**Title**

Elicitation of Quantified Description Under Time Constraints

**Permalink**

https://escholarship.org/uc/item/15g3k76s

**Journal**

Proceedings of the Annual Meeting of the Cognitive Science Society, 41(0)

**Authors**

Briggs, Gordon
Wasylyshyn, Christina
Bello, Paul F.

**Publication Date**

2019

Peer reviewed

# Elicitation of Quantified Description Under Time Constraints

**Gordon Briggs (gordon.briggs@nrl.navy.mil)**
**Christina Wasylyshyn (christina.wasylyshyn@nrl.navy.mil)**
**Paul F. Bello (paul.bello@nrl.navy.mil)**
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
4555 Overlook Avenue SW
Washington, DC USA

## Abstract

Quantity can be expressed in a variety of ways and at different levels of precision. One factor that influences numerical description of elements in a visual scene is how long the scene is observed. We extend a previous incremental model of numerical perception to model quantified description under time constraints. Our extended model predicts that as presentation duration decreases and as the quantity of items to be enumerated increases, the frequency of inexact quantifiers will increase. We conducted two human subject elicitation studies to test these predictions. Our findings were consistent with our model's predictions. Additionally, we demonstrate that our novel model of incremental numerical perception and quantified description closely predicts the precise proportion of exact numerical responses generated by in these experiments.

**Keywords:** numerical language; numerical perception; quantifiers; subitizing; counting; estimation; computational model

## Introduction

Quantity can be expressed in a variety of ways and at different levels of precision. Speakers can use exact numbers to describe quantities (e.g., "there are *twenty-two* guests at the party") or they can use more vague language (e.g., "there are *many* guests at the party"). Many factors influence the form and degree of precision of quantified language a speaker uses, including pragmatic considerations (Cummins, 2015). For example, speakers can express quantities in strategically vague ways for the purpose of influencing behavior (Hesse & Benz, 2018).

In the context of visual scene description, another factor influences quantified language: how long the scene is observed. Research in numerical perception suggests that mental representation of visual quantity is incrementally acquired through temporally extended and attentionally-dependent perceptual processes (Trick & Pylyshyn, 1994; Railo, Koivisto, Revonsuo, & Hannula, 2008). Glancing at a scene allows one to form a less precise representation of quantity, while taking the time to count each relevant item gives one a precise numerical representation of quantity. Using exact numbers when insufficient time has been devoted to complete enumeration often results in incorrect numerical guesses, and psychologists studying numerical perception often rely on analysis of error patterns during exact number elicitation tasks to make inferences about underlying processes and representations (e.g., Mandler & Shebo, 1982).

However, outside psychophysics experiments, people are rarely forced to express themselves using only exact numerical expressions. Some recent work has begun to examine patterns of quantified language usage in more unconstrained situations. In particular, Barr, Deemter, and Fernández (2013) found that when individuals produce quantified reference expressions (QREs), the form of the QRE was dependent on the numerosity of the sets under consideration. When the quantities in each set were large, people tended to produce relational expressions (e.g., "my set is the largest one"). However, when the target set of objects consisted of a small quantity, people tended to produce QREs with exact numerical descriptors despite inexact QREs being sufficient to disambiguate the expression. This result suggests that people balance pragmatic concerns of informativity with minimization of perceptual effort or cost.

The contribution of this paper is two-fold. First, we extend a previous incremental model of numerical perception to model quantified description under time constraints. This provides an explicit model of perceptual cost that was lacking in prior literature. The second contribution of this paper is two novel human subject elicitation studies designed to test the predictions generated by this model. Our extended model predicts that as presentation duration decreases and as the quantity of items to be enumerated increases, the frequency of inexact quantifiers will increase. The findings from the experiments were consistent with our model's predictions. Additionally, we demonstrate that our novel model of incremental numerical perception and quantified description closely predicts the precise proportion of exact numerical responses generated by in these experiments.

