Title:

Individual differences reveal underlying similarities in meaning: quantifier-dependent choices in verification procedure are consistent across individuals

Authors

Malhaar Shah (corresponding author)
University of Maryland at College Park

Tyler Knowlton
Barnard College

Robert Hopkins
University of Maryland at College Park

Justin Halberda
Johns Hopkins University

Paul Pietroski
Rutgers University

and Jeffrey Lidz
University of Maryland at College Park

Acknowledgements

We are grateful to the members of the UMD Psycholinguistics Workshop, audiences at the 37th Human Sentence Processing Conference 2025 at the University of Maryland, College Park and the 51st Annual Meeting of the Society for Philosophy and Psychology 2025 at Cornell University. A special thanks goes to Sebastián Mancha for advice about the statistical analysis.

Corresponding author details

Malhaar Shah

email: mpshah@umd.edu

Address:
1401 Marie Mount Hall
7814 Regents Drive
College Park
MD 20742
United States of America

Individual differences reveal underlying similarities in meaning: quantifier-dependent choices in verification procedure are consistent across individuals

Abstract

Previous literature has investigated how quantificational sentences like *Most of the dots are blue* are understood, using visual tasks which force participants to estimate numbers of dots. In this paper, we report two experiments which show that a participant's performance on a task involving the quantifier *Most* can be predicted by their performance on a 'baseline' task which does not involve *Most*. Furthermore, this task does not need to be visual: an auditory task can be used to establish this prediction. We argue that this phenomenon is partly explained by different speakers sharing a specific, stable way of understanding the quantifier *Most* involving a subtraction operator. Because different speakers all represent *Most* this way, each of their behaviors is predictable, once that speaker's general acuity in non-linguistic number cognition is determined. We draw some consequences for the theories of linguistic meaning.

1 Introduction

Some mental representations are common to us all. For instance, different people likely share a representation of the number seven, at least in terms of the representation's content. Given a stimulus of seven dots and adequate time to reflect, all would agree that there are seven dots, and not six or eight. But behavior is often an imperfect reflection of the mental representations involved in causing the behavior (Chomsky, 1965; Pylyshyn, 1973). For instance, when flashed seven dots on a screen very quickly, two people (or animals) may vary in their internal confidence about the number of dots they saw (Gallistel and Gelman, 2000; Whalen et al., 1999). One person may think *around seven and at least six*, and guess they saw six, whereas another might think *around seven and no more than eight*, and guess they saw eight. But this difference in behavior does not, by itself, mean their representations of the number seven are different.

Likewise, people vary not only in their behavior when it comes to estimating one quantity, but also when it comes to discriminating two quantities (see Halberda and Odic, 2015 for overview). In fact, this relates to their ability to estimate a single quantity. That is, a person's ability to tell whether there are more blue dots on a screen than yellow dots depends on their ability to estimate the number of blue dots and their ability to estimate the number of yellow dots. This ability, like the task in the previous paragraph, relies on the properties of an individual's Approximate Number System (ANS), which is an evolutionarily ancient and modality-independent system for rapidly estimating the number of a group of objects without counting them (Dehaene, 2011). Other systems, like linguistic cognition, interact with the ANS.

Previous work has investigated this interaction through the study of quantifiers like *More* and *Most*, which seem to direct the ANS to operate in different ways. Two groups presented the same visual stimuli – e.g., arrangements of blue and yellow dots – showed different behavior in a sentence verification task, conditioned by the sentence they were verifying (Knowlton et al., 2021; Lidz et al., 2011; Pietroski et al., 2009). One group was presented the target sentence *Most of the dots are blue*, and the other group was presented a different target sentence, *More of the dots are blue*. Despite the logical equivalence between these sentences in contexts with only two sets of items (blue dots and yellow dots), the group who was presented *More of the dots are blue* outperformed the group who was presented *Most of the dots are blue* (Knowlton et al., 2021).

Knowlton et al. (2021) propose, following Lidz et al. (2011), that this is because the sentences are paired with different structured mental representations, i.e., different algorithms (Marr, 1982). Each sentence is paired with a different representation, and moreover, one of these, when deployed for verification, leads to worse performance than the other (see also Wellwood and Hunter, 2023). The sentence in (1a) was paired with (2a), and the sentence in (1b) was paired with (2b):

(1) a. More of the dots are blue

b. Most of the dots are blue

(2) a. THE NUMBER OF BLUE DOTS > THE NUMBER OF NON-BLUE DOTS

b. THE NUMBER OF BLUE DOTS > (THE NUMBER OF DOTS − THE NUMBER OF BLUE DOTS)

Despite these algorithms resulting in the same output for the same input in ideal circumstances, in practice, the additional step of subtraction with (2b) introduces some noise into the calculation (in a way we make precise in Section 2.2). Because of this noisier algorithm, a person using the algorithm in (2b) should *behave* worse on average than one using the algorithm in (2a). Thus, the pattern of results described above is consistent with one group verifying the sentence by representing the stimuli as (2a) when the stimuli were described by *More*, and the other group using the representation (2b) when the stimuli were described by *Most* (see also Figure 3 for expected performance by algorithm).

While this asymmetrical performance in the two conditions suggests distinct representations, it cannot support a claim of consistency in the representation of *Most* across speakers, as it only investigated between-groups performance. Some speakers could represent *Most* one way and others represent *Most* another way, with the differences obscured in aggregation. In fact, semanticists and philosophers of language think psychological variation of this sort is quite plausible (see Travis, 2006 for overview). And it is known that speakers *do* vary

in how confident they are in describing particular arrangements of dots with sentences like *Most of the dots are blue* (Ramotowska et al., 2024). Is such variation in behavior best explained by positing a difference in linguistic representation (say, by ascribing (2a) to some speakers, but (2b) to others)? Or is it better to maintain that the semantic representation in (2b) is shared by all speakers, and the variation stems from extra-linguistic factors?

We report here two within-individuals experiments aimed to answer this question, supporting the claim that different speakers all represent *Most* as (2b). Each experiment had two tasks. Experiment 1 uses an individual's performance on a visual task comparing spatially separated blue and yellow dots to predict their performance on a task with spatially intermixed multicolored dots described with the sentence *Most of the dots are blue*. The first task gives us a baseline – it allows us to measure the participant's ANS acuity (in a sense to be explained in §2). This measure allows us to then predict that participant's performance on the second task, involving the quantifier *Most*. To anticipate the results, the prediction improves when it is assumed that the participant represents *Most* as (2b) rather than (2a). Then, in Experiment 2, to address the potential confound caused by the use of many colors of dots, we use only blue and yellow dots. Exploiting the modality-neutrality of number cognition (Dehaene, 2011), Experiment 2 replaces the first task with an auditory task where participants compare the quantity of high-pitch beeps to the quantity of low-pitch beeps to predict performance on a visual task involving spatially intermixed blue and yellow multicolored dots described with the sentence *Most of the dots are blue*. Analysis of participant-level data for almost all participants suggests ascribing the algorithm in (2b) to the participant in the task involving *Most*. In essence, by being able to factor out the variation between individuals in the use of their *number* cognition, we reveal the consistency in their *linguistic* cognition. Interestingly, this consistency holds at a level finer-grained than informational or truth-condiitional content: speakers are consistent not only in their (ideal) judgments of truth and falsity, but also how the verification procedure that yields this judgment.

2 Background: Modeling individual performance

The logic of the experiments is to use a task that is representative of a participant's general sensitivity to numerosity (alongside the meanings and related algorithms proposed above) to predict that participant's performance in evaluating *Most*. The first task should be set up to provide as reliable a measure as possible of their general acuity with estimating quantities using their ANS.

In this section, we review some of the fundamentals of the psychophysics of the ANS. Those already familiar with this literature may skim or skip Section 2.1. Section 2.2 explains the the predicted psychological difference between the algorithms in (2a) and (2b), and Section 2.3 explains the particular models we assume.

