Semantic verification is flexible and sensitive to context

Jenna Register, Francis Mollica, Steven T. Piantadosi
Brain & Cognitive Sciences
University of Rochester

Abstract

Recent work in linguistics and psycholinguistics has sought to understand the cognitive foundations of semantic meaning for function words like *most* (Pietroski, Lidz, Hunter, & Halberda, 2009; Hackl, 2009; Lidz, Pietroski, Halberda, & Hunter, 2011). This work has put forth a variety of theories about the meaning of *most*, including that it naturally relies on cardinality, its verification process is closely tied to its underlying semantics, and that logically equivalent meanings of *most* have been distinguished using simple psychophysical tasks. Here, we show that subjects' performance in truth judgments tasks for *most* likely relies on a wide family of strategies, with a bias towards using cardinality-based strategies. The choice between these strategies is context-dependent, variable between participants, and sensitive to task-level factors like the number of trials participants are asked to complete. These results indicate that there is unlikely to be a single, simple formalization how *most* is computed, even in controlled laboratory tasks involving dot arrays. Instead, progress may be made by studying the way in which language users flexibly deploy a variety of different verification procedures.

1 Introduction

> "'Well! I've often seen a cat without a grin,' thought Alice 'but a grin without a cat! It's the *most* curious thing I ever saw in my life!"'
>
> — *Lewis Carroll, Alice in Wonderland*

Part of the power and complexity of human language arises from the existence of *function words* which do not refer to observable entities in the world but instead express logical relationships between other elements of sentences. Theories of what function words mean are deeply connected to representational hypotheses in syntax and semantics. For instance, the meaning of the function word *most* is determined in part from the syntactic positions where it can be used—i.e., what type of arguments it requires—and in part by the semantics that distinguishes it from other words of the same syntactic type (e.g. *every*). Traditionally, semantic theories have attempted to define the meaning of *most* using abstract logical characterizations of *what* words mean rather than *how* the meanings are computed. For example, *most* can be captured as a relation between sets (Barwise & Cooper, 1981): most A are B if $|A \cap B| > |A \setminus B|$.

However, there are multiple different psychological processes that could compute this meaning. Recent research in linguistics and psycholinguistics has attempted to evaluate possible algorithms as distinct psychological theories (Pietroski et al., 2009; Lidz et al., 2011; Hackl, 2009). For example, in the cardinality computation for these sets $|\cdot|$, one could count cardinalities exactly, one-to-one match corresponding items, or use estimation in the Approximate Number System (ANS) (Feigenson, Dehaene, & Spelke, 2004; Dehaene, 2011). Each of these provides a distinct psychological hypothesis and distinguishing them allows us to develop concrete, empirical theories of the algorithms supporting language comprehension.

To distinguish such alternatives, Pietroski et al. (2009) manipulated the ease of employing different algorithmic verification procedures in a truth value judgement task. Participants were shown two colors of dots that were either (a) intermixed randomly, (b) arranged in pairs (to encourage pairing strategies) or (c) arranged in a line (to encourage length-based strategies). The images were accompanied by the statement *"Most of the dots are yellow."* and participants were asked to judge whether the statement was true. Their results indicated that subject performance was in accordance with the psychophysical model of the ANS. That is, despite stimuli which were explicitly created to encourage use of a one-to-one correspondence strategy, subjects neglected this possible algorithm and instead used estimation to approximate the sets' cardinality. This is interesting in part because one-to-one correspondence could have allowed a more precise and exact answer, whereas the ANS inherently has more limited accuracy. In an extension of this work, Lidz et al. (2011) put forth two key claims about the meaning of *most*: (i) the meaning of *most* relied critically on a notion of *cardinality*; (ii) speakers use the underlying semantics to determine a "default" verification procedure meaning that one specific algorithm or means of cardinality comparison should be inherently favored over others.

Our experiments were designed to test evaluation of *most* in a neutral behavioral framework that allows the "default" semantics of *most* to be clearly studied. Subjects in Pietroski et al.'s experiments were run through 360 trials, where dots were presented for 200 ms each. It could be the case that running through so many trials may have encouraged subjects to adopt a speed/accuracy tradeoff in order to simply complete the task as quickly as possible. In this case, their behavior would not reflect any inherent semantic properties of *most*, but rather task-based strategizing. To address this, we ran an experiment with a single trial for each participant with unlimited viewing time. Our results indicate that participants deploy a wide variety of cardinality-based semantic strategies depending on the context, suggesting that the psycholinguistic system flexibly adopts a wide variety of semantic strategies (Experiment 1). Semantic comprehension, in turn, rapidly adapts to linguistic context. To explain Pietroski et al.'s observed pattern of ANS-based computation, we then show that participants switch to ANS-based responses when asked to complete many trials (Experiment 2). Experiment 3 shows that subjects have a bias to compute *most* through cardinality but individuals sometimes still used area when these cues are in conflict, further highlighting the flexible and variable nature of semantic verification. Supporting this view, Experiment 4 shows that *most* can be accurately and rapidly computed with a large variety of different input formats, including some where the cardinalities are not available. This shows that the semantic algorithms people find "natural" extend beyond those involving cardinality. Together, these experiments paint a picture that semantic processing is rapid, flexible, and fluidly varies across individuals and contexts. There is unlikely to be a single, simple answer to *how* people compute semantics, even for words whose meaning can be characterized with simple set operations.

2 Experiment 1

Experiment 1 asked participants to verify *most* across varied display contexts (intermixed, paired and lined) designed to favor different strategies, in a single trial with no time limit on presentation time. We use a truth value judgement with a free response prompt asking participants to describe how they generated their answer. Our main hypothesis was that strategy selection should be influenced by context. Therefore, we expect these explicit subjective reports to vary across contexts.

