Making Sense of Number, Bit-by-Bit

by

Samuel J. Cheyette

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Steven T. Piantadosi, Chair
Professor Alison Gopnik
Professor David Whitney

Fall 2021

Making Sense of Number, Bit-by-Bit

Abstract

Making Sense of Number, Bit-by-Bit

by

Samuel J. Cheyette

Doctor of Philosophy in Psychology

University of California, Berkeley

Assistant Professor Steven T. Piantadosi, Chair

Numerosity perception has been studied for at least 150 years and its psychophysics have been well characterized by experimental work. However, the origins of many of its key properties remain obscure. For instance, people estimate the numerosity of small sets (up to four) much more rapidly and accurately than larger sets; people tend to underestimate larger numerosities; estimation precision and accuracy increase with exposure duration. Standard models of numerical estimation do not account for these wide ranging phenomena, with large number estimation typically characterized as a draw from $Gaussian(n, w \cdot n)$, where w is a person's "Weber fraction," and exact small number perception characterized separately, the result of an independent object-file system. Furthermore, the inherently *perceptual* nature of estimation is largely ignored in many accounts of individual differences, which are often considered evidence of disparities in innate mathematical cognition. In my dissertation, I present studies of human behavior and computational models aimed at clarifying the visual mechanisms underlying numerical estimation. Our findings help to understand, and unify, key properties of number psychophysics which have previously been explained in terms of independent mechanisms or with ad hoc modifications to existing theories. For instance, we show how the psychophysics of both small and large number estimation can be unified into a single framework with a common mechanistic origin, and in fact how myriad properties of both (including estimation precision, bias, effects of time)

can be understood as downstream consequences of bounded-optimal perceptual inference.

Contents

Acknowledgments

I owe many people thanks for helping make this thesis possible, through advice, support, and guidance. First and foremost is my advisor, Steve Piantadosi. He has allowed me the freedom to explore many questions, while guiding me toward those that can be meaningfully answered by experiments or formal models. More importantly, he has shown excitement about my ideas and has always taken them seriously, even if they have not always particularly well-formed (in fact, they usually have not). This has been a gift, allowing me to grow from a self-conscious young graduate student into a scientist with the confidence to his own ideas seriously.

I would also like to thank both David Whitney and Alison Gopnik, the other members of my thesis committee, who have been incredibly supportive of me and curious about my work. It has been a privilege to be able to pick their brains about many different topics, some of which were related to my thesis work but just as many not. I have really, truly, enjoyed every conversation about science and philosophy I've had with these two brilliant and kind people.

Finally, I would like to thank my dear friends Fred Callaway and Willa Voorhies, who have been my steadiest source of intellectual and emotional support during graduate school. If not for them, I would most likely be a husk of a shell of a man, cowering in a cave somewhere far away.

1 Introduction

1.1 Background

The human visual system rapidly computes and represents myriad summary statistics about groups of objects (Whitney & Yamanashi Leib, 2018). The work presented in this thesis is a humble attempt to understand the cognitive mechanisms underlying just one such summary statistic: the numerosity of a set. Our visual system can quickly determine the number of objects in a scene without the aid of serial counting, though only approximately, a capacity that emerges early in infancy (e.g. Dehaene, 2011; Hyde et al., 2010; McCrink & Wynn, 2007; Wynn, 1992a; Xu & Spelke, 2000) and is shared with evolutionary ancestors as distant as cephalopods (Yang & Chiao, 2016). The emergence of measurable numerical discrimination ability so early in development, along with the capacity to do approximate arithmetic (Wynn, 1992a), suggest that being able to assess and reason about quantities is functionally quite important to many facets of life — an idea with obvious intuitive appeal to the modern numerate person. The ubiquity of innate numerical abilities across the animal kingdom aligns with this intuition as well: most animals will want to move toward the most abundant source of food, whatever that food may be; and, conversely, most animals will want to flee from the more numerous pack of predators, whomever those predators are.

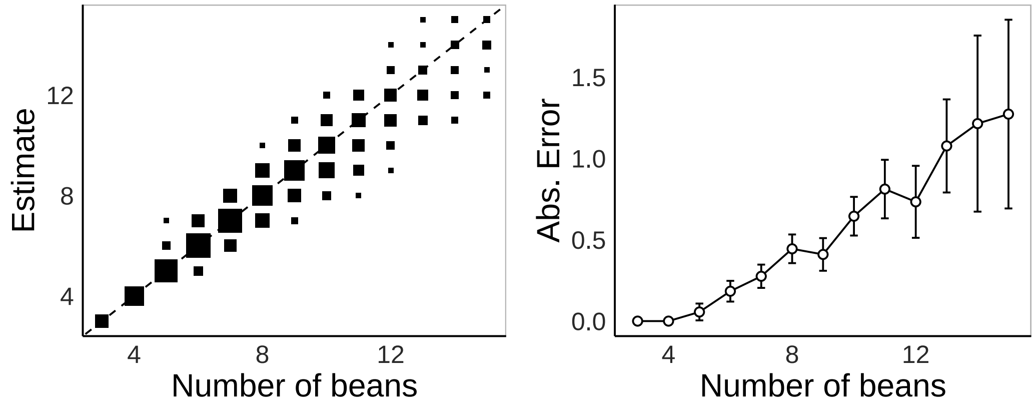
The first estimation experiment

Scientific investigation into human quantity estimation ability was first motivated as an attempt to understand the limits of the human mind in representing multiple objects simultaneously — a subject of interest to philosophers for some time. Citing the musings of various 18th century scholars in his Lectures on Metaphysics and Logic (Vol. 1), Sir William Hamilton (1859) wrote,

