trials to pattern identically to those in Experiment 1, and so the hypotheses diverge only in their predictions of how this pattern will or will not change as the number of colours in the display increases.

Firstly, participants might use the Subtraction Procedure on all trials. In this case we predict responses to be unaffected by the number of colours in the display, and pattern identically to those in Experiment 1 throughout.

Secondly, participants might use the Selection Procedure on all trials. In this case we predict responses to be at chance when the number of colours present is greater than two, because this verification procedure fails.

Thirdly, participants might adopt the most suitable verification procedure for each individual trial. In this case the Selection Procedure will be used on two-colour trials and the Subtraction Procedure elsewhere, so accuracy on trials with more than two colours should be above chance but lower than accuracy on two-colour trials.

5.3. Results and Discussion

Percentage of correct responses for each participant was entered into a 4 condition (2, 3, 4, 5 colours) $\times$ 2 trial type (size-controlled, area-controlled) $\times$ 5 ratio Repeated Measures ANOVA. There was a significant effect of ratio, as participants did better with easier ratios ($F(4,44) = 109.092$, $p < 0.001$), and no effect of condition (number of colours in the stimulus) ($F(3,33) = 7.326$, $p = 0.842$). This pattern of results can be seen in Figure 9. As in Figure 7, we have collapsed across trial type to construct this graph; though here there was a marginal effect of trial type, as

participants did slightly better on size-controlled trials than on area-controlled trials ($F(1,11) = 7.326$, $p < 0.05$).

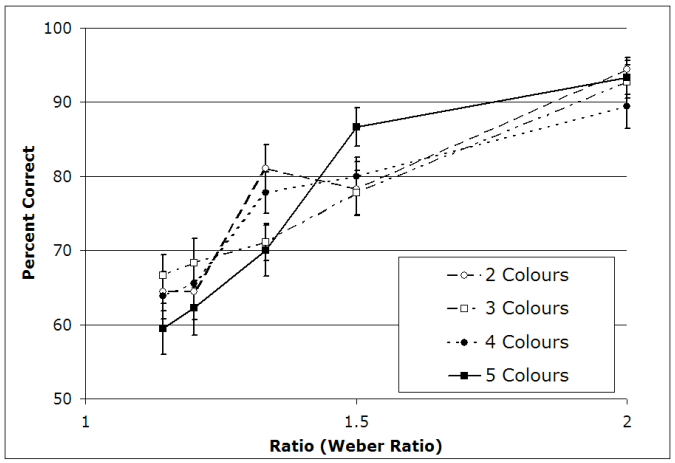


Figure 9: Percentage of responses correct for each trial type as a function of ratio. The ratio referred to as 1:2 in the main text appears at 2 on the x-axis; the ratio referred to as 7:8 appears at $\frac{8}{7} \approx 1.14$.

These results favour the first hypothesis presented in Section 5.2, on which participants use the Subtraction Procedure on all trials. Participants' responses showed the same pattern as in Experiment 1 in all conditions, indicating that one verification procedure is used throughout which is impervious to the heterogeneity (or otherwise) of the non-yellow dots.

Furthermore, the results again matched those predicted by the standard psychophysical model of the ANS extremely closely, as shown in Table 3. Note that the estimations of internal Weber fraction (w) are close to those in Experiment 1 in all four conditions, around 0.3 (see Table 1). If the more accurate Selection Procedure had been used on two-colour trials, we would have expected a significantly lower value for w in the two-colour condition than in the other three conditions, indicating a less noisy computation.[9]

As in Experiment 1, we have not only identified which particular verification procedure was used, but also found more evidence that the choice of verification procedure is not as unconstrained as a truth-conditional theory of sentence meanings would predict. Participants did **not** adopt the most suitable verification procedure for each individual trial, as the third hypothesis in Section 5.2 suggests: accuracy was no better in the two-colour trials than in other trials, indicating that participants did not use the Selection Procedure even when it was psychologically

[9]The use of the Subtraction Procedure provides one possible explanation for the relatively high values of w; see footnote 5. If participants had been using the Selection Procedure, they would have presumably been carrying out a procedure equivalent to that of the *There are more yellow dots than blue dots* task in which a value of w around 0.14 is typically found. Indeed, one might have considered the high value of w in Experiment 1 to be evidence that the Subtraction Procedure was being used there, even before the findings of Experiment 2 showed the same accuracy with more colours present.