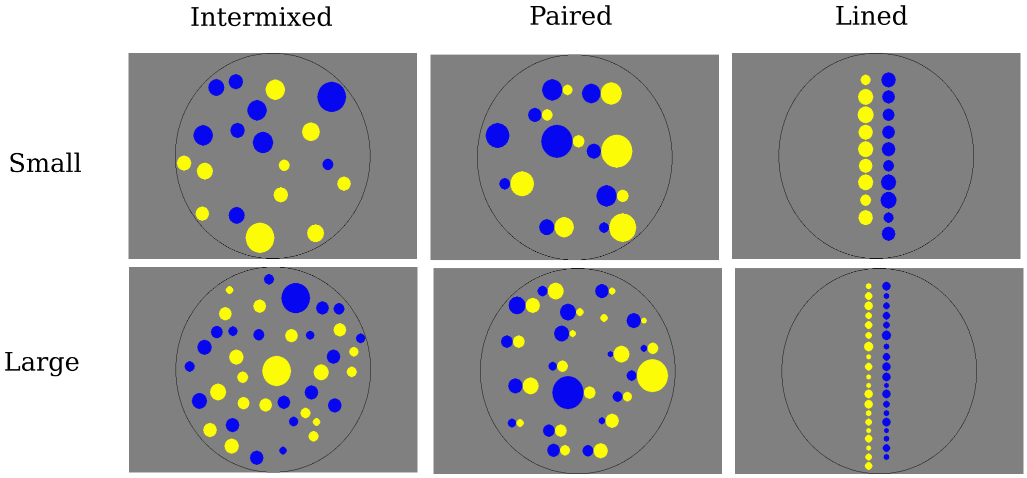

Figure 1: Example Intermixed, Paired, and Lined stimuli for both Large and Small Set Sizes.

Materials Following Pietroski et al. (2009), the stimuli consisted of area-controlled, non-overlapping yellow and blue dots on a gray background. There were three presentation contexts (see Figure 1). The intermixed context randomly scattered dots with a controlled area between colors. The pairing context horizontally matched-up blue and yellow dots scattered throughout the image, with the remaining dot (either blue or yellow) placed without a counterpart. The lined context vertically arranged dots of the same color with one line always containing one more dot than the other. To distinguish exact from approximate strategies, we presented numbers at two ratios, $9 : 10$ ("Small Set") and $19 : 20$ ("Large Set"). For each presentation context (see Figure 1), 10 images were created with a $9 : 10$ ratio (i.e., Small Set) and 10 images were created with a $19 : 20$ ratio (i.e, Large Set). The images were counterbalanced within each presentation context and set size, such that half of the images had more blue dots than yellow and the other half had more yellow dots than blue.

Procedure The experiment was presented online using psiTurk (McDonnell et al., 2012). Participants were randomly presented one of the 20 images. They were asked to judge the the statement *"Most of the dots are Blue. Y/N"* by pressing the "Y" or "N" key on their keyboard. No instructions were provided on how this judgement was to be made, as people's natural strategy was the main object of interest. Accuracy and response time were measured using custom javascript.

Next, participants were given an attention check question, to verify that they were earnest in their participation in the task. We asked participants how many red dots they saw flash during the experiment,

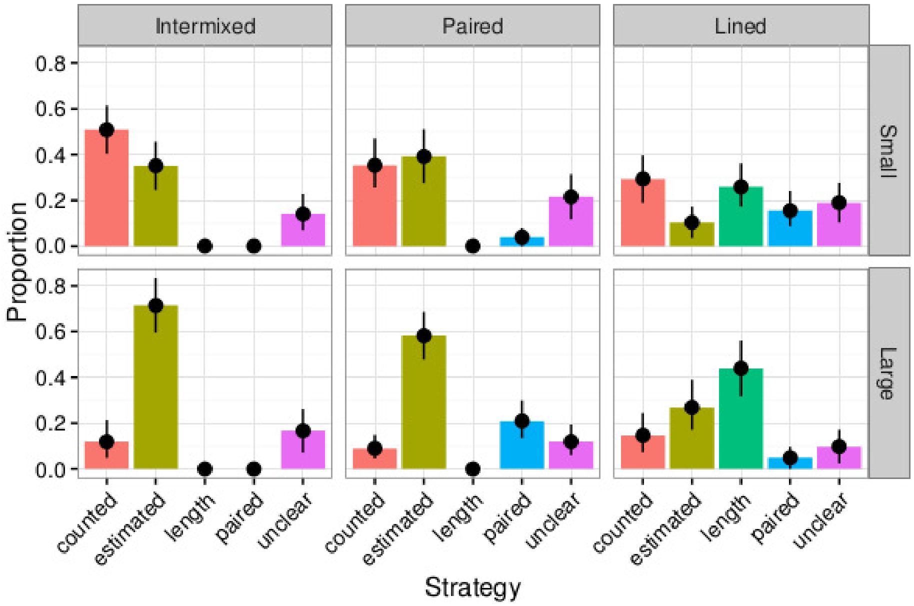

Figure 2: The distribution of strategies across presentation context and set size. Error bars represent bootstrapped 95% confidence bounds on each proportion.

when in fact there were never any red dots. We expected that participants who failed to attend during some part of the experiment would guess that some dots probably flashed and give a nonzero answer. The response times of those participants would not reflect use of their verification procedure and, therefore, we drop them from all analyses. Finally, participants were asked to self-report the strategy they used. Participants could freely describe their strategy in an open text-box.