> Supposing that the mind is not limited to the simultaneous consideration of a single object, a question arises: how many objects can it embrace at once? [...] I find this problem stated and differently answered by different philosophers, and apparently without a knowledge of each other. By Charles Bonnet, the mind is allowed to have a distinct notion of six objects at once; by Abraham Tucker the number is limited to four; while Destutt Tracy again amplifies it to six. The opinion of the first and last of these philosophers appears to me correct. You can easily make the experiment for yourselves, but you must must beware of grouping the objects into classes. If you throw a handful of marbles on the floor, you will find it difficult to view at once more than six, or seven at most, without confusion.

It is not entirely clear what specifically Hamilton proposed to measure here or if Hamilton actually conducted any experiment at all — if he did, he did not provide details or data. That task was instead taken up a decade later by the logician and economist William Stanley Jevons (1871) who, citing this passage as motivation, remarked that "the subject seemed to me worthy of more systematic investigation, and it is one of the very few points which can, as far as we yet see, be submitted to experiment." In addition to its significance as the first scientific investigation into numerosity perception, Jevons' paper is remarkable for its clarity of writing, insight, and anticipation of cognitive psychology in using rigorous methods to test latent properties of the human mind.

In his experiment, Jevons set a small paper box on a tray and repeatedly tossed black beans in its direction. As soon as the beans came to rest, he immediately estimated how many had landed in the box and then recorded his estimate along with the actual quantity. The procedure was repeated 1,027 times. As he wrote, “the whole value of the experiment turns upon the rapidity of the estimation, for if we can really count five or six by a single mental act, we ought to be able to do it unerringly at the first momentary glance.” Though he only tested himself, and though there are some dubious aspects of Jevons’ methodology like that he himself threw the beans rather than another person, his essential findings have stood the test of time and have since been replicated many dozens of times.



(a)

(b)

Figure 1.1: Data from Jevons (1871), courtesy of the R HistData package (Friendly, 2021). Panel (a) shows the distribution of estimates (y-axis) for each actual quantity of beans thrown (x-axis), with the size of the square proportional to how many times that number-estimate pair occurred. Panel (b) shows the absolute error of his estimates as a function of quantity (i.e. $\mathbb{E}_k\left[|n-k|\right]$ for quantities n and estimates k).

Figure 1.1 provides data from Jevons’ experiment, re-plotted with tools unavailable in the 1800s (R Core Team, 2020). The full distribution of estimates as a function of the true number of beans thrown is presented in 1.1a, with the absolute error of estimates given in 1.1b. Jevons drew two main conclusions from these results: the answer to “how many objects can the mind embrace at once” seems to be four — at least, for him — since he was completely unerring in his estimation only when there were three or four beans but not beyond; beyond four, the proportion of inaccurate trials, as well as the magnitude of (absolute) error, grew in proportion with the true number of beans (shown in 1.1b).

Jevons fit a curve to determine how the magnitude of his errors grew with quantity. Specifically, for a number of beans n, he found that the expected absolute difference of his estimates k was well-described by the function,

$$\mathop{\mathbb{E}}_{k}\left[|n-k|\right] \approx \begin{cases} 0, & \text{if } n < 4.5 \\ 0.12 \cdot (n - 4.5), & \text{otherwise.} \end{cases} \tag{1.1}$$

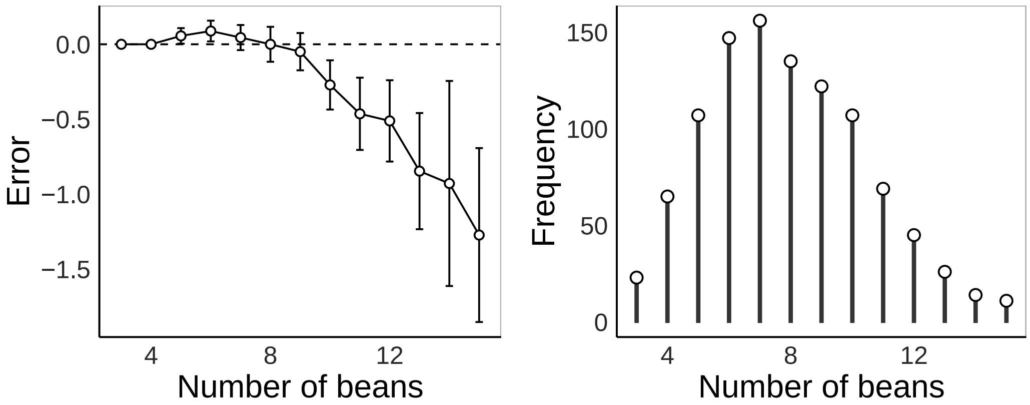
Jevons interpreted this function as saying that between four and five objects (specifically 4.5) can be discerned at once, after which the magnitude of error increases linearly. He wrote,

> This is a purely empirical law, the meaning or value of which I cannot undertake to explain. The most curious point is that it seems to confirm that my own power of estimating the number *five* is not perfect. The limit of complete accuracy, if there were one, would be neither 4 nor 5 but half-way between them.