## Computational Models of Numerical Perception

The perception of numerosity consists of multiple processes, each occurring at different rates and resulting in mental representations of varying precision. Explicit counting provides a slow, but precise, determination of number (Gelman & Gallistel, 1986) rooted in linguistic representation in a phonological buffer (Whalen, Gallistel, & Gelman, 1999). Estimation provides a rapid but less precise judgment of the quantity of a group of objects (Barth, Kanwisher, & Spelke, 2003) rooted in what has become known as the approximate number system (ANS) (Dehaene, 2011). Between these two procedures, a third process, called subitizing, provides both rapid and precise judgments of numerosity, but only for small quantities, from one to typically around four objects (Kaufman, Lord, Reese, & Volkmann, 1949). Consequently, the range of numerosities between one and four has become known as the subitizing range. While debate still continues about the representations and processes underlying subitizing, there are converging lines of evidence that suggest that the object-tracking system (OTS) plays a central role (Feigenson, Dehaene, & Spelke, 2004).

Recently, there has been renewed interest in developing neural models of numerical perception. These models typically focus on accounting for only a single process and form of representation, such as estimation and the ANS (Chen, Zhou, Fang, & McClelland, 2018) or counting and exact number (Fang, Zhou, Chen, & McClelland, 2018). Some researchers have begun to examine the generation of quantified descriptions of visual scenes with varying levels of precision (Pezzelle, Marelli, & Bernardi, 2017). However, these models also generally do not attempt to model the time course of enumeration. The psychophysical literature on numerical perception has shown that within the subitizing range, each additional object requires only 40–100 ms to accurately enumerate, while outside the subitizing range, each additional object requires 250–350 ms to enumerate (Trick & Pylyshyn, 1994). Most existing computational models of numerical perception do not attempt to capture this aspect of enumeration, nor do they provide accounts for how estimates can be refined with additional time.

## An Incremental Model

In contrast with these previous models, Briggs, Bridewell, and Bello (2017) developed a computational model, implemented in the ARCADIA cognitive system (Bridewell & Bello, 2016), that models temporally extended numerical perception and integrates various forms of numerical representation. The model contains components that implement three distinct numerical representation systems: the ANS, the OTS, and the phonological buffer. We will denote this model as `INP-Guess` (incremental numerical perception and guessing).

`INP-Guess` operates by first ascertaining an approximate, noisy estimate of quantity by deploying visual attention toward the entire group of items to be enumerated.[1] Subsequently, serial attention is deployed to each individual item in the group. This process of serial attention first fills up the visual short-term memory (vSTM) slots in ARCADIA's object-tracking system. If there are no more items to be enumerated or no more available slots in vSTM, then a lexical representation of the quantity of relevant items in vSTM is subvocalized within the system's phonological buffer. After this point, subsequent serial focus to new items is accompanied by subvocalization of the next count word in the counting sequence.

*Numerical Guessing.* If the visual scene ends, or enumeration otherwise ends, the model merges both the results from the ANS and the lexicalized count into a single numerosity judgment. If time allows for an explicit count to be fully generated (i.e., all items had received individual attentional focus), the explicit count is recorded. Otherwise, an educated guess is made:

$$Guess(n_c, n_e, w) = n_c + sample(\mathcal{N}(n_e - n_c, \sqrt{w \cdot (n_e - n_c)}))$$

where $w$ denotes the Weber fraction of the ANS, $n_e$ denotes the number of items that collectively received attentional focus during estimation, and $n_c$ denotes the number of items that received individual focus during subitizing and counting.

As more items are individually attended to, an exact representation of a lower bound on the number of visual items in the scene increases. The equation above reflects the variance of the noisy numerosity representation upon which linguistic description is based decreasing as this lower bound increases. Note, we are not proposing that serial deployment of attention directly affects the variance of the representation produced by the ANS. Rather, what we are proposing is that the partial exact, lexical representation of number and the noisy ANS representation are merged (Briggs, Bridewell, & Bello, 2017), such that the resulting merged representation of numerosity will have lower variance when more items have received serial focus of attention. If there is enough time to devote attention to each item individually, then $n_c = n_e$ and guess is equal to the lexicalized count $n_c$.

While the precise time $T_{attend}$ it takes to fully attend to $n$ items individually within the `INP-Guess` model depends on multiple task-related factors, we can formulate a mathematical approximation of the time required in a simple case (i.e., a single-task involving enumeration of all items in a visual scene):

$$T_{attend}(n) \approx T_f \cdot min(r_s, n) + \prod_{max(r_s,n) \leq i \leq n} T_{subvocal}(i) + T_f$$

where $T_f$ denotes the time necessary to attend to encode a single item into vSTM, $r_s$ denotes the subitizing limit, and $T_{subvocal}(i)$ denotes the time necessary to subvocalize the $i$-th count word. Based on the original parameter values used by Briggs and colleagues (2017), we set the following values: $T_f = 50$ms, $r_s = 4$. Subvocalization time $T_{subvocal}(i)$, varies by number and is based on the formula from Huss and Byrne (2003).