Participants Three hundred and sixteen participants (Intermixed: $N = 99$, Paired: $N = 118$, Lined: $N = 99$) were recruited via Amazon Mechanical Turk and paid for their participation, which lasted approximately a minute. Twelve participants were excluded because they failed an attention check. Participants were not allowed to complete this experiment twice or to complete any of the other experiments presented in this paper.

2.1 Results and Discussion

Based on similarities across the verbal report, four common strategies were identified and coded by the first author: estimation, counting, pairing and using length. Participants who did not fall into these groups were marked as "unclear." Importantly, the *distribution* of reported strategies, shown in Figure 2, varied across Presentation Context ($\chi^2(8) = 108.15$, $p < 0.001$) and Set Size ($\chi^2(4) = 39.405$, $p < 0.001$), indicating that strategy deployment is influenced by context. For the Small Set size contexts, use of counting to *exactly* determine cardinality was often natural for participants (almost 50% in the intermixed condition). The Large-vs-Small Set size difference strongly suggests that use of counting vs. estimation is sensitive to the set size to be counted, with participants making a sensible strategic choice to estimate larger sets. The sensitivity to set size can also be seen in the Paired contexts, where the Small Set size does not give rise to pairing strategies as often as the Large Set size. Note that in each context different individuals typically reported using a variety of strategies.

Matching Pietroski et al.'s general conclusion, participants show an overall preference for cardinality-based strategies, even in contexts that encourage pairing. The Lined and Paired strategies only occur in the

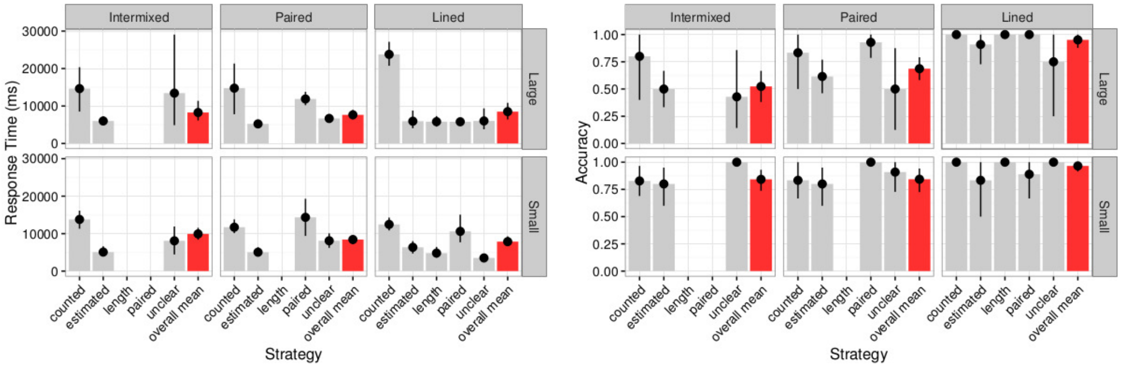

Figure 3: Response times and accuracies for each strategy across presentation context and set size. The average collapsing across strategy is given in the right of each cell in red. Error bars represent bootstrapped 95% confidence bounds.

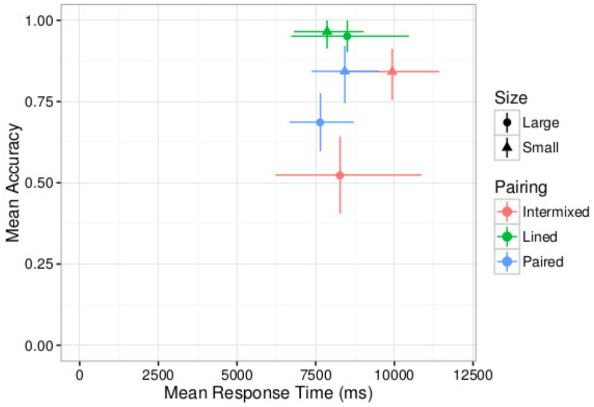

Figure 4: Mean Accuracy against Mean Response Time. Note that across Presentation Context and Set Size, accuracy ranges from chance to ceiling, while RT remains statistically insignificant.

Lined and Paired conditions respectively[1]. While pairing becomes more attractive in the Large condition ($\sim 20\%$), response patterns show a bias towards using cardinality-based strategies.

Response time and accuracy showed no variation across conditions when collapsing across reported strategy ($F(5) = 0.93$, $p > 0.4$), but did show differences across strategy in an omnibus analysis of deviation ($\chi^2(5) = 6.78$, $p < 0.001$). Detailed patterns of analysis within strategy and context show that that estimators respond faster ($t = -11.19$, $p < 0.001$) but less accurately ($t = -3.248$, $p < 0.01$) than counters. Note that the response time and accuracy for each strategy is very similar across contexts. Participants who reported estimating, for instance, take about the same time and achieve the same accuracy regardless of the presentation format or cardinality. We take this to suggest that the subjective-report likely reflects genuine strategy.

Moreover, the RT and accuracy patterns indicate that participants take advantage of a speed-accuracy trade-off, prioritizing speed over accuracy when selecting a strategy. Figure 4 shows how accuracy varies across different Presentation Contexts and Set Sizes, whereas RT does not differ significantly. This shows that participants were unwilling to spend more or less time than about 8 seconds on a trial, yet were relatively insensitive to losing accuracy.

[1] As the absence of a strategy itself is a significant change in a distribution, we conducted our analysis of the distribution removing length strategies from consideration and our conclusions hold for both context ($\chi^2(6) = 28.11$, $p < 0.001$) and set size ($\chi^2(3) = 38.74$, $p < 0.001$).