Condition	Correlation (R^2)	Internal Weber Fraction (w)
2 Colours Present	0.9480	0.29
3 Colours Present	0.9586	0.32
4 Colours Present	0.9813	0.28
5 Colours Present	0.9625	0.32

Table 3: The high values of R^2 (close to 1) indicate a high correlation between the predictions of the ANS model, for the given value of the internal Weber fraction (w), and the pattern of results for each condition.

feasible, despite its being more accurate than the alternative Subtraction Procedure. Not only the does meaning of *Most of the dots are yellow* appear to have a bias towards ANS-based verification procedures over correspondence-based procedures, but even within the range of ANS-based procedures that would compute the appropriate truth condition (albeit stochastically) there appear to be asymmetries. Participants insisted on approximating the cardinality of the set of non-yellow dots as "the dots minus the yellow dots", despite the availability of a more direct and more accurate alternative procedure, suggesting that the subtraction sign in the conventional expression of the relevant truth condition should be taken to carry some "verificational weight". Subtleties abound here concerning the exact nature of this "verificational weight", given the relationship between set subtraction and cardinality subtraction and the equivalence of $|\mathrm{DOT} - \mathrm{YELLOW}|$ and $|\mathrm{DOT} - (\mathrm{DOT} \cap \mathrm{YELLOW})|$ and $|\mathrm{DOT} \cap \overline{\mathrm{YELLOW}}|$; for more detailed discussion see Lidz et al. (2008). But the main point is that no asymmetry between verification procedures at all is predicted if sentence meanings are verification-agnostic truth conditions, unstructured functions from worlds to truth values.

6. Conclusion

In this paper we have argued against the claim that a competent speaker's understanding of a sentence is exhaustively characterised by a truth condition. To do so we have presented evidence of asymmetries in speakers' willingness to use various verification procedures: in Experiment 1, an apparent bias to use algorithms approximating a cardinality comparison rather than those based on one-to-one correspondence, and in Experiment 2, an insistence on an indirect method of approximation. These asymmetries would be surprising if the only constraint on the choice of verification procedures for a sentence was the requirement that the procedure must implement the sentence's truth condition.

References

Cordes, Sara, Rochel Gelman, Charles R. Gallistel, and John Whalen: 2001, 'Variability signatures distinguish verbal from nonverbal counting for both large and small numbers', *Psychonomic Bulletin & Review* **8**, 698–707.

Cresswell, Max J.: 1985, *Structured Meanings: The Semantics of Propositional Attitudes*. MIT Press, Cambridge, MA.

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

Feigenson, Lisa: 2005, 'A double dissociation in infants representation of object arrays', *Cognition* **95**, B37–B48.

Feigenson, Lisa, Stanislas Dehaene, and Elizabeth Spelke: 2004, 'Core systems of numbers', *Trends in Cognitive Science* **8**, 307–314.

Hackl, Martin: in press, 'On the Grammar and Processing of Proportional Quantifiers: 'Most' versus 'More Than Half'', *Natural Language Semantics*.

Halberda, Justin and Lisa Feigenson: in press, 'Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, 6-year-olds and adults', *Developmental Psychology*.

Halberda, Justin, Jeffrey Lidz, Paul Pietroski, and Tim Hunter: 2007, 'Language and number: Towards a psychosemantics for natural language quantifiers'. Talk at 4th Hopkins Workshop on Language. Oct 12-14, Baltimore, MD.

Halberda, Justin, Michele Mazzocco, and Lisa Feigenson: in press, 'Individual differences in nonverbal estimation ability predict maths achievement', *Nature*.

Halberda, Justin, Sean F. Sires, and Lisa Feigenson: 2006, 'Multiple Spatially Overlapping Sets Can Be Enumerated in Parallel', *Psychological Science* **17**, 572–576.

Halberda, Justin, Len Taing, and Jeffrey Lidz: 2008, 'The age of 'most' comprehension and its potential dependence on counting ability in preschoolers', *Language Learning and Development* **4**, 99–121.

Lidz, Jeffrey, Justin Halberda, Paul Pietroski, and Tim Hunter: 2008, 'Comparison, Subtraction and the Psychosemantics of 'most''. Unpublished ms.

Nieder, Andreas and Earl K. Miller: 2004, 'A parieto-frontal network for visual numerical information in the monkey', *Proceedings of the National Academy of Sciences of the USA* **101**, 7457–7462.

Pica, Pierre, Cathy Lemer, Véronique Izard, and Stanislas Dehaene: 2004, 'Exact and approximate arithmetic in an Amazonian indigene group', *Science* **306**, 499–503.

Pietroski, Paul, Justin Halberda, Jeffrey Lidz, and Tim Hunter: 2008, 'Beyond Truth Conditions: An investigation into the semantics of 'most''. Unpublished ms.

Treisman, Anne and Stephen Gormican: 1988, 'Feature analysis in early vision: Evidence from search asymmetries', *Psychological Review* **95**, 15–48.

Wolfe, Jeremy M.: 1998, 'Visual memory: What do you know about what you saw?', *Current Biology* **8**, R303–R304.

Wynn, Karen: 1992, 'Addition and subtraction by human infants', *Nature* **358**, 749–750.