2.1 Measuring ANS acuity with Weber fractions

A participant's ANS acuity can be described in terms of a single parameter, w, the Weber fraction (see Halberda and Odic, 2015; Odic and Starr, 2018 for overview). Essentially, w reflects the rate at which the amount of internal 'noise' increases as the numerosity being represented increases. This parameter is subject to variation between individuals (Chesney et al., 2015; Halberda et al., 2008, 2012), though a given individual's w is fairly stable.

To elaborate on the interpretation of w, consider first a standard way of modeling the ANS. Representations built by the ANS are often modeled as Gaussian tuning curves on a mental number line (Feigenson et al., 2004; Gallistel and Gelman, 2000). The mean of each curve is centered around the numerosity being represented and the standard deviation of those curves increases at a constant rate (linearly) with the number being estimated. A representation of a larger number is then 'noisier' than the representation of a smaller number, as represented by the width of the curve. This feature is known as scalar variability and captures the fact that estimating smaller numerosities is easier than estimating larger ones. Answering a question like *How many things were there?* can be thought of as sampling from the distribution centered on the true answer, and sampling from a narrower Gaussian is more likely to lead to the right answer than sampling from a wider one.

Given this model, w represents the rate at which the standard deviation of the tuning curves increases with the number being estimated. This is shown in Figure 1. A person with a w of 0.3 (shown in the shaded activations) will lose precision more quickly than a person with a w of 0.2 (shown in the hollow activations). So if asked *How many things were there?* and the true answer is 12, the participant with the larger w is more likely to mistakenly sample 16 than a person with a smaller w.

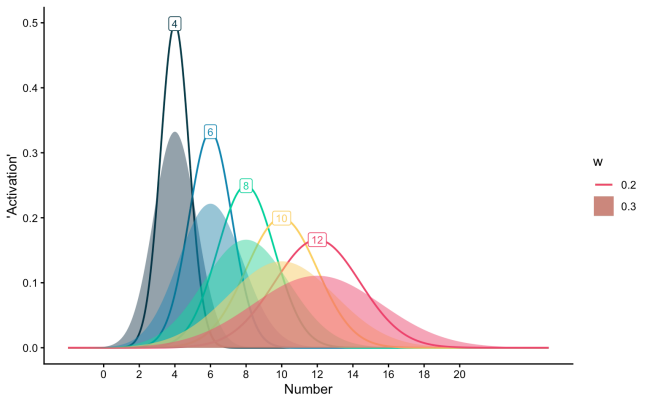

Figure 1: Ideal activation as number increases ($w = 0.2$ in hollow curves; $w = 0.3$ in shaded curves)

Given that successful discrimination of two estimates depends on the quality of each individual estimate, a smaller w also means a better ability to discriminate two quantities when the ratio between them gets smaller. When the ratio between the quantities is high (say, 2:1), they are easily discriminable for most people, regardless of their personal ANS acuity. That is, making a 6 vs. 12 comparison won't change too much if your w is .2 or .3, as there will be relatively little overlap in both cases. Likewise, When the ratio is low (say, 7:8), quantities become more difficult to distinguish for everyone.

Estimated accuracy at distinguishing two numbers (assuming participants make ideal use of their ANS) thus varies by both w and ratio, as plotted in Figure 2.

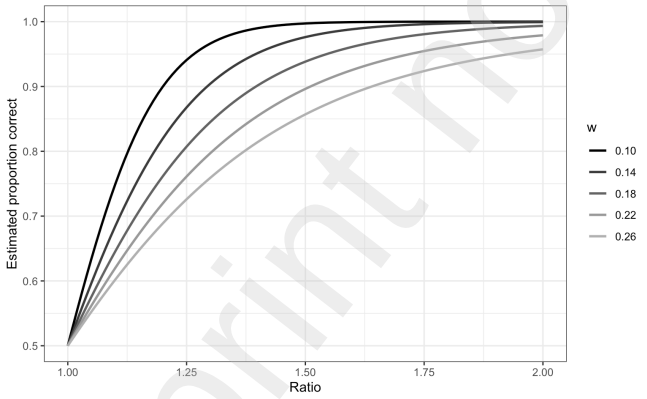

Figure 2: Ideal performance by w and ratio

This model assumes that all participants are, in ideal circumstances, at chance when there are exactly the same number of things being compared (a 1:1 ratio). It also assumes the same performance regardless of what sorts of things are being enumerated and compared. In Experiment 1, the things being compared are always dots; in Experiment 2, the things being compared are sometimes auditory pips and sometimes dots. Because the ANS is modality neutral (Barth et al., 2005; Dehaene, 2011; Feigenson et al., 2004; Izard et al., 2009), the ideal curve should be the same for both comparisons of pips in audition and comparisons of dots in vision.

In either case, the Weber Fraction w determines how well the participant performs as the ratio increases. As shown in Figure 2, a participant with a w as low as 0.1 reaches ceiling performance when the blue dots are in a 3:2 ratio to the non-blue dots (or vice versa). A participant with w of 0.26 is estimated to get only 95% of trials right even when the high pips are in a 2:1 ratio to the low pips (or vice versa).

The curves in Figure 2 are determined as follows. The estimate a participant makes of a quantity n (of blue dots, high beeps, etc.) is modeled, as noted earlier, as a Gaussian with a mean of the actual quantity and a variance that scales with that mean as a function of w. In particular, the variance for a given Gaussian is $w \cdot n$, where w is that individual's Weber fraction. Discrimination between two quantities can then be modeled as subtraction of the two Gaussian estimates.[1] Given this above model of estimation of a quantity, comparison of the blue dots and the non-blue dots (i.e., telling whether the blue dots outnumber the non-blue dots) is modelled by (3):

(3)
$$\frac{1}{2} \cdot erfc\left(\frac{n_{\text{blues}} - n_{\text{non-blues}}}{\sqrt{2} \cdot w \cdot \sqrt{n_{\text{blues}}^2 + n_{\text{non-blues}}^2}}\right)$$

Similarly, in an auditory task, using subtraction of the two Gaussians to comparing high pips and low pips, estimated performance at a ratio is given by the cumulative error function given in (4):

(4)
$$\frac{1}{2} \cdot erfc\left(\frac{n_{\text{highs}} - n_{\text{lows}}}{\sqrt{2} \cdot w \cdot \sqrt{n_{\text{highs}}^2 + n_{\text{lows}}^2}}\right)$$

We call the algorithm instantiated by this formula 'Direct Comparison' (DC), and it corresponds to the specification of content given in (2a).

[1] This subtraction should not be confused with the extra step of subtraction posited in the meaning of *Most*. The subtraction in this model is just a way of centering the distribution on zero, with the area to the right of zero representing the chance of answering 'true'.

2.2 Total Subtraction, an alternative to DC

The algorithm in (4) is not the only one available. There is also the specification of content in (2b). If participants instead compare the number of blue dots not to the number of yellow dots but to the total number of dots, their performance will be modeled by (5), which we call Total Subtraction (TS):

(5)

$$\frac{1}{2} \cdot erfc\left(\frac{n_{\text{blues}} - (n_{\text{total}} - n_{\text{blues}})}{\sqrt{2} \cdot w \cdot \sqrt{n_{\text{blues}}^2 + n_{\text{total}}^2}}\right)$$

The TS algorithm replaces the $n_{non-blues}$ argument with $n_{total} - n_{blues}$, making it less reliable than the Direct Comparison algorithm. Namely, it involves representing a bigger number (n_{total}), which is associated with a higher variance. Simply put, arriving at the number of things that aren't blue by subtracting the number of blue things from the total number of things introduces more variance than directly selecting the non-blues. As the algorithms differ in their efficacy, a participant's performance will differ depending on which algorithm they use, even when holding w constant. We plot the ideal performance of a participant with $w = 0.1$ by algorithm in Figure 3.