2.2 Conclusion

Experiment 1 demonstrates that there is a distribution of strategies for verifying the meaning of *most* that are naturally adopted in an unconstrained task. The results show that the deployment of strategies is influenced by the context and set size of the stimulus.

3 Experiment 2

Experiment 2 tested if participants' strategy choices vary with the number of trials they are asked to complete. Strategic choice of strategy predicts that over many trials participants should resort to a fast strategy, potentially indicating that findings that *most* rely on the ANS are driven by experimental design choices.

3.1 Methods

We presented participants with either four or twenty trials of the Small Intermixed context from Experiment 1, measuring response time, accuracy, and reported strategy.

Materials Twenty-four stimuli for the Small Intermixed context of Experiment 1 were generated for Experiment 2 using the procedure from Experiment 1.

Procedure The procedure was the same as Experiment 1 with a few exceptions. The instructions stated, *"This experiment consists of [n] trial(s) followed by 2 questions. On each trial, you will be presented with an image containing different colored dots and a sentence. Please judge this sentence to be true or false. If the sentence is true press the Y key. If the sentence is false press the N key."* Subjects were randomly assigned to either four or twenty trials. Stimuli were presented in random order for each subject. As in Experiment 1, the self-report response was asked at the end of the behavioral task—i.e., after all of the trials were completed. In the results, we include data from Experiment 1 as a comparison condition.

Participants One hundred and seventy five participants (Four Trials: $N = 89$, Twenty Trials: $N = 86$) were recruited via Amazon Mechanical Turk and paid for their participation. Twenty seven participants were excluded because they either failed the attention check as described in Experiment 1, or performed exactly at chance in a counterbalanced task with unrealistically short reaction times (indicative of holding down one key). Participants were not allowed to complete this experiment twice or to complete any of the other experiments presented in this paper.

3.2 Results and Discussion

Figure 5 shows that as the number of trials increases, the proportion of counters decreased, suggesting that participants were more willing to use time-intensive strategies for shorter tasks and switch to faster strategies that sacrifice accuracy for longer tasks. In fact, 10 participants explicitly reported switching to estimation after one or few trials; *e.g. "I counted a couple of images then tried to make an educated guess for the rest."* A chi squared test showed that reported strategy selection was affected by the number of trials in the task ($\chi^2(5) = 732.53$, $df = 12$, $p < 0.001$).

Next, we analyzed how accuracy and response time changed over trials, shown in Figure 6. These show a decreasing RT with more trials, likely indicating a switch to ANS-based responses, although differences are apparent even early in the experiment. A linear regression predicting response time from condition (1, 4, 20) and trial number showed an effect of trial number ($\beta = -1540$, $t = -7.2$, $p < 0.001$), condition ($\beta = -295$, $t = -10.09$, $p < 0.001$), and a significant interaction ($\beta = 70$, $t = 6.60$, $p < 0.001$) in the expected directions.

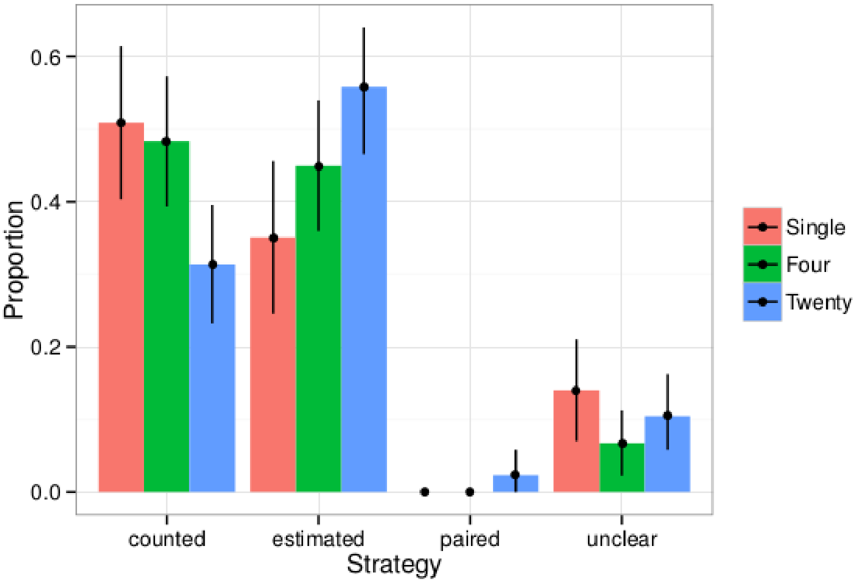

Figure 5: The distribution of strategies used across Four and Twenty Trial tasks. Error bars represent bootstrapped 95% confidence bounds on each proportion.

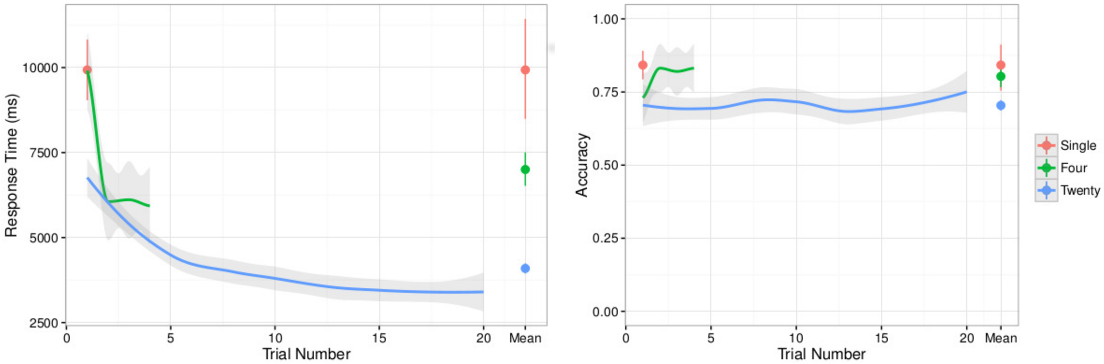

Figure 6: Mean response time and accuracy for each trial for both the Four and Twenty Trial tasks. The shading reflects the standard error of the mean. The floating points represent the mean collapsing across trials with bootstrapped 95% confidence bounds. Means for the one-trial condition run in Experiment 1 (above) have been re-plotted here for comparison.