Therefore, the number of items $n_c$ that can be individually attended to in `INP-Guess` within a time window of $T$ can be approximated as:

$$n_c(T) \approx \operatorname*{argmax}_{i \geq 0} \{i | T_{attend}(i) \leq T - T_{estimate}\}$$

where $T_{estimate}$ denotes the time necessary for the initial estimation of quantity within the visual scene.

Briggs and colleagues (2017) demonstrated that the `INP-Guess` model could account for the bilinear reaction time

[1] We refer the reader to the original model paper for details about how visual attention is realized in the ARCADIA system.

curve in enumeration (Trick & Pylyshyn, 1994). Additionally, the `INP-Guess` model could account for the pattern of error in studies of subitizing during conditions of divided attention (Railo et al., 2008).

However, while being able to capture the pattern of error in numerical guessing during tasks with time and attentional constraints is desirable in a model of numerical perception, it is an incomplete account of quantified language use. When speaking with one another, people are faced with a variety of communicative norms. Communicating exact numbers in cases of noisy numerical representations would likely violate these norms, including prohibitions against communicating without adequate evidence (Grice's Maxim of Quality) and failure to be informative about one's own certainty (Grice's Maxim of Quantity) (Grice, 1975). The use of inexact quantified description (e.g., "there are between 4 to 7 items", "there are about 6 items", etc.) is one way to satisfy these communicate norms. While the wide range of quantified language provides ample opportunity for investigation by researchers in semantics and pragmatics (Cummins, 2015), an even more basic question arises: how can we model when people decide to use inexact quantified description?

## Extension to Inexact Language

To model when people use inexact quantified descriptions, we propose a simple extension to the `INP-Guess` model, which we will denote as `INP-Hedge`. Instead of sampling a single guess value, `INP-Hedge` obtains $\nu$ distinct guesses, which we will denote as the multiset $G = \{g_1, ..., g_\nu\}$. This corresponds to a collection of values an individual may find plausible. If all the guesses in $G$ are the same ($g_1 = g_2 = ... = g_\nu$), then we would predict an exact numerical description is generated equivalent to the value of these guesses. Otherwise, we would predict an inexact numerical description is generated, which can be derived from the set of guesses. For instance, consider a set of guesses $G = \{6, 8, 8\}$. Potential ways to linguistically describe this set are "about eight" or "six to eight."

How particular forms of inexact quantified description are generated is a question beyond the scope of this paper. Here, we do not attempt to model the distribution of specific forms of inexact description. In the `INP-Hedge` model we currently generate two forms of inexact expression: hedged numbers (e.g., "about eight") and intervals (e.g., "six to eight"). If the majority of the sampled guesses are equal to a value $X$, then `INP-Hedge` produces a hedged number expression anchored in this majority guess (i.e., "about $X$"). Otherwise, the model produces an interval response ("between $X$ and $Y$" ), where $X$ corresponds to the minimum guess and $Y$ corresponds to the maximum guess.

This quantified description mechanism is still preliminary, and we will discuss how the `INP-Hedge` model can direct future work on inexact quantifier realization in the general discussion below. Overall, `INP-Hedge` assumes that speakers would detect the uncertainty of their mental representation of quantity by considering multiple plausible exact number responses, and then elect to hedge their quantified descriptions. Thus, the `INP-Hedge` model predicts that the limits of human perceptual performance would influence language usage, because there may be insufficient time to completely eliminate uncertainty about quantity through serial attention. Specifically, the `INP-Hedge` model predicts that enumeration duration and numerosity have the following effects on quantified language:

(P1) In the subitizing range (quantity 1-4), participants will predominately use exact quantifiers.

(P2) For indefinitely long presentation durations, participants will use exact quantifiers.

(P3) As presentation duration decreases, the frequency of inexact quantifiers will increase.

If people elect to use inexact quantified language to avoid being incorrect, as we hypothesized above, we can propose one additional prediction:

(P4) In difficult duration/quantity pairings, participants that responded by describing quantity using inexact quantifiers will indicate higher confidence in the correctness of their response vs. participants that responded in the same duration/quantity condition with exact numbers.