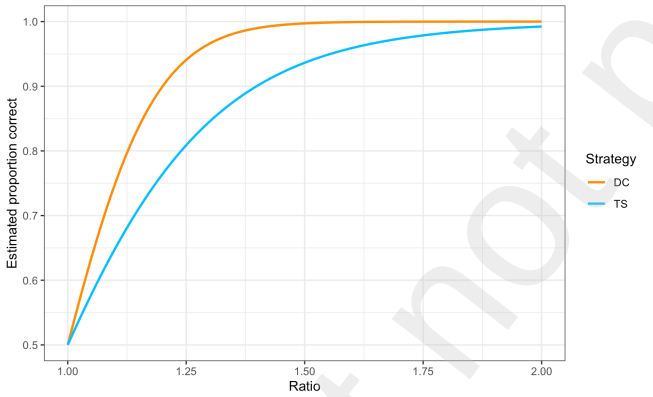

Figure 3: Ideal performance by algorithm (w = 0.1)

So if a participant's performance at comparing two cardinalities is worse than expected, there are two possibilities: their w is worse than expected, or they have chosen a worse algorithm. As a reminder, based on previous work, our hypothesis is that, because *Most* has a semantic representation that invokes the superset of dots/beeps, *Most* biases a participant towards the worse Total Subtraction algorithm.

The Total Subtraction algorithm still shows sensitivity to w, as shown in Figure 4, but the curves generated are quite different to those for the Direct Comparison algorithm, which were shown in Figure 2.

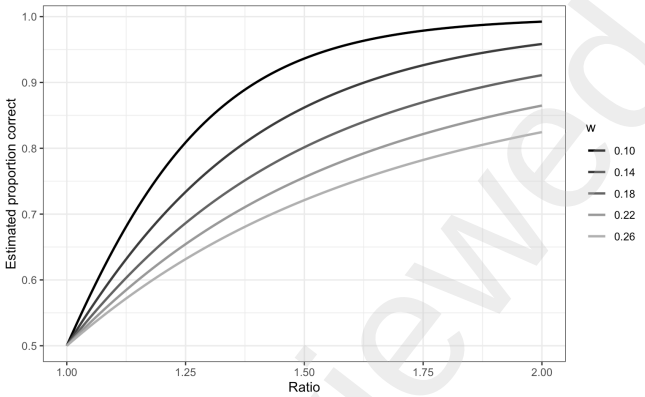

Figure 4: Ideal performance by w using Total Subtraction

Thus there are (at least) two variables at play: choice of algorithm, and w. If the observed data from a participant is represented by the crosses, two very similar curves may be drawn to fit the data: one using DC with a high w, another using TS with a low w. This is shown in Figure 5 and highlights the importance of getting an independent measure of w (if one cares about distinguishing whether participants use DC or TS).

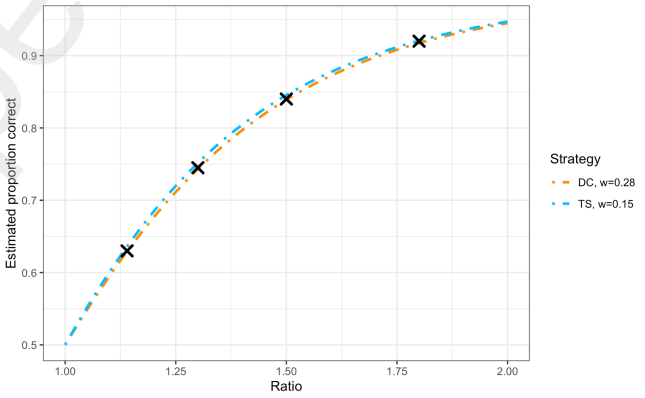

Figure 5: Similarity between the two algorithms with different values of w

In fact, we think there is a third relevant variable. Sometimes, participants will provide the wrong answer because they have failed to make ideal use of their capacity to estimate quantities. We call this 'guess rate' parameter g, though it is not intended to model a participant's rate of guessing *per se* – simply the proportion of failure that is not due to w or choice of algorithm (e.g., blinking during a trial). Intuitively, the guess rate represents the percentage of time that participants answer not using their ANS but by flipping a coin. Formally, g is incorporated into the models as follows:

(6) Direct Comparison model with guess parameter:

$$(1-g)\cdot\left(\frac{1}{2}\cdot erfc\left(\frac{n_{\text{blues}}-n_{\text{non-blues}}}{\sqrt{2}\cdot w\cdot\sqrt{n_{\text{blues}}^2+n_{\text{non-blues}}^2}}\right)+\frac{g}{2}\right)$$

(7) Total Subtraction model with guess parameter:

$$(1-g)\cdot\left(\frac{1}{2}\cdot erfc\left(\frac{n_{\text{blues}}-(n_{\text{total}}-n_{\text{blues}})}{\sqrt{2}\cdot w\cdot\sqrt{n_{\text{blues}}^2+n_{\text{total}}^2}}\right)+\frac{g}{2}\right)$$

A high g explains why participants are below ceiling at the highest ratios. The modeling assumptions here are (i) that participants will be at or near 100% on the easiest ratios if they exclusively use their ANS and (ii) that even with high guess rates, participants will still be at 50% correct if the ratio is 1:1. To illustrate, performance with the Direct Comparison algorithm as g varies is shown in Figure 6.

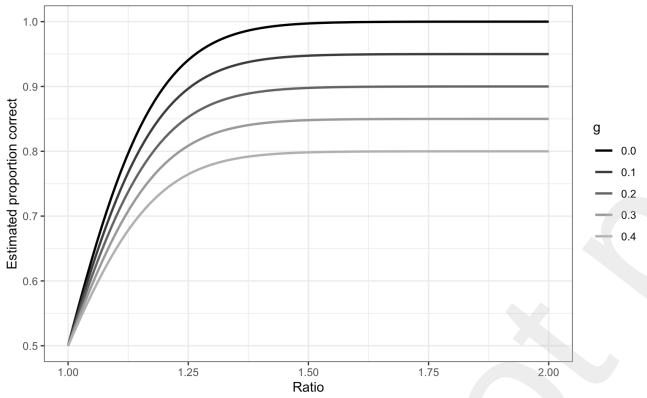

Figure 6: Ideal performance by g ($w = 0.1$)

In sum, a participant's error pattern in a number estimation comparison task results from a combination of their internal ANS acuity (w), the algorithm they choose to deploy (DC vs. TS), and the amount of time that they guess instead of relying on their ANS (g).

2.3 Modeling procedures

With these details established, we can recapitulate the logic of the experiments below. We take a participant's performance on one task to be representative of their general acuity for discriminating quantities ($w_{observed}$). This is their 'baseline' task: in Experiment 1, the baseline task is a visual task with *More*, in Experiment 2, the baseline task is an auditory task with *More*. We use *More* because *More of the dots are blue* is interpreted in these contexts as *More of the dots are blue than not* (Heim, 2000; Wellwood, 2019), and therefore invites the Direct Comparison strategy (Knowlton et al., 2021; Odic et al., 2013).

Seeing as their performance on a separate task, involving *Most*, is usually not identical to their performance on this 'baseline' task, we then ask how to explain their performance on the task involving *Most*. Specifically, we ask whether their performance the *Most* task is better fit by a model that assumes participants use the most reliable algorithm available to them – Direct Comparison – but that the initial estimate of w was wrong (i.e., that the noise in estimating w is what explains difference in model fit from condition to condition), or whether we should maintain that $w_{observed}$ is a good estimate but that participants change the algorithm they use when evaluating a sentence with *Most* instead of *More*.