3.3 Conclusion

The number of trials participants are asked to complete influenced strategy selection. These results are consistent with Experiment 1's finding of flexible deployment of strategies in the face of task demands. The results suggest that single-trial experiments should be strongly preferred for future work examining default mechanisms of linguistic interpretation.

4 Experiment 3

Perhaps the most central claim of the prior literature on the semantics of *most* is that it concerns the cardinality of sets. In a similar spirit, animal behavioral work has highlighted the importance of number in perceptual tasks. For instance, Cantlon and Brannon (2007) ran a matching task with macaques with direct conflict between cardinality and cumulative surface area. Monkeys saw a single image, and were then prompted to match this image with either an image with the same number of items or an image with the same total area. Their results indicated that even monkeys that were inexperienced with numerical tasks preferred matching to number than to surface area, suggesting the primacy of number over other correlated dimensions. We employed a similar paradigm to distinguish between a semantics of *most* that is based in cardinality as opposed to correlated dimensions like area. While the stimuli in Experiments 1 and 2 included area-controlled yellow and blue dots, Experiment 3 uses stimuli with conflicting area and number ratios between the yellow and blue dots to test whether *most* is sensitive to area over and above effects of number.

4.1 Methods

Materials Twenty stimuli (see Figure 7) were created using a similar process to that of Experiments 1 and 2. The area ratio between blue and yellow dots was manipulated. We used a 1:3 ratio for *number* and for *area.*

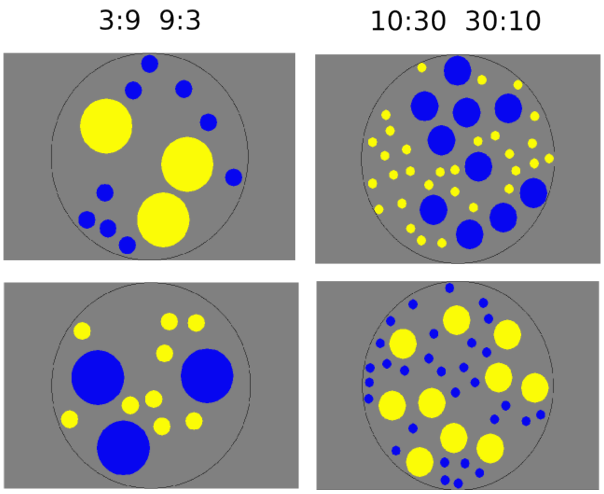

Figure 7: Example 1:3 in number and 3:1 in total areas

Procedure The procedure was the same as Experiment 1 with a few exceptions. Each participant still only received a single trial, and were randomly assigned to one of the following conditions: Area, Either, or Number. Stimuli were randomly selected from the same set across all conditions but the prompt that participants responded to was changed for each condition. For the Area condition, participants were asked to respond to the question "Most of this is yellow. (Y or N)". For the Number condition, "Most of these are yellow. (Y or N)". For the Either condition participants were given an ambiguous question, "Which has most? Yellow or Blue?". The critical response then is the proportion of number responses in the Either condition as it reflects the bias for cardinality over other dimensions inherent in "most." As in Experiment 1, the self-report response was asked at the end of the task.

Participants Three hundred and two participants (Area Bias: $N = 100$, Either Bias: $N = 102$, Number Bias: $N = 100$) were recruited via Amazon Mechanical Turk and paid for their participation, which lasted approximately one minute. Twenty-four participants were excluded because they failed the same attention check as described in Experiments 1 and 2.

4.2 Results and Discussion

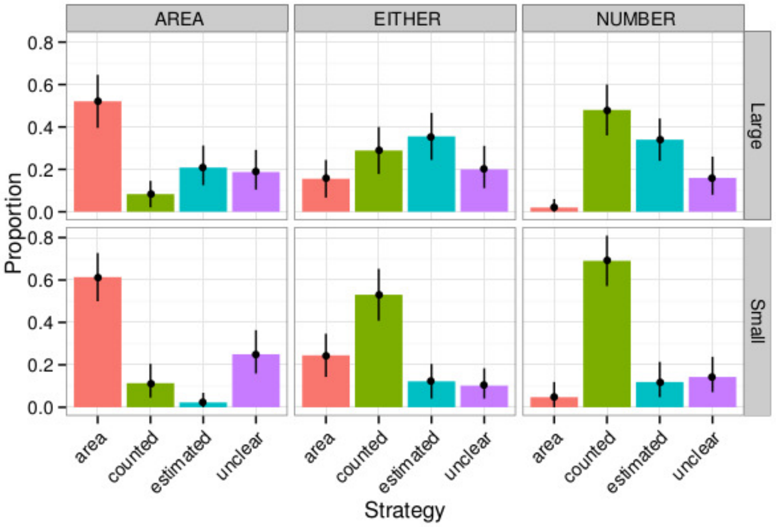

Figure 8: The distribution of strategies across presentation context and set size. Error bars represent bootstrapped 95% confidence bounds on each proportion.