# Experiment 1

To test our model's predictions, we conducted an online numerical perception and quantified language elicitation experiment. Participants viewed videos in which varying quantities of black dots were presented for varying durations.

## Method

*Participants.* Thirty-nine participants (mean age = 36.3; 19 females, 19 males, and 1 other) volunteered through the Amazon Mechanical Turk online platform (Paolacci, Chandler, & Ipeirotis, 2010). All but one participant reported being native English speakers.

*Design, procedure, and materials.* We manipulated the duration of stimulus presentation of dot clusters and the quantity of elements (dots) presented. Three possible presentation durations were used: 200 ms, 1000 ms, and an indefinite amount of time (dots remained on the screen while participants responded). Three possible quantity ranges were used: $[1-4]$, $[5-8]$, and $[9-12]$. The dot clusters in each video were randomly arranged, and four videos were produced for each specific numerosity, yielding 16 unique videos for each stimulus duration and numerosity condition (4 videos per number with 4 possible numbers per quantity range). Participants were presented with one video from each of these duration/quantity conditions in a random order, viewing nine videos in total. Videos were 512x512 pixels in dimension with a light grey

| | Stimulus Duration | | |
|---|---|---|---|
| Num. Range | 0.2s | 1.0s | $\infty$ s |
| $[1-4]$ | 94.9% | 97.4% | 100.% |
| $[5-8]$ | 61.5% | 87.2% | 100.% |
| $[9-12]$ | 46.2% | 53.8% | 100.% |

Table 1: Percentage of responses categorized as EXACT-NUM by stimulus duration and numerosity range conditions in Experiment 1.

background. A dark grey fixation cross appeared for one second before the cluster of dots. A masker grid was displayed following the stimulus interval (except in the indefinite enumeration time condition). After a video had concluded, participants were asked to complete the following sentence, being as accurate and precise as possible:

"In the above video, there are _______ black dot(s)."

Additionally, participants were asked to report their confidence in their completed description (1 = very unsure to 5 = very confident). Because we were primarily interested in investigating the precision of quantified description of a visual scene, we used a free-response, sentence-completion task instead of a completely free-response task. This was done to encourage quantified description and avoid descriptions of dot clusters based on other attributes, such as spatial arrangement (e.g., "I see a group of dots shaped like a constellation of stars").

## Results

*Analysis.* The expressions used to complete the description were categorized into five types: (1) exact numbers, which we will denote as EXACT-NUM (e.g., "In the video above, there are *four* black dot(s)"); (2) hedged numbers, denoted as HEDGED-NUM (e.g., "In the video above, there are *about ten* black dot(s)"); (3) intervals, denoted as INTERVAL (e.g., "In the video above, there are *five to seven* black dot(s)"); (4) vague quantifiers, denoted as VAGUE-Q (e.g., "In the video above, there are *several* black dots(s)"); and (5) other miscellaneous expression, denoted as OTHER (e.g., "In the video above, there are *groups of* black dot(s)"). Two annotators classified each response. High inter-annotator agreement was found (Cohen's $\kappa$ = .945). The proportions of exact numerical responses (EXACT-NUM) for each duration and numerosity condition are reported in Table 1. All four predictions were supported by the data.

Consistent with P1, 114 of 117 (97.4%) of responses in conditions within the subitizing range were EXACT-NUM responses. Additionally, consistent with P2, only exact descriptions were used when participants had an unlimited amount of time to enumerate. Consistent with P3, the number of exact responses decreased from 117 out of 117 in the indefinite duration condition to 93 out of 117 with one second of duration, yielding a significant difference (Fischer exact test, $p < .001$). The number of exact responses further decreased from 93 out of 117 to 79 out of 117 in the 200ms duration condition, though this difference was only marginally significant (Fischer exact test, $p = .054$).

Finally, Wilcoxon-signed rank tests indicated confidence ratings were significantly lower for exact responses than inexact responses for quantity ranges $[5-8]$ ($p = .039$) and $[9-12]$ ($p < .001$) at 200 ms presentation time and quantity range $[9-12]$ ($p = .007$) at 1000 ms presentation time, supporting our final prediction (P4).