Specifically, we use model comparison to choose between a three parameter model that uses different algorithms depending on the task (**Change-Algorithm** in (8)) and a four parameter model that uses the 'better' algorithm and allows w to vary with the task (**Same-Algorithm** in (9)). Guess rates (g) are fit independently (i.e., one for each task) in both models, as they are likely to be dependent on a number of variable factors (e.g., how interesting you find the task, tiredness, etc.). DC and TS refer to the Direct Comparison and Total Subtraction models specified above. P_i refers to the probability of giving a correct answer on trial i; $blues_i$ refers to the number of blue dots in that trial; $non\text{-}blues_i$ refers to the number of non-blue dots in that trial, etc.

(8) **Change-Algorithm (3 parameters):**

$$P_i=\begin{cases}(1-g_{\text{more}})\cdot DC(w, blues_i, non\text{-}blues_i)+\dfrac{g_{\text{more}}}{2} & \text{if task} = More\\ (1-g_{\text{most}})\cdot TS(w, blues_i, total_i)+\dfrac{g_{\text{most}}}{2} & \text{if task} = Most\end{cases}$$

(9) **Same-Algorithm (4 parameters):**

$$P_i=\begin{cases}(1-g_{\text{more}})\cdot DC(w_{\text{more}}, blues_i, non\text{-}blues_i)+\dfrac{g_{\text{more}}}{2} & \text{if task} = More\\ (1-g_{\text{most}})\cdot DC(w_{\text{most}}, blues_i, total_i)+\dfrac{g_{\text{most}}}{2} & \text{if task} = Most\end{cases}$$

In (8), w is fit based on the 'baseline' (*More*) task and used again in the *Most* task. Guess-rates are fit independently for each task, as noted above. Importantly, the choice of algorithm also varies across tasks (TS for *Most*; DC for *More*). This model thus encodes the idea that *Most* will be verified using the TS algorithm and that variation between tasks is explained by (i) this change in algorithm (ii) the differing guess rate.

In contrast, (9) fits separate w parameters for each task in addition to separate guess rates. The algorithm, however, remains constant across both tasks: The more precise algorithm (DC) is used in both cases. This model thus assumes that participants always use the same algorithm across tasks and that any differences between them are explained by an ANS acuity that varies between tasks (alongside a guess rate that varies).

It is psychologically implausible to interpret the Same-Algorithm model (9) as saying that people's w's change from day to day; w is roughly fixed for an individual. One can think of it producing two *estimates* of w, rather than hypothesizing two different ws. Seeing as the prior is that w does not vary, it means that there should be a penalty for providing two different estimates. The second comment is that we do not incorporate this penalty into our model selection. This means that we are *underpenalizing* the Same-Algorithm model.

Moreover, the Change-Algorithm model incorporates the constraint that changing from the DC algorithm to the TS algorithm can only make them *worse* (see Figure 3). As a result, if a participant performs worse on the baseline task than on the 'Most' task, the model will not be able to fit the data as straightforwardly as the Same-Algorithm model (unless the low performance in the baseline task is explained entirely by a high guess rate). Thus, the Change-Algorithm model is constrained in ways that the Same-Algorithm model is not.

Despite the fact that we underpenalize Same-Algorithm, and despite the fact that the Change-Algorithm model faces heavier constraints, we show below that model selection typically favors the Change-Algorithm model anyway (in both Experiment 1 and Experiment 2). This strongly suggests that $w_{observed}$ is a good estimate of participants' internal ANS acuity and that differences in performance between tasks are better explained by assuming that participants have changed the algorithm they use between tasks. Specifically, when evaluating *Most*, they opt to use the procedure in (2b/5).

All stimuli, data, and analysis scripts (for both experiments) are available in an OSF repository here.

3 Experiment 1

3.1 Methods and design

Experiments were conducted in person using PsychoPy Version 3 (Peirce et al., 2019). Participants signed a consent form, and sat in front of a 24 inch monitor (1920×1080; ≈ 53.1 cm wide and 29.9 cm high). Viewing distance was unconstrained but was around 60cm. 40 participants whose native language is English were recruited from [University] SONA system, and were awarded course credit. They signed up for two different days, on average 7 days apart. 11 participants were excluded for failing to respond to more than 24 trials (10% of total trials).

The target sentence on day one was always *Most of the dots are blue* and the target sentence on day two was always *More of the dots are blue*. On both days, participants saw 240 trials with 10 practice trials. On each trial, the target sentence was presented for 1500ms seconds before disappearing. The stimulus was then flashed for 1000ms. Participants responded 'True' or 'False' by pressing a button on the keyboard (either 'J' or right shift for True; either 'F' or left shift for false.). They could respond as soon as the stimulus was presented, but also for 1500ms after the stimulus vanished. Responses after this time elapsed were not counted. After each response, a prompt saying 'Press Spacebar to continue' would allow them to advance to the next trial.

Blue dots were present on every trial. The ratio of the blue dots to the other dots was one of three ratios: 2:1, 4:3, or 8:7 in trials where the target sentence was true; in trials where the target sentence was false, the ratio of blue dots to non blue dots was 1:2, 3:4, or 7:8. On half of the trials for each ratio, dots were 'area' controlled, meaning the total number of blue pixels was the same as the total number of non-blue pixels. On the other half of trials, dots were 'size' controlled, meaning the average size for the blue dots matched the average size for the non-blue dots. The trials were randomly ordered, in that the correct answer, ratio of blues:non-blues, and the control type used varied randomly.

On day one, with *Most*, participants saw the blue dots and the non-blue dots intermixed spatially. The blue dots could occupy the left 80% of the screen; the non-blue dots could occupy the right 80%. There were 5 colors of dots on the screen: each non-blue dot was assigned a color randomly out of yellow, magenta, red, and green. This made a visual selection of the group of all dots as straightforward as possible.

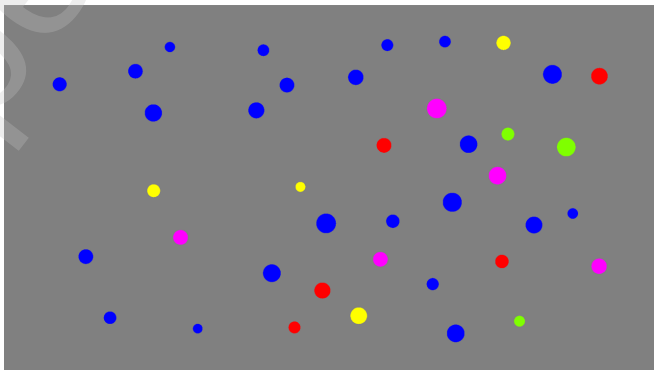

Figure 7: Stimuli for *Most of the dots are blue* (Exp. 1)

On day two, with *More*, participants saw the blue dots and the non-blue dots separated spatially. The blue dots were always on the left and the non-blue dots were always on the right. The non-blue dots were always yellow. This made the comparison of the two groups as straightforward as possible, allowing for a reliable estimate of a participant's acuity.

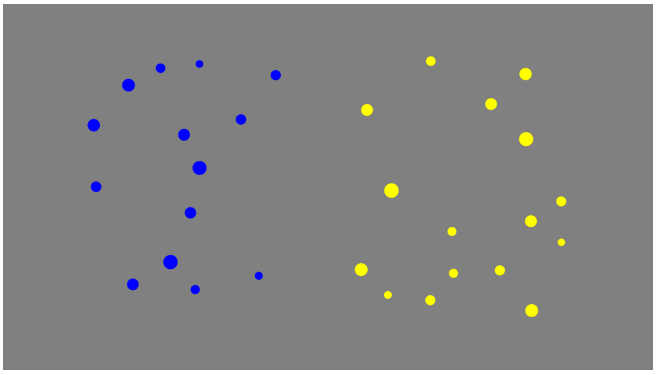

Figure 8: Stimuli for *More of the dots are blue*

Stimuli were varied from condition to condition for two reasons. The first reason is that the task with *More*/spatial separation of the dots is intended to provide the cleanest estimate of w possible. Making the task artificially more difficult would have led to a poorer measurement of the participant's w, as the task would be harder. The second reason is that, building on previous work (Knowlton et al., 2021; Lidz et al., 2011; Pietroski et al., 2009), providing a stimulus where the totality of dots are less salient would mean that the subtraction-based meaning of *Most* is unfairly disadvantaged. *Most*'s meaning, in (2b), requires (an estimate of) the total number of dots. If that set is very difficult to see, it might lead participants to ignore the meaning and choose a more convenient algorithm for addressing the experimental task. Instead, by presenting the spatially-overlapping array of multicolored dots, we make the collection of all dots salient, and therefore, do not disadvantage the stimuli for use with *Most*. We acknowledge the potential confound of using many colors, and resolve this in Experiment 2.