Figure 8 shows the distribution of strategies among the different conditions. We expected to see area strategies used in the context with the area prompt ("Most of this is yellow"), either area or cardinality strategies used in the context with the Either prompt ("Which has most? Yellow or Blue?"), and cardinality strategies in the context with the Number prompt ("Most of these are yellow."). This matches the pattern in Figure 8, with area used in the Area prompt and counting and estimating used in the Number prompt. Importantly, in the Either prompt, there is a bias to use counting and estimating with a stronger bias to count small numbers. Overall, we find a significant effect of context on strategy selection ($\chi^2(6) = 87.073$, $p < 0.001$). Similar to Experiment 1, we also find an effect of Size on strategy selection ($\chi^2(3) = 22.033$, $p < 0.001$), particularly that the larger set size seems to call for estimation as the more appropriate strategy, even in these single trials.

By design, participant responses to the *most* prompt directly reflected whether they used number or area to respond. Figure 9 shows the proportion of participants who used a Number strategy (counting or estimating) to verify *most* in all three conditions. Strategy selection was sensitive to the context, as we found the lowest proportion of number users in the Area context, the highest proportion of number users in the Number context, and an intermediate in the Either context. A chi squared test reveals that the proportion of number users was significantly dependent on the prompt presented ($\chi^2(2) = 64.332$, $p < 0.001$). Critically, the pattern of results shows that there is a bias toward number, with participants responding according to Number nearly 50% of the time, even in the Area condition. At the same time, in the ambiguous Either condition, roughly 25% of participants did use area in responding, showing that the flexibility about the semantics of *most* across participants spans both numerical and quantity interpretations. Often, subjects

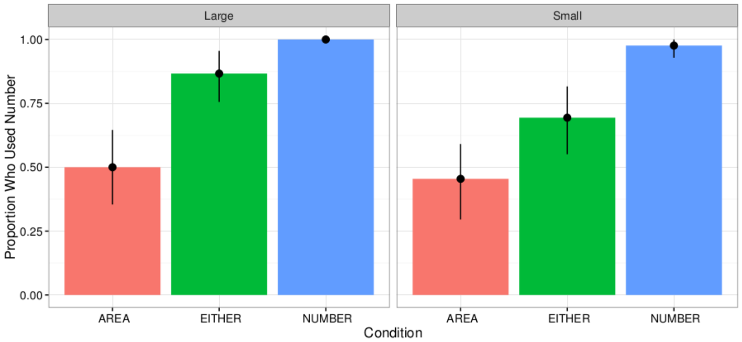

Figure 9: The proportion of participants who used Number (explicitly determined by answer and not subjective report) to verify *most* across the different conditions. Error bars represent bootstrapped 95% confidence intervals on each proportion.

had direct access to these choices: verbal reports included statements like *"I thought the word* most *implied area. The yellow dots seemed to cover the most area."* and *"The three yellow dots were much larger in area. Although there were numerically more small blue dots I felt the yellow was* most*"*.

4.3 Conclusion

Experiment 3 shows that while cardinality is a preferred semantics for *most*, in ambiguous contexts participants will often answer in accordance with Area cues that conflict with Number.

5 Experiment 4

The hypothesis that cardinality is the default or most natural meaning for *most* predicts a processing advantage for comparisons relying on cardinality as compared to these alternatives. The absence of such an advantage would support flexible, rapid deployment of the multiple semantic interpretations of *most*, akin to the multiple strategies available to the cardinality sense of *most* explored in Experiments 1, 2 and 3. In Experiment 4, we tested this processing advantage in a simple sentence verification task that manipulated the input format provided to participants.

5.1 Methods

We created a variety of scenarios that set up a *most* comparison (see Table 1). After reading a scenario, each participant was prompted with a truth value verification of a sentence with *most*. We measured accuracy and response time in search of a processing advantage—i.e., faster responses and/or increased accuracy, for the (cardinality-dependent) numerical condition compared to several cardinality-independent alternatives.

Procedure Participants were provided with the following instructions: *"This experiment consists of 1 trial, followed by 2 questions. In this trial, you will be presented with scenario and a sentence. Please judge this sentence to be true or false. If the sentence is true press the Y key. If the sentence is false press the N key."* The task displayed a scenario item (e.g. *"70% of the children in the class like Math, and 30% of the children like History."*, and the participant pressed the space bar to reveal a prompt (e.g. *"Most of the children like math."*). This was to ensure that response time is recorded *after* reading the scenario items, which each differed in length. After making a judgement, participants proceeded to an attention screening question. Each participant received a single item from one of the 11 categories (Area, Estimation, Greater than Half, etc.).

Contextual Cue	Scenario	Prompt
Numerical	“49 musicians play the accordion, and 21 musicians play the harp.”	“Most musicians play the harp.”
Area cues	“I was learning about the planet Earth.”	“Most of the planet is oceans.”
Estimation	“I was learning about the United States.”	“Most people in the US live in cities.”
Greater than Half	“More than half of the lights are off. The rest are on.”	“Most of the lights are off.”
Less than Half	“Less than half of the lights are off. The rest are on.”	“Most of the lights are off.”
Numerical Ratio	“70% of the children in the class like Math, and 30% of the children like History.”	“Most of the children like Math.”
OneToOne Pairing (plus one)	“For every book on the top shelf, there is a book on the bottom shelf. You add one book to the bottom shelf.”	“Most of books are on the top shelf.”
OneToOne Pairing (in 70/30 ratio)	“3 guests that come to my party bring both a cake and a pie. Then 4 friends show up with just a pie.”	“Most of the desserts at my party are cakes.”
Prototype	“I was talking about children with my friend.”	“Most children like to do homework.”
Reverse Numerical Ratio	“30% of the children in the class like Math, and 70% of the children like History.”	“Most of the children like History.”
Set Difference	“There are 5 more blue cars than there are red cars.	“Most of the cars are blue.”