Intriguingly, only 23 out of the 39 participants (59%) used inexact quantifiers. The other 16 out of 39 (41%) only used exact number responses, guessing in cases of uncertainty. As previously discussed, we would expect a participant who uses inexact quantified descriptions to generate these inexact descriptions to express uncertainty and avoid incorrectness when exact enumeration is difficult. Therefore, we would predict the EXACT-NUM responses from participants who used only EXACT-NUM responses to be less accurate than those generated by participants who switch between exact and inexact expressions. The data supported this prediction. The accuracy of EXACT-NUM responses from participants who used only exact expressions (69.4%) was found to be lower than those from participants who sometimes used inexact expressions (91.8%), yielding a significant difference (Mann-Whitney U test, $p < .001$).

## Model Fit

While our four main predictions were supported by the data, we also sought to test how well the INP-Hedge model predicts the precise frequency of EXACT-NUM responses in each stimulus duration and numerosity condition. We ran the INP-Hedge model[2] ten times on each video from Experiment 1. The proportion of exact quantifier responses produced by INP-Hedge, compared with the human data from Experiment 1, is found in Figure 1. The results of the human data were highly correlated with the proportion of exact/inexact quantifiers selected by the INP-Hedge model (Spearman's $\rho$ = 0.938). Though the correlational fit is high, we can see that INP-Hedge underestimates the amount of EXACT-NUM responses in the hardest duration and numerosity conditions. Specifically, INP-Hedge underestimates exact responses in numerosity ranges $[5-8]$ and $[9-12]$ by 43% and 28%, respectively. Exact responses during medium durations (1000 ms) for larger numerosities ($[9-12]$) are also underestimated by 33%.

What could explain this underestimation? Recall that a sizable portion (41%) of participants only gave EXACT-NUM descriptions. That is to say, about 4 out of every 10 participants guessed an exact number response when they were uncertain about the quantity of dots. To account for this, we revisited the Experiment 1 videos, running the INP-Guess model four times on each video and the INP-Hedge model six times (replicating the mixed set of strategies found in human participants). Not only was the correlational fit of this

[2] Setting $\nu = 3$.

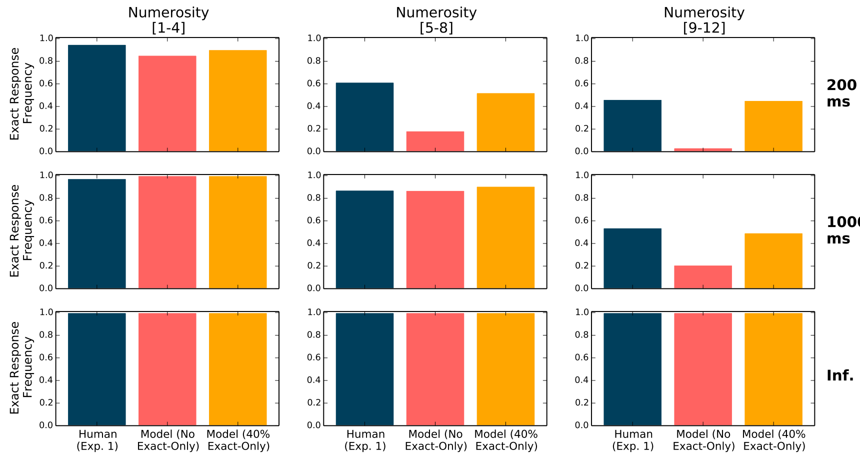


Figure 1: Predicted proportions of EXACT-NUM responses based on our model (pink) and our model adjusted for the number of participants that give only exact-number responses (yellow), compared with proportion of EXACT-NUM responses from human data in Experiment 1 (blue).

mixed-model improved (Spearman's ρ=.956), but the precise predictions about exact response frequency are much closer to the observed frequencies from the human data. Proportion differences are reduced during short stimulus durations (200ms) to 9% and 1%, for numerosity ranges $[5-8]$ and $[9-12]$, respectively. Finally, the difference in exact response proportion is reduced to approximately 4% for larger quantities ($[9-12]$) during medium durations (1000ms).

## Experiment 2

Roughly 40% of participants in Experiment 1 only gave exact numerical responses. This is in line with other studies, where a subset of participants use only exact numerical expressions for all items. For instance, in a QRE elicitation task, about 20% of participants used only exact numerical expressions (Barr, Deemter, & Fernández, 2013). However, unlike in Barr and colleagues (2013), participants in Experiment 1 did not have unlimited amounts of time to view stimuli. Therefore, it seems likely that participants are limiting their set of potential quantified response forms *a priori*. One possible explanation is that because each trial in Experiment 1 involved a question asking the participant to rate the confidence of their numerical expression, participants may have felt less pressure to hedge uncertainty about the observed quantity in the language of the numerical expression itself. Rather, participants may have been more inclined to guess an exact number, then hedge their uncertainty in the confidence question. In Experiment 2, we sought to eliminate this possibility.