In fact, this result can be understood naturally as an informational capacity limit, as Chapters 3 and 4 will explain. Jevons, of course, was writing without the benefit of information theory or psychological theory regarding capacity limits, and so his insightful remark that his limit of complete accuracy in numerosity perception was “between 4 and 5” could not be expressed in those terms.

Jevons made one other observation worth noting here, which is that he tended to slightly overestimate smaller quantities and significantly underestimate larger quantities. This can be seen in Figure 1.2a, which shows the average (signed) error as a function of quantity. He made a connection between this tendency and the frequency distribution of beans actually thrown into the bucket (shown in Figure 1.2b), commenting:

> There is a clear tendency to over-estimate small numbers and to under-estimate large numbers. There is an evident inclination toward those medium numbers which most frequently recurred: how far this discredits the experiments I cannot undertake to say, but it is an instance of that inevitable bias in mental experiments against which it is impossible to take complete precautions.



(a) (b)

Figure 1.2: Data from Jevons (1871). Panel (a) shows the average (signed) error as a function of the number of beans. Panel (b) shows the distribution of the true number of beans thrown.

This is an astonishingly prescient remark, and may be the first clearly interpreted scientific evidence that our minds rationally adapt to the statistics of the environment, even absent any conscious thought or deliberate intention. Though he intended it as a potential limitation of his experiment, an extension of this basic idea can help explain why estimation error is zero for small quantities and grows linearly beyond — provided in Chapters 3 and 4.

Number psychophysics

Nearly 80 years after Jevons' paper, Kaufman et al. (1949) delineated three distinct modes of enumeration: counting, subitizing, and estimating. Counting, of course, is a learned, serial procedure for exactly determining how many objects are in a set of arbitary size. Subitizing and estimation, on the other hand, are distinguished from counting as innate processes of enumeration that can operate in parallel over a visual field. Subitizing, a term they coined from the Latin word *subitare* meaning "to arrive suddenly," is the fast and accurate mode of apprehending small quantities. They considered the quantity of sets with six or fewer members to be "subitized" rather than "estimated." Estimation, on the other hand, is a less accurate and somewhat slower mode of determining the numerosity of larger sets. They speculated, based on data regarding participants' reaction times, accuracy, and confidence in quantity estimation, that there is some mechanistic distinction between subitizing and estimation. They wrote,

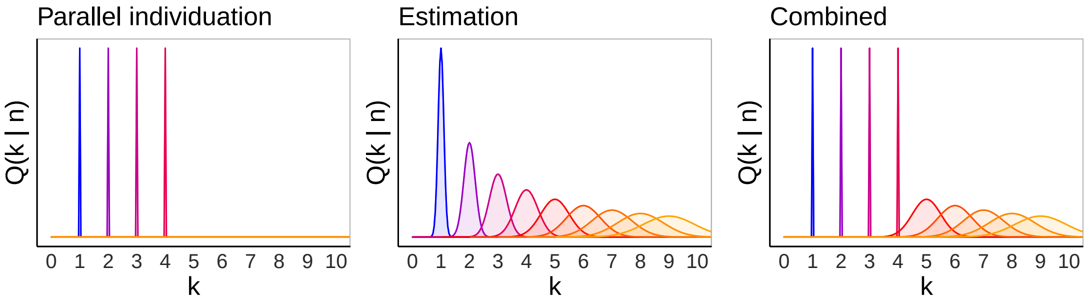
> The two terms differ in meaning, because to produce the process of *estimating* we present more than 6 dots; to produce *subitizing* we present 6 or less. This difference is surely an identifiable difference in operations. It might be a trivial difference, but the results tell us that it is not. If no discontinuities had appeared in the results, no distinction between subitizing and estimating could have been drawn.

Though the subitizing range is now considered to be four rather than six[1] — which really should have been the limit Kaufman et al. (1949) chose based on their data and that of Jevons (1871)[2] — the idea that there are two operational modes of determining a set's quantity (other than counting) is widely accepted. Furthermore, Kaufman et al.'s suggestion that these modes reflect different underlying cognitive mechanisms has gained widespread support as well. On one prominent account, we have two innate systems that allow us to represent numerical information (Dehaene, 1997; Feigenson et al., 2004; Trick & Pylyshyn, 1994). The first is the "parallel individuation" system, which allows us to attend to and track up to four objects. This slot-like tracking mechanism is what allows for rapid, exact enumeration of small quantities. The second is the "approximate number system," which is a noisy, analog system for representing numerosity in sets when their size exceeds the limits of the parallel individuation system.

The psychophysics of estimation resulting from the parallel individuation system and the approximate number system are illustrated in Figure 1.3. Each line represents the probability density $Q(k \mid n)$ over estimates (k) given a number of objects presented (n). Given $n = 1...4$ objects, the parallel individuation system exactly tracks them and thus estimation will be perfectly accurate, illustrated by the delta function probability density curves in Figure 1.3a. However, it is unable

[1] It is interesting that the debate has always been whether four or six is the limit, never five.

[2] Curiously, Kaufman et al. (1949) did not cite Jevons (1871).



