Uniformity of verification strategies within and across individuals:
Predicting performance with *most*-sentences from performance with *more*-sentences

Malhaar Shah (Univ. of Maryland), Tyler Knowlton (Univ. of Delaware),
Justin Halberda (Johns Hopkins), Paul Pietroski (Rutgers), Jeffrey Lidz (Univ. of Maryland)

Mainstream semantic theory identifies the meanings of sentences with their truth-conditions. This makes it unable to distinguish between truth-conditionally equivalent representations like (2a) and (2b), both of which provide the same answers (given a context) for the sentence in (1).

(1) Most of the dots are blue.
(2) a. #(blue dots) > #(non-blue dots) b. #(blue dots) > #(dots) − #(blue dots)

With this in mind, recent psycholinguistic work has argued for the explanatory virtues of a level of representation that distinguishes (2a-b) [e.g., A-D]. For instance, given that the Approximate Number System represents larger quantities with more 'noise' and that (2b) necessarily involves larger numbers than (2a), the strategy in (2a) is a superior algorithm for evaluating the sentence in (1). But even in contexts where (2a) is viable, participants have been shown to eschew it in favor of the inferior (2b) [D]. Only by distinguishing (2a-b) can this finding be explained.

That said, support for this idea comes from between-subjects experiments, leaving open the question of how consistent such results are within individuals. Here, we remedy this concern by showing that **participants who are asked to evaluate *most*-sentences perform as expected given (i) the verification strategy in (2b) and (ii) an independent measure of their personal numerical approximation acuity.** That is, we find remarkable consistency with respect to the strategies used to verify proportional sentences (in English) both across and within individuals.

Participants (n=30) performed two speeded truth-judgment tasks on different days. On Day 1, they judged the *most*-sentence in (1) relative to 100 multi-colored dot displays shown for 1 second each; on Day 2, they judged the sentence *more of the dots are blue* relative to 100 blue and yellow dot displays shown for 1 second each (Fig. 1). Difficulty in both tasks was varied by modulating the ratio of blue dots to non-blue dots (closer ratios are harder to reliably distinguish; e.g., 180 vs. 200 is far harder than 10 vs. 20, despite the absolute difference being larger).

Participants generally performed better with *more* than *most*, replicating prior work [D]. Our question is whether this reflects uniform behavior across subjects, modulo individual differences in approximation abilities. Namely, by assuming (i) that participants use the strategy in (2a) for *more* and the strategy in (2b) for *most*, and (ii) that they have a stable numerical approximation acuity, can we (iii) predict their performance on *most* from their performance on *more*?

By and large, we find that we can (Fig. 2). We modeled individual performance as a function of two parameters: numerical estimation acuity (which captures a participant's robustness to the difficulty of the task) and propensity to guess (which captures their rate of answering incorrectly on trivially easy trials). We compared two models: a four-parameter model that allows both acuity and guess rate to vary between days and assumes the superior algorithm in (2a) for both *more* and *most* versus a three-parameter model that allows guess rate to vary across days but uses a single numerical estimation acuity and requires the inferior (2b) algorithm be used for *most*. The relative likelihood of the three-parameter model over the four-parameter model was at least .7 for every participant tested and the average relative likelihood of the three-parameter model was .809 (Fig. 3). This suggests that the model that assumes (2a) for *more* and (2b) for *most* (along with a single parameter capturing numerical estimation acuity) is preferred.

This result supports the idea that when evaluating sentences like {*most*/*more*} *of the dots are blue*, speakers rely on two truth-conditionally equivalent yet psychologically distinct representations. Given this starting assumption, we were able to consistently predict within-subjects performance on *most*-sentences and explain between-subject variation in terms of individual differences in numerical cognition. Such consistency within and between individuals supports the view that speakers represent meanings at a finer grain than truth-conditions and that psycholinguistic methods are essential for elucidating the detail of these representations.

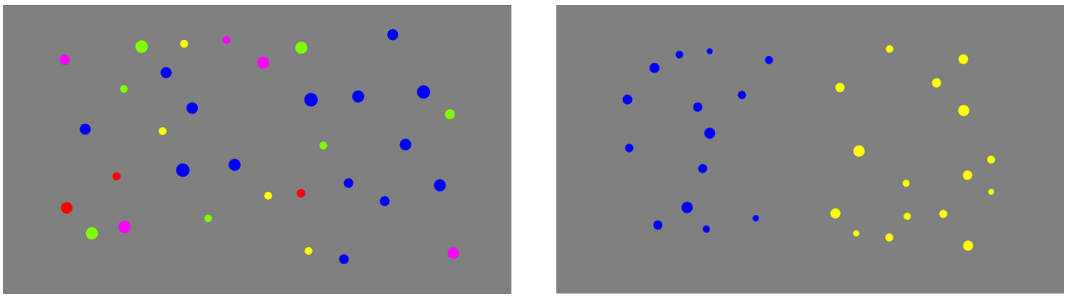

Figure 1. Left: stimuli for the *most* condition; Right: stimuli for the *more* condition