3.2 Results

For each participant, we fit Change-Algorithm and Same-Algorithm using Maximum Likelihood Estimation. We report Bayesian Information Criterion (BIC) results for each participant below. A lower BIC means a better fit; thus the BIC value of Same-Algorithm − the Information Criterion value of Change-Algorithm are positive in cases where Change-Algorithm is preferred.

These results are shown in Table 1, with estimates of w in each task. Both of these estimates are arrived at using the Direct Comparison algorithm, specified in (4). As a reminder, the Same-Algorithm model (whose success results in the Direct Comparison algorithm being selected) is underpenalized; there is no penalty for estimating two different values for w_1 and w_2. We show each participant's performance by ratio in Figure 9.

For 23 out of 29 participants, our model selection metric preferred Change-Algorithm, the model in which participants change algorithm. A Wilcoxon signed rank exact test indicated that ΔBIC is positive ($p < 0.001$).

3.3 Discussion

The results of this experiment confirm our main prediction: given within-individual baselines, almost every participant's performance in the task involving *Most* is better modeled by the subtraction strategy, i.e., the semantic representation in (2b). than the representation in (2a).

Participants for whom Same-Algorithm was preferred seem to have underperformed on the task involving the spatially separated dots/*More*, setting the wrong baseline for their performance.

However, it could be that the visual stimulus drove the choice to use the Total Subtraction algorithm. The Total Subtraction algorithm requires a representation of the set of all dots. On the other hand, the Direct Comparison algorithm requires a representation of the set of blue dots and the set of non-blue dots for comparison. But because there are many non-blue colors, it could be that the set of non-blues *per se* was less visually salient than the set of all dots, making the Total Subtraction algorithm preferable on purely visual grounds rather than the quantifier forcing this choice. While we think there are reasons to doubt this (Lidz et al. 2011 found no effect of number of color), we remedy this in Experiment 2, where only two colors of dots were presented.

4 Experiment 2

Experiments were conducted online on Cognition, and were implemented in jsPsych (de Leeuw et al., 2023). 48 participants living in the United States who self reported their first language as English and were between the age of 18 and 30 were recruited from Prolific, and were compensated monetarily. It was required that they were not using a tablet or a mobile phone for the experiment.

The experiment was split into two sections: a visual task and an auditory task. There were 126 visual trials (120 response trials; 6 practice trials) and 66 auditory trials (60 response trials; 6 practice trials); 192 trials overall (180 response trials; 12 practice trials). 11 participants were excluded for failing to respond at all to one or more tasks. 18 participants were excluded for achieving less than an average of 60% accuracy on either task, or for not responding to 10% of trials or more (participants often dropped the experiment after meeting the threshold for monetary compensation). Two further participants were excluded for taking an exceptionally long time to complete the task. This left 17 participants.

Auditory trials offered us a control unavailable in the purely visual setup of Experiment 1. As the ANS is modality-neutral, a person's w is estimable from an auditory task, while still

Participant no.	w_1 (baseline task)	w_2 ('Most' task)	ΔBIC	Algorithm selected
1	0.15	0.17	10.7	TS
2	0.15	0.25	5.83	TS
3	0.18	0.15	21.1	TS
4	0.16	0.56	7.99	TS
5	0.14	0.21	4.43	TS
6	0.15	0.38	6.39	TS
7	0.13	0.14	6.76	TS
8	0.19	0.23	0.435	TS
9	0.16	0.44	7.06	TS
10	0.14	0.19	6.25	TS
11	0.42	0.47	2.02	TS
12	0.16	0.24	4.23	TS
13	0.13	0.42	7.44	TS
14	0.19	0.23	10.9	TS
15	0.25	0.69	6.53	TS
16	0.44	0.40	-1.73	DC
17	0.10	0.18	5.34	TS
18	0.33	0.24	-3.58	DC
19	0.27	0.33	1.34	TS
20	0.12	0.22	5.52	TS
21	0.17	0.27	4.99	TS
22	0.28	0.28	-3.32	DC
23	0.14	0.17	6.61	TS
24	0.15	0.24	4.66	TS
25	0.12	0.25	5.47	TS
26	0.18	0.19	-4.06	DC
27	0.15	0.15	-5.44	DC
28	0.11	0.28	6.04	TS
29	0.17	0.33	5.31	TS

Table 1: Results from Experiment 1 by participant.
w_1 estimated by DC-algorithm on baseline task. w_2 estimated by DC-algorithm on 'Most' task.