### Method

*Participants.* Forty participants (mean age = 35.4; 19 females and 21 males) volunteered through the Amazon Mechanical Turk online platform. All participants reported being native English speakers.

*Design, procedure, and materials.* The experimental design, procedures, and materials were identical to Experiment 1, except in two respects. First, the confidence question was eliminated. Second, the number of trials each participant completed was increased to 18 (two trials per numerosity range and stimulus duration condition). Videos were randomly sampled from each stimulus duration and numerosity category without replacement.

### Results

As with Experiment 1, two annotators used the same labels to categorize all the expressions participants produced. Inter-annotator agreement was again high (Cohen's $\kappa = .938$). Table 2 lists not only the proportion of exact quantified descriptions, but the precise counts of each type of expression found. We found that 14 out of 40 participants (35%) in Experiment 2 used only EXACT-NUM expressions, compared with the 16 out of 39 (41%) in Experiment 1. While the proportion of participants using only exact expressions slightly decreased in Experiment 2, this difference is not statistically significant (Fischer's exact test, $p = .647$) given the number of participants in each study. Consistent with Experiment 1, we also found that EXACT-NUM responses from participants who used only exact descriptions were less accurate (61.9%) than those from participants who sometimes used inexact expressions (88.9%), yielding a significant difference (Mann-Whitney U test, $p < .001$).

*Predictions.* Aside from P4, which could not be tested as there was no confidence data in this experiment, the main predictions were also still supported. Consistent with P1, 233 out of 240 (97.1%) responses in the subitizing range were exact. Also, 235 out of 240 (97.9%) responses in the

| Response Type | [1 − 4] 0.2s | 1.0s | ∞ | [5 − 8] 0.2s | 1.0s | ∞ | [9 − 12] 0.2s | 1.0s | ∞ |
|---|---|---|---|---|---|---|---|---|---|
| EXACT-NUM | 75 | 79 | 79 | 45 | 67 | 80 | 35 | 36 | 76 |
| HEDGED-NUM | 2 | 0 | 0 | 16 | 8 | 0 | 13 | 15 | 1 |
| INTERVAL | 1 | 0 | 0 | 9 | 4 | 0 | 14 | 13 | 0 |
| VAGUE-Q | 2 | 1 | 1 | 10 | 0 | 0 | 17 | 14 | 3 |
| OTHER | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 0 |
| Exp. 2: Exact % | 93.8 | 98.8 | 98.8 | 56.3 | 83.8 | 100. | 43.8 | 45.0 | 95.0 |
| Model (INP-Hedge): Exact % | 85.3 | 100. | 100. | 18.4 | 86.9 | 100. | 3.4 | 20.9 | 100. |
| Model (40% INP-Guess): Exact % | 90.3 | 100. | 100. | 52.2 | 90.6 | 100. | 45.3 | 49.4 | 100. |

Table 2: Counts of response types by stimulus duration and numerosity conditions for Experiment 2.

indefinite duration condition were exact, consistent with P2. Consistent with P3, the number of exact responses decreased from 235 out of 240 in the indefinite duration condition to 182 out of 240 with one second of duration, yielding a significant difference (Fischer exact test, $p < .001$). The number of exact responses further decreased from 182 out of 240 to 155 out of 240 in the 200ms duration condition, yielding a significant difference (Fischer exact test, $p = .009$).

*Model Fit.* The pattern of exact/inexact response was similar to Experiment 1. Correlation of the proportion of exact is high (Spearman's $\rho$ = 0.926). Underestimation of usage of exact numerical expressions remains in the post-subitizing range for short stimulus durations (200ms) and larger numerosity ranges. Specifically, INP-Hedge underestimates exact responses in numerosity ranges $[5-8]$ and $[9-12]$ by 38% and 40%, respectively. Exact responses during medium durations (1000 ms) for larger numerosities ($[9-12]$) are also underestimated by 24%. Our mixed model, (40% INP-Guess, 60% INP-Hedge) increases model correlation (Spearman's $\rho$ = 0.944), while reducing this observed underestimation. Proportion differences are reduced during short stimulus durations (200ms) to 4% and 3%, for numerosity ranges $[5-8]$ and $[9-12]$, respectively. Finally, the difference in exact response proportion is reduced to approximately 4% for larger quantities ($[9-12]$) during medium durations (1000ms).