(a) (b) (c)

Figure 1.3: Distributions $Q(k|n)$ of responses (k) given a number of objects presented (n) under three models. Probabilities (y-axis) of estimates (x-axis) are shown for numerosities 1-9 (colors). Panel (a) shows the form of a precise estimation system, panel (b) shows the form of a scale variable estimation system, and panel (c) shows them combined.

to represent sets beyond $n = 4$. The approximate number system (exemplified by 1.3b), on the other hand, is analog, continuous, and unbounded but represents each subsequent numerosity with decreasing precision. A popular model of the approximate number system assumes that estimates are drawn from,

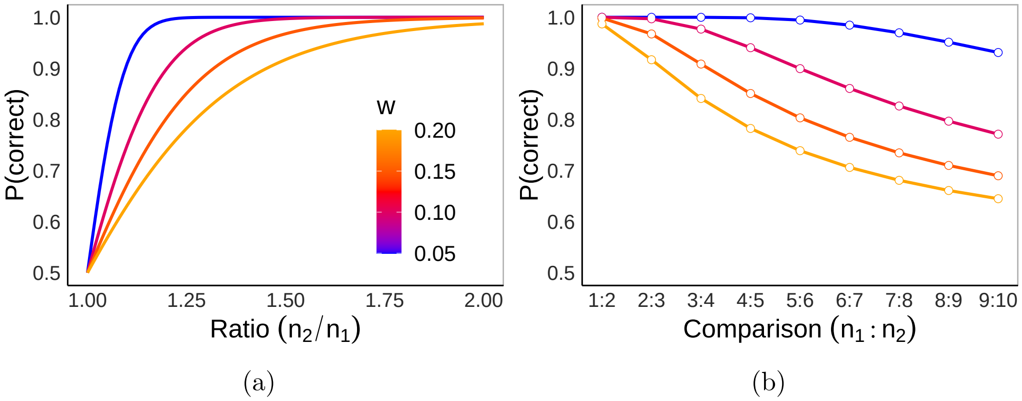
$$k \sim \mathcal{N}\left(n, w \cdot n\right). \tag{1.2}$$

This model of large number estimation has the standard deviation of estimates increasing at a rate w per object shown. This constant, w, is called a person's "Weber fraction."

Is subitizing consistent with Weber's law?

Weber's law, from which the term "Weber fraction" derives, states that the discriminability of two sensory stimuli is governed by their ratio (Weber, 1834), and it has been found to approximately hold for many sensory stimuli, such as brightness, loudness, and length, among others (e.g. M. Treisman, 1964). A consequence of Weber's law is that as the magnitude of a stimulus s increases, it becomes increasingly difficult to distinguish it from the magnitude $s + \delta$, since the ratio $s/(s + \delta)$ approaches 1 as s increases. Instead, a fractional change in s,

$w \cdot s$, results in constant discriminability, since the ratio $s/(w \cdot s)$ is constant. For this reason, (1.2) is a model of numerosity perception consistent with Weber's law. Figure 1.4a illustrates the probability that a person would choose be able to determine that a stimulus n_2 is greater than another stimulus n_1 as a function of the ratio n_2/n_1 and that person's Weber fraction.



(a)

(b)

Figure 1.4: Weber's law states that the ability to discriminate between two sensations depends on their *ratio* rather than the *difference* between them. The *Weber fraction* is a person- and sensation-specific constant, w, which governs how easily discriminable magnitudes are at a given ratio. Panel (a) shows the probability that a person can accurately discriminate two magnitudes as a function of their ratio, given a Weber fraction w (each line); panel (b) gives the probability a person could reliably distinguish between quantities n and $n + 1$ for $n = 1 \ldots 10$.

It is widely accepted that discrimination of large quantities follows Weber's law[3] — though as Testolin and McClelland (2021) point out, this is only approximately true. However, one difficulty in empirically determining whether small number discrimination is consistent with Weber's law is peculiar to the domain of number: because the cardinality of a set is a discrete quantity, there is no way to fully dissociate *ratios* from *magnitudes*. This is a particular problem for small quantities, since the a ratio of 2:1 is the smallest that one can test against sets of cardinality 1. For this reason, the question of whether people's near-exact estima-

[3] This does not hold for very large quantities, which seem to be perceived in terms of their texture density (Anobile et al., 2016; Ross & Burr, 2012). In fact, very large quantities follow different psychophysics, consistent with square-root scaling of internal noise rather than a linear scaling.

tion of small quantities is consistent with Weber's law was not definitively settled until recently. Figure 1.4b illustrates the difficulty: even if Weber's law holds in the small-number regime, small quantities can be distinguished with near-exact precision given low enough w.

However, Revkin et al. (2008) performed a particularly well-controlled study definitively showing that the psychophysics of small-quantity estimation differs from that of large-quantity estimation. They ran a quantity estimation task with two conditions. In the first condition, only 1-8 items were shown on each trial; in the second condition, 10-80 items were shown, but only the deciles (10, 20, 30...). In both conditions, participants were informed of the possible quantities involved in advance, and could only respond by pressing 1-8 on their keyboard in both conditions (1 corresponding to 10, 2 to 20, etc... in the second condition). The reason this is such a nice experiment is that the numerical ratios involved are matched — the only difference between the two conditions is the actual quantities. To the extent that the necessarily large ratios between small quantities is the reason why they are easy to estimate and discriminate, then, there should be no difference in performance between the two conditions. However, they found that there was in fact a large difference between the two conditions in the 1-4 and 10-40 ranges: people could estimate 1-4 with near-exact precision but not 10-40. This study therefore shows that small-quantity psychophysics differ fundamentally from large-quantity psychophysics.