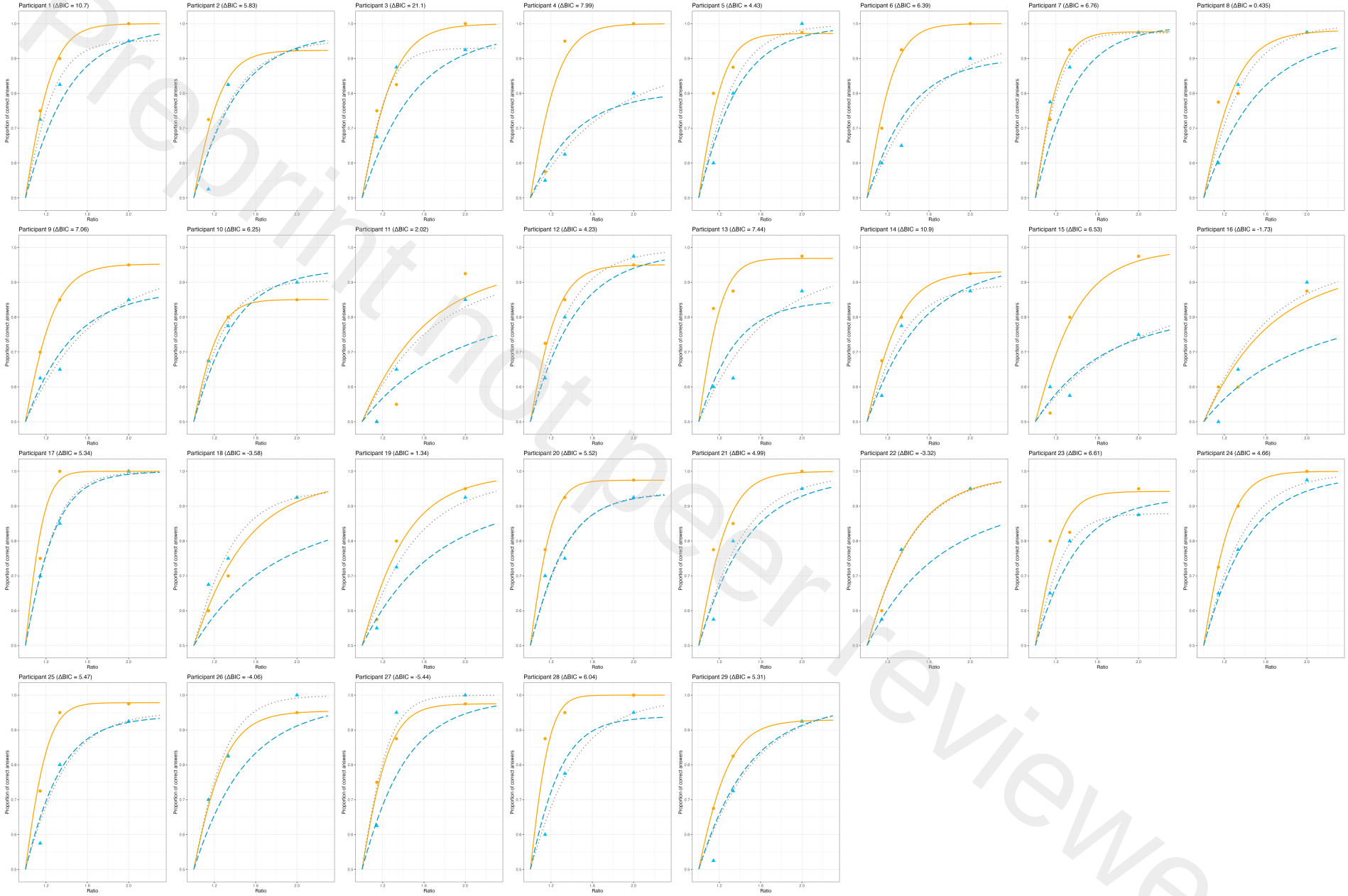

Figure 9: Experiment 1 results: performance across participants with ΔBIC.

serving as a predictor of performance in a visual task. Because auditory trials are most transparent when the streams are presented sequentially rather than simultaneously binaurally (Tokita and Ishiguchi, 2012), the cleanest estimate of w in the auditory task is with sequential presentation. By reducing the number of colors in the *Most*/Visual task to two, we also made it possible for participants to rely on the same algorithm (DC) for both parts of the experiment.

4.1 The visual task

In the visual task, participants were presented with a target sentence, which was always *Most of the dots are blue*, for 1000ms. Then, the stimulus of blue and yellow dots was flashed for 250ms. Typical stimuli are presented in Figure 10. The participant had 2500ms to respond by pressing 'J' if they thought the sentence accurately described the stimulus and 'F' otherwise. After the response window, a prompt saying 'Press SPACE to continue' appeared, after which the next trial would begin.

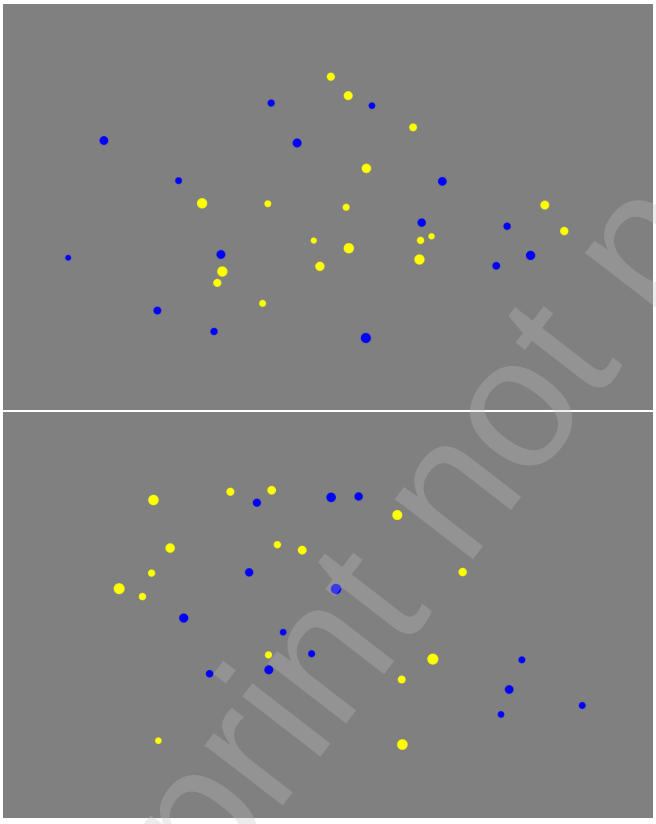

Figure 10: Typical stimuli for visual task (verifying the sentence *Most of the dots are blue*)

The dots appeared in the middle 80% of the screen, but could only overlap in 80% of the available region. On half the trials, the blue dots occupied the left region of the screen, and the yellow dots appeared on the right; in the other half of trials, this was flipped. The ratio of the blue dots to yellow dots was one of five ratios ratios: 2:1, 3:2, 4:3, 6:5, or 8:7 in trials where the target sentence was true; in trials where the target sentence was false, the ratio of blue dots to yellow dots was 1:2, 2:3, 3:4, 5:6, or 7:8. The smaller group of dots could be between 5 dots and 17 dots, as long as they were in the appropriate ratio to the larger group. Thus the largest number of dots on the screen at any time is 51 (17 dots in 2:1 ratio), and the smallest is 11 dots (5 dots in 5:6 ratio). On half of the trials for each ratio, dots were 'area' controlled, meaning the total number of blue pixels was the same as the total number of non-blue pixels. On the other half of trials, dots were 'size' controlled, meaning the average size for the blue dots matched the average size for the non-blue dots. Each condition was repeated once.

There were 120 trials with 6 practice trials. The trials were randomly ordered, in that the correct answer, ratio of blues:non-blues, which dots appeared on the left, and the control type used varied randomly.

4.2 The auditory task

In the auditory task, participants were presented with a target sentence, which was always *Which has more? The highs or the lows?*, for 1000ms. Then, a stimulus of a sequence of auditory pips was played. The pips were either of 400Hz or 700Hz. After the sequence ended, a cross appeared on screen and the participant had 2500ms to respond by pressing the Up arrow key if they thought the high beeps were more numerous, and the Down arrow otherwise. After the response window, a prompt saying 'Press SPACE to continue' appeared, after which the next trial would begin.

All pips were presented binaurally. On half the trials, the high pips were all played first, and the low pips played after the last high pip; in the other half of trials, this was flipped. The ratio of one group of pips to other was again always 1:2, 2:3, 3:4, 5:6, or 7:8. The smaller group of pips could be between 8 dots and 16 dots, as long as they were in the appropriate ratio to the larger group. Thus the largest number of pips a participant heard is 32 (16 pips in 2:1 ratio), and the smallest is 20 pips (8 pips in 2:3 ratio). On half of the trials for each ratio, pips were 'total duration' controlled, meaning the sum of the durations of the high pips was the same as the sum of the durations of the low pips. On the other half of trials, pips were 'average duration' controlled, meaning the average duration of a high pip matched the average duration of a low pip.

There were 60 auditory trials with 6 practice trials. The trials were again randomly ordered. There were fewer auditory trials than visual trials because these trials, involving sequences of pips, are much longer.

4.3 Results

For 14 out of 17 participants, analysis reveals that participants changed algorithm – even though the Direct Comparison algorithm was visually salient. Instances where Same-Algorithm was preferred are underlined.

Estimates for w_1 and w_2 using the Direct Comparison algorithm and results of model selection are shown by participant in Table 2. Performance across ratio by participant is shown in Figure 11.

A Wilcoxon signed-rank test indicated again that ΔBIC is positive, favoring the Change-Algorithm model overall ($p < 0.05$).

4.4 Discussion

Experiment 2 reveals that participants still use the subtraction strategy when given *Most*, even when the visual stimulus presented has only two colors (meaning that the non-blues are a visually salient set, making Direct Comparison viable). That is, participants are not forced into the Total Subtraction algorithm on visual grounds alone. Furthermore, we see that a participant's performance on an auditory task can be used to set up predictions for their performance on a visual task – meaning that w measured is a reflection of modality-neutral number cognition and not a property of visual cognition. Linguistic quantification seems then to interface with modality-neutral number cognition.