## General Discussion

The results of our two quantified language elicitation experiments demonstrate that the use of precise quantified language to describe visual scenes decreases with decreased viewing time or increased stimulus quantity. Future computational models of quantified description of visual scenes, regardless of how they are implemented, need to account for this phenomenon to fully capture human quantified language use. We contend that to make sense of these results, one must view numerical perception as a temporally extended process in which uncertainty is reduced by additional perception of the visual scene. Both computational models we presented above, INP-Guess and INP-Hedge, account for this reduction of uncertainty by a proposed model of serial deployment of attention to individual items in the visual scene.

While we have demonstrated that a combination of these psychologically grounded and attention-driven models of numerical perception and quantifier use is able to closely fit human patterns of quantified language use under time constraints, many open questions still remain. One question raised by our quantifier elicitation experiments (and results from Barr and colleagues, 2013) is how do people decide to constrain the set of quantifiers they elect to even consider generating? Because our elicitation experiment contained relatively small quantities of visual items (1-12), participants may have felt that the degree of potential error in exact number guessing to be acceptable. With this explanation, increasing the number of potential visual items (e.g., 50-120) may reduce the proportion of participants giving only exact responses. Likewise, task motivation and context would potentially affect quantifier use. Situations were precision is critical and error may lead to highly negative consequences are likely to elicit more exact quantified language.

This work raises another series of questions regarding the realization of inexact quantified language. If people do consider multiple forms of inexact quantified language, how do people choose the precise language to use in a particular context? The mechanism for selecting different forms of quantified description in our proposed model is still rudimentary. However, it does begin to make some predictions. For instance, the current model would predict that the bounds of interval expressions would increase proportionally with reduced enumeration time or increased numerosity. In future work, we hope to address a variety of these open questions and hypotheses.

## Acknowledgments

We would like to thank Hillary Harner, Andrew Lovett, Will Bridewell, Derek Brock, Brian McClimens and Sangeet Khemlani for their feedback and discussions pertaining to this project. We would also like to thank Kevin Zish, Kalyan Gupta, and the Knexus Research Corporation for their assistance in supporting these studies. Additionally, we would like to thank the reviewers for helpful feedback in improving the presentation of this work. This work was supported by a NRC Research Associateship award to GB, and AFOSR MIPR grant F4FGA07074G001 and ONR grant N0001416WX00762, both awarded to PB. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

## References

Barr, D., Deemter, K., & Fernández, R. (2013). Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of the 14th European Workshop on Natural Language Generation* (pp. 157–161).

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*(3), 201–221.

Bridewell, W., & Bello, P. F. (2016). A Theory of Attention for Cognitive Systems. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems* (pp. 1–16). Evanston, USA.

Briggs, G., Bridewell, W., & Bello, P. F. (2017). A computational model of the role of attention in subitizing and enumeration. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1672–1677). London, UK.

Chen, S. Y., Zhou, Z., Fang, M., & McClelland, J. L. (2018). Can Generic Neural Networks Estimate Numerosity Like Humans? In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 202–207). Madison, WI.

Cummins, C. (2015). *Constraints on Numerical Expressions* (Vol. 5). Oxford University Press.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. New York, NY: Oxford University Press.

Fang, M., Zhou, Z., Chen, S. Y., & McClelland, J. L. (2018). Can a Recurrent Neural Network Learn to Count Things? In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 360–365). Madison, WI.

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

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Hesse, C., & Benz, A. (2018). Giving the wrong impression: Strategic use of comparatively modified numerals in a question answering system. In *Proceedings of The Conference on Natural Language Processing (KONVENS)* (pp. 148–157). Vienna, Austria.

Huss, D., & Byrne, M. (2003). An ACT-R/PM model of the articulatory loop. In *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 135–140).

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*, 498–525.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*, 1–22.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Pezzelle, S., Marelli, M., & Bernardi, R. (2017). Be precise or fuzzy: Learning the meaning of cardinals and quantifiers from vision. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Vol. 2, pp. 337–342).

Railo, H., Koivisto, M., Revonsuo, A., & Hannula, M. M. (2008). The role of attention in subitizing. *Cognition*, *107*, 82–104.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.