How should we interpret these distinct psychophysical modes? In their final paragraph, Revkin et al. (2008) wrote,

> In conclusion, although our study provides evidence against estimation as the underlying mechanism of subitizing, the question remains open as to whether subitizing relies on a domain-specific numerical process or on a domain-general cognitive process. One hundred years after the discovery of subitizing, its mechanisms remain as mysterious as ever — but it is now clear that they are not based on a Weberian estimation process.

One interesting thing to note about this statement is, as in Kaufman et al. (1949), the finding of a dissociation in performance between small- and large-quantities is taken as evidence of distinct *mechanisms*. However, differences in empirically measured psychophysical properties do not necessarily imply differences in mechanisms — this is actually an inductive leap. Chapters 3 and 4 describe why such an inference, while intuitively appealing, is not justified.

Number as a primary perceptual attribute

The study of numerosity perception, while quite old by the standards of psychology, has retained substantial interest in recent decades. There are several reasons for this, including: 1) that basic numerical abilities seem to be present even in uncomplicated animals and are observed early in human development; 2) that number psychophysics are unique in showing a discontinuity; and 3) that number seems to be a *primary perceptual attribute* (Burr & Ross, 2008; Ross & Burr, 2010). Number can be considered a primary perceptual attribute because it can be estimated even when holding correlated dimensions constant (e.g. the density and total area of objects) and because numerical percepts are susceptible to adaptation.

Upon reflection, this is quite surprising given how far removed a set's quantity seems to be from direct sensory input. Specifically, computing the number of objects in a scene would seem to require a map of individual items in space, which by itself is a computationally complex problem involving image segmentation and object detection. Yet, remarkably, people can determine the number of items in an image given as little as 16 milliseconds of exposure (Inglis & Gilmore, 2013), albeit with low precision. Equally remarkable is that, as the adaptation studies show, repeated exposure to large numerosities decreases one's perception of a novel set's quantity (and vice-versa), even when holding other dimensions constant (Anobile et al., 2014, 2016; Burr & Ross, 2008; Ross & Burr, 2010) — just as with much simpler perceptual attributes like luminance.

The abstractness of number has led naturally to skepticism that it could be a primary quality of perception. Some researchers have maintained that perception of numerosity is actually *indirect*, and that numerical estimates are actually the product of extrapolation from less abstract continuous features that strongly correlate with numerosity, such as surface area or texture density (Dakin et al., 2011; Durgin, 2008). Others posit the existence of a “general magnitude system” in which the representations of density, area, and numerosity are all combined, inseparably flattened onto common magnitude scale (Gebuis et al., 2016; Gebuis & Reynvoet, 2012a; Lourenco & Longo, 2010; Sokolowski et al., 2017). Studies have in fact found that numerical estimates are biased by the texture density and total surface area of items in a display (Aulet & Lourenco, 2021; Dakin et al., 2011; Gebuis & Reynvoet, 2012a; Lourenco & Longo, 2010), which has been taken as evidence against numerosity as a primary perceptual attribute and in favor of a “general magnitude system”.

However, the finding that numerical estimates are biased by other cues is not at all surprising given the exceptionally high empirical correlation between number, area, and density in natural scenes (e.g. Piantadosi & Cantlon, 2017) and the visual system’s propensity for making efficient use of natural scene statistics in general (Olshausen & Field, 1996; Simoncelli & Olshausen, 2001) and correlated dimensions in particular (Bates & Jacobs, 2020; Orhan & Jacobs, 2013). But more importantly, there is by now overwhelming evidence that number, area, and density can all be estimated independently without training (Anobile et al., 2018; Anobile et al., 2014; Cicchini et al., 2016; Ferrigno et al., 2017; Yousif & Keil, 2019); that from an early age numerosity specifically seems to be privileged in salience over area and texture density (Anobile et al., 2019; Ferrigno et al., 2017) and represented with higher precision than either as well (Cicchini et al., 2016); and that numerosity perception is susceptible to adaptation even controlling for correlated dimensions (Arrighi et al., 2014; Burr & Ross, 2008; Fornaciai et al., 2016; Ross & Burr, 2010).

The “number sense” idea

While many non-human animals have the capacity to approximately estimate, discriminate, and manipulate quantities, humans are the only species to use exact, symbolic systems for representing and manipulating numbers — a development which has made possible the inventions of science, mass industry, and civilization (see O’Shaughnessy et al. (2021) for a review of the cultural origins of number systems). However, innate numerical abilities have garnered attention from researchers interested in the historical origins of number knowledge and symbolic mathematics, as well as the developmental trajectory of numeracy and mathematical understanding in children. The term “number sense” became a catchphrase broadly connoting a link between innate and learned numerical abilities. It has also been used frequently to suggest the more specific idea that inexact quantity representations are a precursor to, and may play an important developmental role in, acquiring mathematical knowledge. For instance, in his popular book *The Number Sense*, Dehaene (1997) writes that,

> The foundations of arithmetic lie in our ability to mentally represent and manipulate numerosities on a mental ‘number line’, an analogical representation of number; and that this representation has a long evolutionary history and a specific cerebral substrate.