There are a few reasons that three participants were better fit by the Same-Algorithm model. For one, the Same-Algorithm model is generally underpenalized, as mentioned earlier. Another is that Same-Algorithm is preferred when the participant underperformed on the baseline task (the spatially-separated dots task in Experiment 1, and the auditory task in Experiment 2). This leads to the model setting the wrong baseline for their performance in the 'Most' task. Additionally, in Experiment 2, some participants may have found the auditory task more difficult than the visual task as the stimuli are presented over a longer stretch of time (see also Tokita and Ishiguchi, 2012 et seq.).

5 General discussion

The experiments discussed above aimed to find out whether different English speakers share the structured mental representation they pair with the sentence *Most of the dots are blue*. Both experiments revealed evidence that speakers are consistent in pairing them with the representation (i.e., algorithm) in (2b), repeated here:

(2) a. THE NUMBER OF BLUE DOTS > THE NUMBER OF NON-BLUE DOTS

b. THE NUMBER OF BLUE DOTS > (THE NUMBER OF DOTS − THE NUMBER OF BLUE DOTS)

That speakers would be consistent in pairing (2b) with the quantifier *Most* is surprising from a number of perspectives. First, it should be emphasized that the algorithms in (2) yield the same results, given a description of a scene. That is, the semantic representations in (2) share truth-conditions. Natural language semanticists and philosophers of language often hold that *all* that can be understood about a word's meaning is its contribution to the sentence's informational content, or its truth-conditions (e.g., Davidson, 1967, Lewis, 1970). But it seems that we need to capture distinctions at a finer grain than truth-conditions.

There is an alternative 'mentalist' view (following terminology in Williams 2015). This view takes sentence meanings to be structured concepts (Chomsky, 1986, 2000; Jackendoff, 1992; Katz and Fodor, 1963), or instructions to access concepts (Knowlton et al., 2021; Pietroski, 2017), internal to individual speakers' minds. Under this view, a structured mental representation is the meaning of a sentence. This allows for finer-grained distinctions than the truth-conditional theory. The representations in (2) have differing structures. But a purely truth-conditional view cannot see these as more than notational variants (Yalcin, 2018). Our empirical results support the view that such differences can make a difference psychologically, even if not logically, and that this is an important generalization across speakers which should be statable in a theory of meaning.

Chierchia & McConnell-Ginet (1990) discuss the competition between these two views. By representational theories, they mean what we have called 'mentalist' theories.

> "Proponents of representational theories [...] have focused on understanding as a matter of what interpreters can infer about the cognitive states and processes, the semantic representations, of utterers. You understand us, on this view, to the extent that you are able to reconstruct semantic representations like the ones on which we have based what we say. Communicative success depends only on matching representations and not on making the same links to situations." (p. 15ff)

The evidence we have found is quite consistent with such a view. This means that the possibility of communicative success does not militate for the objectivist view of semantics, as is often thought: there is reason to believe that, in some cases, different speakers share the structured mental representation they pair with sentences.

We want to emphasize that it does not follow from the mentalist view that speakers *always* use the algorithm in (2b) to verify sentences with the quantifier *Most*. The meaning of a sentence is only a part of what goes into determining what a speaker does when verifying it. Sometimes, other procedures may be more obvious or convenient. Suppose a task were set up so that the larger group of dots were always presented on

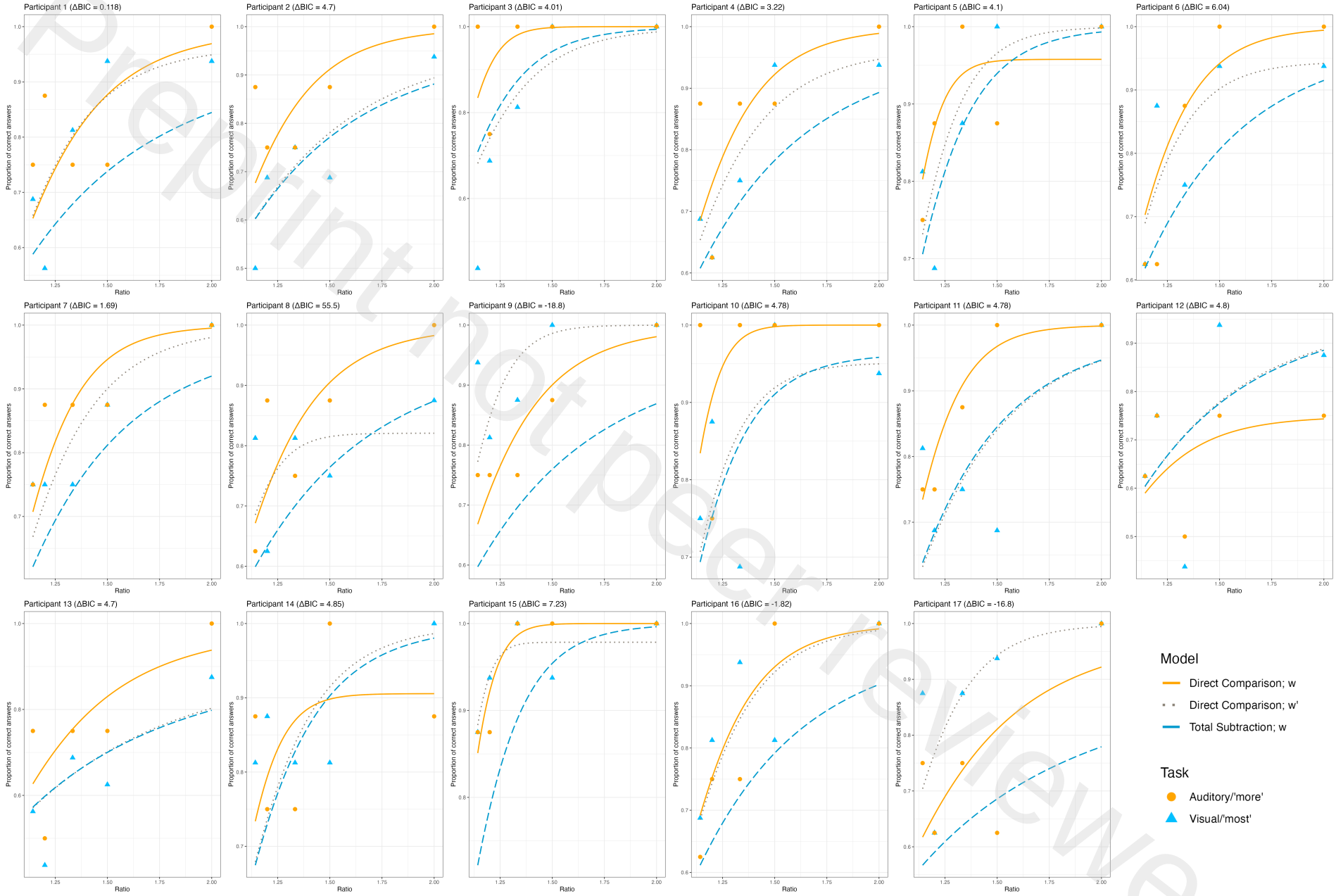

Figure 11: Experiment 2 results: performance across participants with ΔBIC.