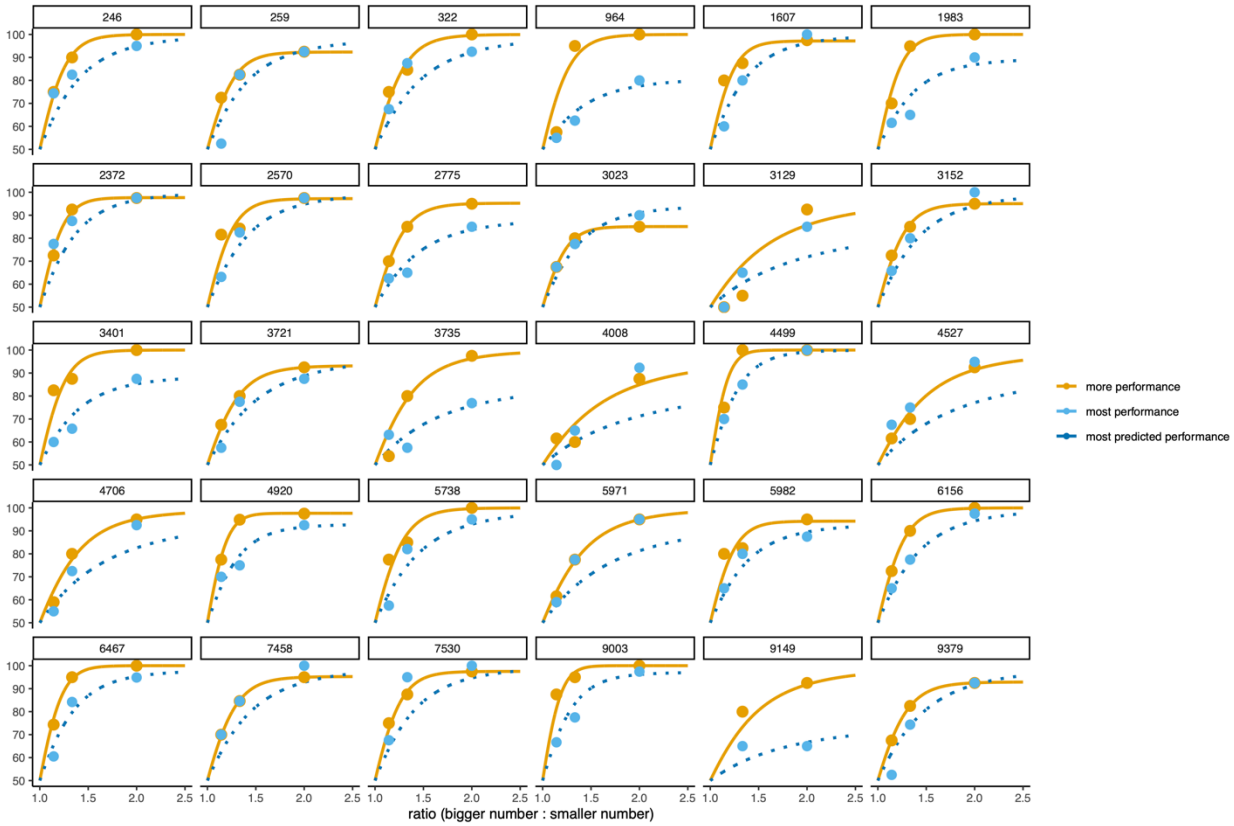

Figure 2. Performance in both conditions, by subject, with predicted *most* performance based on the subtraction algorithm in (2b) and each participant's numerical approximation acuity (fixed across both conditions) and propensity for guessing (variable across conditions)

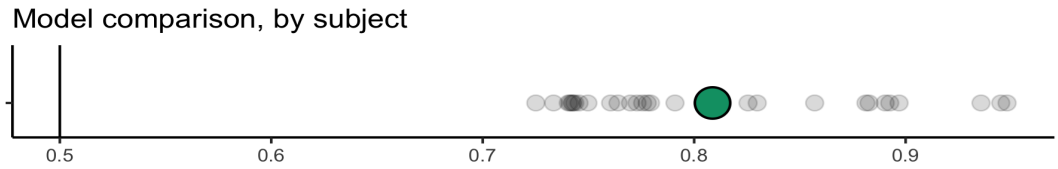

Figure 3. Relative likelihood of the 3 parameter model over the 4 parameter model

Refs. [A] Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009) The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind & Language*, *24*(5), 554–585. **[B]** Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011) Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256. **[C]** Hackl, M. (2009) On the grammar and processing of proportional quantifiers: most versus more than half. *Natural language semantics*, *17*, 63-98. **[D]** 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.