The origins of the (often somewhat vague) supposition that counting and arithmetic have their “foundation” in non-verbal, innate quantity representations can be traced to theories developed by Rochel Gelman and Charles Gallistel in the mid-1970s. On their early account of the relationship between pre-verbal number representations and symbolic number systems, people are actually born with an innate understanding of exact quantities even if their perception of numerosity is only approximate (Gallistel & Gelman, 1992; R. Gelman & Gallistel, 1978). On their account, then, learning to count is essentially the process of mapping one’s non-verbal integer representations to number words. However, subsequent studies have demonstrated that learning the count list is actually a pre-requisite to representing and manipulating exact quantities — i.e., an understanding of integers is *constructed from* rather than *mapped to* number words (Cheung et al., 2017; Le Corre & Carey, 2007; Piantadosi et al., 2012; Sarnecka & Carey, 2008; Schneider, Feiman, et al., 2021; Schneider, Pankonin, et al., 2021; Wynn, 1992b).

For instance, it is now well-established that children do not exhibit an understanding of the “cardinal principal” — that each successive number in the count list maps to one more item in a set — until well after they have mastered the count list (Le Corre & Carey, 2007; Wynn, 1992b). Furthermore, even after a child achieves an understanding of the cardinal principal — often assessed by their being able to provide n items when asked for "n" — there will generally be a long gap before they seem to map larger number words to their innate quantity representations. A key observation that supports this view is that in number estimation tasks, younger children who have only recently mastered the "give-n" task will provide estimates that do not monotonically increase with the number of objects shown, particularly for larger sets (Le Corre & Carey, 2007). The evidence therefore implicates a limited role, if any, for inexact, non-verbal number representations in learning to count.

However, confusing matters, Halberda et al. (2008) found that performance in a non-symbolic number discrimination task correlates strongly with high school mathematical achievement as measured by a standardized test (the TEMA-2) — over and above many other factors such as IQ. Some subsequent studies have replicated the general finding of a positive correlation between estimation acuity and mathematical achievement (Bonny & Lourenco, 2013; Feigenson et al., 2013; Halberda et al., 2012; Libertus et al., 2011, 2013; Mazzocco et al., 2011; Starr et al., 2013b; Wagner & Johnson, 2011). However, most did not control for the range of factors that Halberda et al. (2008) did, leaving the primary determinants of early mathematics learning unresolved. In fact, a number of studies have found that the relationship between estimation acuity and math achievement is highly dependent on whether other number-related abilities and non-numeric cognitive factors, such as inhibitory control, are controlled (Caviola et al., 2020; Fuhs & McNeil, 2013; Holloway & Ansari, 2009; Kolkman et al., 2013; Lyons et al., 2014; Price et al., 2012). Studies that *have* controlled for other factors have largely failed to find any association between mathematical achievement and estimation acuity (Caviola et al., 2020; Holloway & Ansari, 2009; Kolkman et al., 2013; Lyons et al., 2014; Price et al., 2012) or highlighted the importance of additional factors in addition to non-symbolic numerical acuity, like knowledge of the counting list and domain-general cognitive factors (Mou et al., 2018).

Others have reported causal evidence in support of this link from training studies, with even brief training on a non-symbolic task seemingly improving arithmetic abilities (DeWind & Brannon, 2012; Hyde et al., 2014; Park & Brannon, 2013, 2014; Wang et al., 2016; Wilson et al., 2009). The results from these studies are mixed, potentially due to weaknesses in the training efficacy (Lindskog & Winman, 2016), and include a recent failure to replicate training effects (Szkudlarek et al., 2021). A thorough critique and review of the literature by Szűcs and Myers (2017) found that prior studies often had low power and high false positive rates, uncritically cited other papers, and did not evaluate strong alternative hypotheses, concluding that "there is no conclusive evidence that specific ANS training improves symbolic arithmetic."

1.2 Scope of this thesis

The literature on numerical cognition in general, and visual numerosity perception in particular, is quite substantial — and the psychophysics of numerosity perception are accordingly well-characterized. However, there is still significant ongoing debate about how to interpret the observed psychophysics and a basic lack of clarity regarding the underlying mechanisms. There are, in addition, some widely-held assumptions about the mechanisms supporting numerosity perception that, on closer examination, are weakly supported or even untested. These include the ideas that: 1) quantity estimation operates in parallel across an entire scene; 2) people have an intrinsic Weber fraction which is a static measure and reflective of their intrinsic "number sense"; 3) perception of small quantities is always exact; 4) underestimation is driven by mis-calibration of one's response scale; 5) a discontinuity in estimation error implies multiple mechanisms; 6) the approximate number system itself, rather than lower-level perceptual uncertainty, accounts for the psychophysics of estimation; and 7) numerosity must be represented as an analog magnitude on a continuous scale because numerical discrimination follows Weber's law. This thesis contains experiments, models, and analyses that are centered on understanding the visual mechanisms supporting numerosity perception. The findings presented here challenge some widely held assumptions, including all of the ones listed above, and offer new ways of understanding some of the most puzzling aspects of numerosity perception.