Participant no.	w_1 (baseline task)	w_2 ('Most' task)	ΔBIC	Algorithm selected
1	0.24	0.21	0.118	TS
2	0.21	0.36	4.7	TS
3	0.10	0.20	4.01	TS
4	0.20	0.22	3.22	TS
5	0.10	0.15	4.1	TS
6	0.18	0.17	6.04	TS
7	0.17	0.22	1.69	TS
8	0.21	0.12	55.5	TS
9	0.22	0.13	-18.8	DC
10	0.10	0.15	4.78	TS
11	0.15	0.28	4.78	TS
12	0.20	0.35	4.8	TS
13	0.29	0.52	4.7	TS
14	0.12	0.20	4.85	TS
15	0.10	0.07	7.23	TS
16	0.19	0.20	-1.82	DC
17	0.32	0.18	-16.8	DC

Table 2: Results from Experiment 2 by participant.
w_1 estimated by DC-algorithm on baseline task. w_2 estimated by DC-algorithm on 'Most' task.

the left of the screen. To verify the sentence *Most of the dots are blue*, speakers would not need to bother to estimate the number of dots at all – they could simply look at whether the blue dots occupied the left of the screen or not. It is only in situations where there is no reason to pick an easier strategy that speakers will comply with the semantic representation of the sentence (for discussion, see Lidz et al., 2011).

Another caveat is in order. While we have found evidence for shared mental representations for *Most*, this need not be the case for every word (or even every quantifier). We have found evidence for shared structure only in a narrowly circumscribed area (quantifiers and their interaction with number cognition). For words like *furniture*, there is evidence for variation both between and within individuals: speakers change their minds from day to day on whether, say, *furniture* applies to curtains (McCloskey and Glucksberg, 1978; see also Marti et al., 2023). While it could be that even these have some, abstract, shared level of representation (Pietroski, 2018; Quilty-Dunn, 2021), the experiments reported here are independent of such claims.

Furthermore, even the consistency in quantifier terms is surprising. In most situations, we do not have any positive reason to think that different speakers share underlying structure. In Quine's (1960) vivid metaphor, "Different persons growing up in the same language are like different bushes trimmed and trained to take the shape of identical elephants. The anatomical details of twigs and branches will fulfill the elephantine form differently from bush to bush, but the overall outward results are alike." But here, it seems that speakers do not vary in pairing (2b) with *Most*. To continue Quine's metaphor, it seems that speakers are alike not only in the elephantine form, but in the organization of their inner twigs and branches.

How does this come to be? Given that the algorithms in (2) will yield the same answer in a given situation, it is not obvious how a child learning English regularly comes to pair (2b), and *not* (2a), with *Most*. We doubt that the syntactic structure of the sentences used to express the thoughts is any sure clue. Even if the underlying structure of quantificational sentences quite divorced from its surface form (e.g., May, 1985), the difference in algorithm is not predictable from a syntactic difference between *More of the dots are blue* and *Most of the dots are blue*.

(10) a. More As are B
= #(As that are B) > #(As that aren't B)

b. Most As are B
= #(As that are B) > #(As) $-$ #(As that are B)

This underscores the point we began with. Mental representations (including those which serve as the meanings of sentences) have structure. Identifying *which* structured mental representations are actually used in cognition is a matter of empirical investigation. Considered purely *a priori*, there is a wide variety of plausible candidates.

Focusing even on the narrow case of the English quantifier *Most*, where there are (at least) two plausible candidates for its representation, an algorithm we called Direct Comparison, and an algorithm we called Total Subtraction, we have found that there are surprising empirical consistencies in the mental representations used across speakers. Speakers consistently pair a strategy using subtraction with *Most*. They could have been inconsistent, or they could have consistently paired Direct Comparison to *Most*. But this was not the case. This suggests that there are interesting psychological generalizations about linguistic meaning, and that structured mental representations should be invoked to play an explanatory role in our theories of sentence-meaning and its acquisition.

References

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract Number and Arithmetic in Preschool Children. *Proceedings of the National Academy of Sciences*, *102*(39), 14116–14121.

Chesney, D., Bjalkebring, P., & Peters, E. (2015). How to Estimate How Well People Estimate: Evaluating Measures of Individual Differences in the Approximate Number System. *Attention, Perception, & Psychophysics*, *77*(8), 2781–2802.

Chierchia, G., & McConnell-Ginet, S. (1990). *Meaning and Grammar: An Introduction to Semantics*. MIT.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Praeger.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Davidson, D. (1967). Truth and Meaning. *Synthese*, *17*(1), 304–323.

Dehaene, S. (2011). *The number sense: how the mind creates mathematics* (Rev. and updated ed). Oxford university press.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core Systems of Number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Gallistel, C. R., & Gelman, R. (2000). Non-Verbal Numerical Cognition: From Reals to Integers. *Trends in Cognitive Sciences*, *4*(2), 59–65.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number Sense across the Lifespan as Revealed by a Massive Internet-based Sample. *Proceedings of the National Academy of Sciences*, *109*(28), 11116–11120.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual Differences in Non-Verbal Number Acuity Correlate with Maths Achievement. *Nature*, *455*(7213), 665–668.

Halberda, J., & Odic, D. (2015). The Precision and Internal Confidence of Our Approximate Number Thoughts. In *Mathematical Cognition and Learning* (pp. 305–333, Vol. 1). Elsevier.

Heim, I. (2000). Degree Operators and Scope. *Semantics and Linguistic Theory*, *10*, 40.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn Infants Perceive Abstract Numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382–10385.

Jackendoff, R. (1992). *Semantic Structures*. MIT Press.

Katz, J. J., & Fodor, J. A. (1963). The Structure of a Semantic Theory. *Language*, *39*, 170–210.

Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., & Lidz, J. (2021). Linguistic Meanings as Cognitive Instructions. *Annals of the New York Academy of Sciences*, *1500*(1), 134–144.

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

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

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press.

Marti, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent Diversity in Human Concepts. *Open Mind*, 1–14.

May, R. (1985). *Logical Form: Its Structure and Derivation*. MIT Press.

McCloskey, M. E., & Glucksberg, S. (1978). Natural Categories: Well Defined or Fuzzy Sets? *Memory & Cognition*, *6*(4), 462–472.

Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2013). Young Children's Understanding of "More" and Discrimination of Number and Surface Area. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 451–461.

Odic, D., & Starr, A. (2018). An Introduction to the Approximate Number System. *Child Development Perspectives*, *12*(4), 223–229.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in Behavior Made Easy. *Behavior Research Methods*, *51*(1), 195–203.

Pietroski, P. (2017). Semantic Internalism. In J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (2nd ed., pp. 196–216). Cambridge University Press.

Pietroski, P. (2018). *Conjoining Meanings: Semantics without Truth Values*. Oxford University press.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind & Language*, *24*(5), 554–585.

Pylyshyn, Z. W. (1973). The Role of Competence Theories in Cognitive Psychology. *Journal of Psycholinguistic Research*, *2*(1), 21–50.

Quilty-Dunn, J. (2021). Polysemy and Thought: Toward a Generative Theory of Concepts. *Mind & Language*, *36*(1), 158–185.

Quine, W. V. O. (1960). *Word and Object*. The MIT Press.

Ramotowska, S., Haaf, J., Van Maanen, L., & Szymanik, J. (2024). Most Quantifiers Have Many Meanings. *Psychonomic Bulletin & Review*, *31*(6), 2692–2703.

Tokita, M., & Ishiguchi, A. (2012). Behavioral Evidence for Format-Dependent Processes in Approximate

Numerosity Representation. *Psychonomic Bulletin & Review*, *19*(2), 285–293.
Travis, C. (2006). Psychologism. In E. Lepore & B. C. Smith (Eds.), *The Oxford Handbook of Philosophy of Language* (pp. 103–26). Oxford University Press.
Wellwood, A. (2019). *The Meaning of More*. Oxford University Press.
Wellwood, A., & Hunter, T. (2023). Linguistic Meanings in Mind. *Behavioral and Brain Sciences*, *46*, e289.
Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal Counting in Humans: The Psychophysics of Number Representation. *Psychological Science*, *10*(2), 130–137.
Williams, A. (2015). *Arguments in Syntax and Semantics*. Cambridge University Press.
Yalcin, S. (2018). Semantics as Model-Based Science. In D. Ball & B. Rabern (Eds.), *The Science of Meaning* (1st ed., pp. 334–360). Oxford University Press.