Table 1: Example items in Experiment 3, manipulating the form of information relevant to verifying *most*.

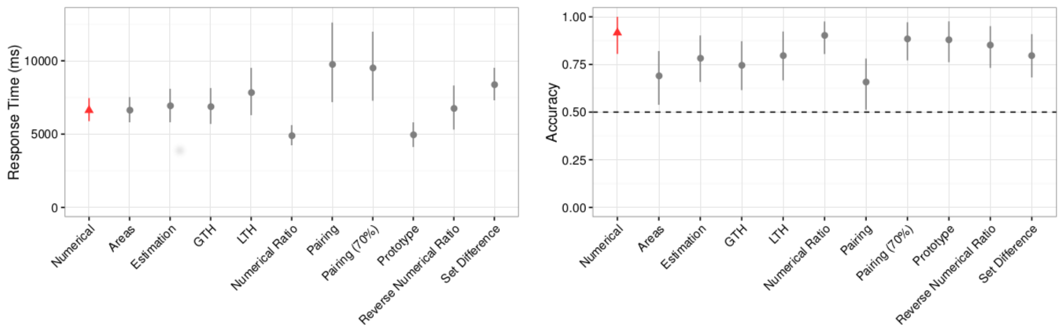

Figure 10: Mean response time and accuracy for each contextual cue to the meaning of *most*. Error bars show bootstrapped 95% confidence intervals

Participants Four hundred and thirty-eight participants (Area Cues: N = 39, Estimation: N = 41, Greater Than Half: N = 39, Less Than Half: N = 39, Numerical: N = 36, Numerical Ratio: N = 41, OnetoOnePairing: N = 41, OnetoOnePairing(70%): N = 35, Prototype: N = 42, Reverse Numerical Ratio: N = 41, Set Difference: N = 44) were recruited via Amazon Mechanical Turk and paid for their participation, which lasted approximately one minute. Twenty seven participants were excluded because they failed the same attention check as described in Experiment 1. Participants were not allowed to complete this experiment twice.

5.2 Results and Discussion

Figure 10 shows accuracy and response time to the verbal task. If there was a processing advantage for exact cardinalities, we would expect the *numerical cue* to have a higher accuracy and faster response times compared to the alternative cardinality-independent prompts (e.g. numerical, area, prototype). In fact, this value is numerically indistinguishable from most of the other alternatives (see Supplemental Tables). The results show that participants perform above chance of 50% for all contextual cues to the meaning of *most* (all $t > 10$, all $p < 0.001$) indicating that various senses of *most* can be accurately verified. The results show no significant differences for accuracy and response time between the numerical cue and other cues, with a few exceptions. Participants were less accurate than numerical cues for *area cue*($\beta = -0.224$, $t = -2.49$, $p < 0.05$), *one-to-one pairing cue* ($\beta = -0.258$, $t = 2.89$,$p < 0.01$), and *greater-than-half cue*($\beta = -0.17$, $t = -1.92$, $p = 0.055$). Participants differed in response time from numerical cues for both pairing conditions. This difference in time was likely due to re-reading the scenario and the prompt, which were lengthier than other cues[2].

Conclusion Overall, these results indicate that *most* can very naturally be computed even when the relevant cardinalities cannot even be explicitly determined. There are few apparent processing disadvantages to non-numerical *most*, even including cases where cardinalities cannot be computed. These results suggest that the psychological processes supporting *most* are unlikely to be numerical at their core.

6 General Discussion

Our results have evaluated several conceptual and experimental factors where semantic theory interfaces with psychological processes. Our work was motivated by key claims from prior literature on *most*, namely that the meaning of *most* relied critically on a notion of *cardinality* and that the semantics of most is closely related to its verification procedures.

With respect to the first claim, we find a general bias to use cardinality, but substantial variation such that some individuals use other cues in ambiguous contexts (Experiment 3) and participants are rapid and accurate at verifying *most* even when the relevant cardinalities cannot be determined (Experiment 4). These results generally show that semantic understanding may tend to be based in numerical ideas, but its execution is much more flexible and context-dependent than a single logical or semantic characterization would predict.

For the second claim, Experiments 1, 2, and 3 showed evidence for a context-sensitive mechanism for verifying *most*. In these experiments, we found that subjects switched procedures most obviously consistent with an optimization of their time. Importantly, though, different subjects appeared to use different strategies even in the same context (10 participants explicitly reported a switch in strategy in Experiments 2). This means that the linkage between the semantics of most and the verification procedure people deploy is not simple—instead it involves strategic choices that depend on the context and format in which cardinalities are presented. From the view of language processing, a tight coupling between verification and semantic representation would be surprising. In many cases, language processing can be seen to be highly adaptable and context-sensitive (e.g., Olson, 1970; Hanna & Tanenhaus, 2004), meaning that we should not expect semantic, compositional, or lexical processing measures to give direct access to underlying semantic representation, independent of pragmatic and contextual considerations. In what could turn out to be a historical analogy, early theories of parsing proposed that default strategies were driven by syntactic representation (e.g, Frazier, 1979). Later work, however, revealed that many information sources are immediately available to parsing systems, and behavioral responses are quite sensitive to these other factors. *Most* may be similar in that its processing patterns tell us more about pragmatics, inference, and context than its underlying logical form.