Chapter 2 examines the visual mechanisms of large-number estimation[4]. We

[4] This work was published in *PNAS*, as Cheyette and Piantadosi (2019), and can be found at https://www.pnas.org/content/116/36/17729.short.

ask how people's perception of numerosity changes over time and as a function of what they view. We present data from an estimation task where participants' visual fixations were recorded with an eye-tracker and test how participants' visual attention mediates their estimates in a model-driven data analysis. We find that perception of quantity is the result of a serial accumulation process operating over saccades: as participants fixate on more objects, their quantity estimates increase and the variance of their estimates decreases. This finding contrasts with the standard picture of estimation as a static process, as embodied in myriad feedforward neural network models of numerosity perception, and of Weber fractions as a simple index of a person's number sense.

Chapter 3 addresses the origins of small- and large-number psychophysics, centering on the curious discontinuity in estimation error between four and five objects[5]. Using information-theoretic methods, we derive the optimal way to represent numerosity given their natural "need frequency" distribution, along with an informational capacity limit. We show that differential patterns of errors for large and small numbers (including exactness in the subitizing range and Weber's law for high numbers), an underestimation bias in mean estimates, and approximately Gaussian-shaped response distributions all arise from this optimization. We present data from four numerical estimation tasks that support key predictions of the model, such as a gradient shrinkage of the subitizing range and concurrent decrease in precision of large number estimates as exposure time is decreased or the color contrast of the objects is lowered.

Chapter 4 investigates whether number psychophysics arise from one or more "number systems," as is commonly believed, or if they have their origins in lower-level perceptual processing [6]. We propose an adaptation of the model developed in Chapter 3, which optimally represents objects in space (rather than quantities) subject to an information capacity constraint. Quantities are only *implicitly* represented in this model, unlike in the direct optimization of numerosity perception in Chapter 3. However, we show that many of the important psychophysical phe-

[5] This work was published in *Nature Human Behaviour*, as Cheyette and Piantadosi (2020), and can be found at https://www.nature.com/articles/s41562-020-00946-0.

[6] A paper containing an early version of this work won the modeling prize in Perception and Action at the *Cognitive Science Conference*, where it was published in the proceedings as Cheyette et al. (2021), and can be found at https://escholarship.org/uc/item/9hk7s32c.

nomena associated with numerical cognition — including subitizing and Weber's law — can be derived as downstream consequences. We report the results of two experiments — a non-numerical spatial memory task and a numerical estimation task — which show that participants' beliefs about numerosity are constrained and ultimately determined by their ability to locate and track objects, consistent with predictions of the model. This work casts doubt on prominent theories regarding the role of domain-specific number systems in determining the psychophysics of estimation and suggests a re-interpretation of the notion that number itself is a primary perceptual attribute.

2 The visual mechanisms of numerical estimation

2.1 Introduction

From infancy, humans are able to estimate and compare quantities (e.g. Dehaene, 2011; Hyde et al., 2010; McCrink & Wynn, 2007; Xu & Spelke, 2000), an ability shared with our close and distant evolutionary relatives (e.g. Cantlon, 2012; Meck & Church, 1983; Yang & Chiao, 2016). There is ongoing debate over whether and how innate numerical abilities underpin the development of symbolic mathematical reasoning in humans (Dehaene, 2011; Feigenson et al., 2004; Halberda et al., 2008; Starr et al., 2017); however, the defining feature of innate numerical estimation is that it is *inexact*, providing approximate representations of numerosities which are likely useful in a variety of evolutionary contexts (e.g. Cantlon, 2012; Gross et al., 2009; Piantadosi & Cantlon, 2017; Yang & Chiao, 2016). The precision of numerical estimation and discrimination is often quantified in terms of a *Weber fraction*, w, which is a real number denoting how the noise in a representation scales with numerosity. Specifically, one popular psychophysical model of estimation assumes that a number n is represented by a Gaussian with mean n and standard deviation $w \cdot n$, so that a lower w implies a higher fidelity system.

The mechanisms supporting innate numerical estimation are often contrasted with other mechanisms for computing numerosity, such as counting and subitizing (Anobile et al., 2014; Burr et al., 2010; Revkin et al., 2008). Counting, for instance, is dependent on intentional, serial enumeration of a set; approximate estimation, in contrast, is often viewed as parallel, rapid, and automatic. This view is supported by response times, where counting takes around 300ms per enumerated item but approximate number computations can take as little as 16ms independent of the number of objects (Inglis & Gilmore, 2013). Additionally, researchers have identified populations of neurons that respond similarly for sequentially and simultaneously-presented numerosities in monkeys (Nieder et al., 2006), which has been taken as evidence that approximate number representations are not the result of sequential processing.

However, recent evidence has muddied the simple picture of numerical estimation. Several studies have shown that individuals' Weber fractions are highly task-dependent, differing between estimation and discrimination tasks (e.g. Guillaume & Gevers, 2016; Price et al., 2012). In fact, Weber fractions have poor retest reliability even when measured using the same task (Inglis & Gilmore, 2014). Numerical estimates have also been found to be influenced by non-numerical features of stimuli, such as the degree of clustering in a scene (Im et al., 2016). Finally, the precision of numerical estimates is known to improve as stimuli are presented for a longer duration (Inglis & Gilmore, 2013), suggesting that estimation may involve some type of temporal process.