These results paint a picture where the most fruitful way to study such semantic language comprehension is to investigate the rich, context-sensitive interpretative ability people have. These results also provide an important lesson for behavioral studies of semantics or other domains: the apparent semantic processes in language understanding are difficult to separate from task demands. Moving from theoretical and abstract characterizations of linguistic meaning to empirically supported psychological processes is likely to require integration with many—perhaps most—areas of cognitive psychology.

[2] These pairing cues also read like math problems, which might have rustled some jimmies.

References

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219. doi: 10.1007/BF00350139

Cantlon, J. F., & Brannon, E. M. (2007). How much does number matter to a monkey (macaca mulatta)? *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(1), 32.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics.* OUP USA.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307 - 314. doi: http://dx.doi.org/10.1016/j.tics.2004.05.002

Frazier, L. (1979). *On comprehending sentences: syntactic parsing strategies.* Indiana University Linguistics Club.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, *17*(1), 63–98. doi: 10.1007/s11050-008-9039-x

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256. doi: 10.1007/s11050-010-9062-6

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., & Gureckis, T. (2012). *psiturk (version 2.1.2)[software]. new york, ny: New york university.*

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological review*, *77*(4), 257.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language*, *24*(5), 554–585. doi: 10.1111/j.1468-0017.2009.01374.x

7 Supplemental

7.1 S1. Experiment 4 Regression Tables

This section presents the regression tables from Experiment 4, for a more in-depth look at the comparisons made. Table 3 shows a linear regression modeling Response Time by Contextual Cue, compared to the baseline Numerical Cue (explicit cardinality). Table 2 shows logistic regression for Accuracy by Contextual Cue compared to the baseline Numerical Cue.

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
Numerical (Intercept)	6624.42	812.07	8.157	3.86e-15 $***$
Areas	-13.24	1126.14	-0.012	0.9906
Estimation	280.95	1112.88	0.252	0.8008
Greater Than Half	236.10	1126.14	0.210	0.8340
Less Than Half	1208.66	1126.14	1.073	0.2838
Numerical Ratio	-1713.81	1112.88	-1.540	0.1243
Pairing (One Plus)	3133.32	1112.88	2.816	0.0051 $**$
Pairing (70%)	2911.90	1156.62	2.518	0.0122 $*$
Prototype	-1701.44	1106.67	-1.537	0.1249
Reverse Numerical Ratio	118.95	1112.88	0.107	0.9149
Set Difference	1751.86	1095.00	1.600	0.1104

Table 2: Regression Table showing differences in Response Time by Contextual Cue with a dummy-coded baseline of Numerical (cardinality independent). P-values are not corrected for multiple comparisons.

Starting with response time data, we fit a linear regression model (see Table 2) and generally find no significant differences between the *numerical cue* and the other cues with two exceptions. Both of the *pairing cues* ("one plus" and "70%") have longer response times than those of *numerical cues.* We believe that this difference in time is due to re-reading the "scenario" and the "prompt", which were lengthier than other cues. Notably, the trend in our data is that participants using the *numerical ratio* and *prototype* contextual cues respond faster when verifying *most* than the participants using the other cardinality-independent contexts of use, perhaps indicating that these have a more natural verification than even numerical stimuli.

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
Numerical (Intercept)	0.91667	0.06499	14.104	< 2e-16 $***$
Areas	-0.22436	0.09013	-2.489	0.01318 $*$
Estimation	-0.13618	0.08907	-1.529	0.12703
Greater Than Half	-0.17308	0.09013	-1.920	0.05549 .
Less Than Half	-0.12179	0.09013	-1.351	0.17731
Numerical Ratio	-0.01423	0.08907	-0.160	0.87316
Pairing (One Plus)	-0.25813	0.08907	-2.898	0.00395 $**$
Pairing (70%)	-0.03095	0.09257	-0.334	0.73827
Prototype	-0.03571	0.08857	-0.403	0.68699
Reverse Numerical Ratio	-0.06301	0.08907	-0.707	0.47970
Set Difference	-0.12121	0.08764	-1.383	0.16736

Table 3: Regression Table showing differences in Accuracy by Contextual Cue with a dummy-coded baseline of Numerical (cardinality independent). P-values are not corrected for multiple comparisons.

Turning to accuracy, we find that participants perform above chance—i.e., 50%, for all contextual cues to the meaning of *most* (all $t > 10$, all $p < 0.001$) indicating that various senses of *most* can be accurately verified. Next, we fit a logistic regression model (see Table 3) and generally find no significant differences between the *numerical cue* and the other cues with three exceptions. First, participants are less accurate using the *area cues* than using the *numerical cues* ($\beta = -0.224$, $t = -2.49$, $p < 0.05$). This difference might be a result of the external knowledge required to verify these specific prompts (e.g., the prompt "Most of Colorado is water" requires knowledge of specific geography). Second, participants were less accurate using the *one-to-one pairing cue* than using the *numerical cue* ($\beta = -0.258$, $t = 2.89$,$p < 0.01$). This difference is likely driven by either the fact that these conditions did not have approximately 70% ratios (since they are by definition one more), or a pragmatic tendency to not consider one more appropriate for *most* (in contrast to, e.g. "about half"). This also supports the results of Experiment 1, where we find a reticence to accept pairing strategies when verifying *most.* Lastly, we find a trend that participants are less accurate using the *greater-than-half cue* ($\beta = -0.17$, $t = -1.92$, $p = 0.055$).