Despite this, prior computational models of estimation have built speed and parallelism into their architecture. For instance, many of the dominant models of the so-called "approximate number system" (ANS) are feedforward neural network models where input is processed in parallel and instantaneously (e.g. Dehaene & Changeux, 1993; Stoianov & Zorzi, 2012; Testolin, Dolfi, et al., 2020; Verguts & Fias, 2004; Zorzi & Testolin, 2018). The objective of the present study is to critically evaluate the simple picture of numerical estimation as a rapid and entirely parallel process. In particular, we aim to capture the possible sequential mechanisms involved in numerical estimation using behavioral experiments and model-driven analysis. We present a new model and behavioral data from two experiments that challenge the standard parallel perception theory. Our results lend support instead to an account of estimation that involves sequential integration across visual fixations.

We ran estimation (Experiment 1) and discrimination (Experiment 2) tasks in which participants made non-symbolic numerosity judgments at different exposure durations. Critically, we collected visual fixation data using an eye-tracker so that we could measure how participants' ANS estimation was influenced by their path of visual fixations. We show that ANS estimates are the result of a serial accumulation process (Gallistel & Gelman, 2000), such that estimates increase as a function of foveation. We present an analysis that quantifies the contribution of foveal, peripheral, and multiply-fixated dots in an array which supports this interpretation. Our results suggest that individual differences in estimation acuity may reflect differences in cognitive processes that are not directly related to number, including attention and visual processing speed. This dependence on non-numerical factors may explain why studies that train people's ANS yield mixed results in transferring to mathematical knowledge (Hyde et al., 2014; Lindskog & Winman, 2016; Park et al., 2016; Park & Brannon, 2013).

2.2 Experiment 1

Since the visual mechanisms supporting the ANS have not been explored in detail, we first used the simplest paradigm possible in order to understand ANS estimation. Figure 2.1 illustrates the sequence of displays shown on each trial. After viewing a fixation cross, participants were shown an array of randomly placed dots on a screen which were noise-masked after a short time. They were then prompted to enter an estimate in Arabic numerals. Subjects were not given feedback and thus had no push to re-calibrate their response scale.

Methods

Participants

27 adult subjects (15 female, 12 male) from the University of Rochester community were recruited to participate in the task. The participants' ages ranged from 18-29 ($M = 21.4$).

Materials

The screen subtended approximately 38° of participants' visual field left-to-right and 26° top-to-bottom. The eye-tracker was a Tobii T60XL model, providing a

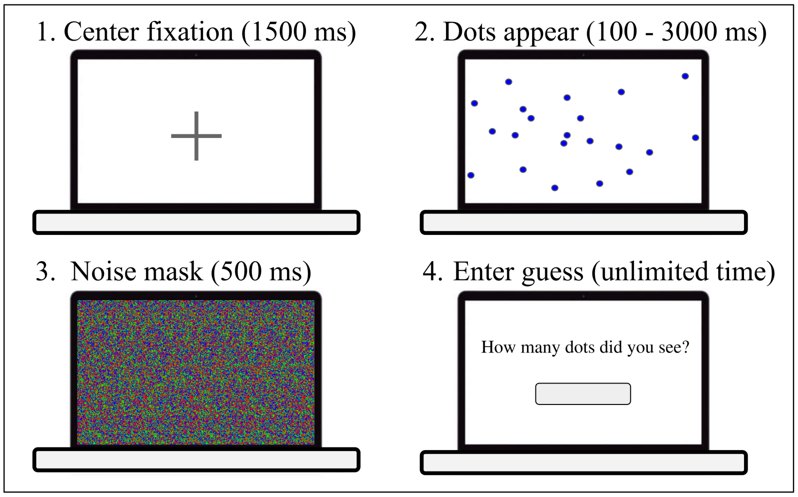

1. Center fixation (1500 ms)
2. Dots appear (100 - 3000 ms)
3. Noise mask (500 ms)
4. Enter guess (unlimited time)
How many dots did you see?

Figure 2.1: Each of the four panels represents one stage of a trial in the estimation task in their order. **Panel 1:** A fixation cross appears for 1500 milliseconds. **Panel 2:** The fixation cross is removed and dots appear on the screen for between 100 milliseconds and 3 seconds depending on the condition. **Panel 3:** The display is masked by noise for 500 milliseconds. **Panel 4:** A prompt appears asking for an estimate of the number of dots shown.

readout of 60 samples per second. We used built-in Tobii software to calibrate participants to the eye-tracker. The computer screen was 24 inches, with an aspect ratio of 16:10 and screen resolution of 1920x1200 pixels. Each dot had a radius of 10 pixels. The density of the dots in the images ranged from $0.01 - 0.07\ dots/deg^2$. The dots were placed on the screen at random locations, only constrained to be non-overlapping. Participants entered their numerical estimates using a keyboard attached to the computer. The experiment was designed using the Python library Kelpy (Piantadosi, 2012).

Design

The experiment consisted of 64 total trials, with four blocks of 16 trials each, in which participants viewed arrays of dots and estimated their quantity. Each 16-trial block contained one of the four different time conditions each subject underwent: $100ms$, $333ms$, $1000ms$, and $3000ms$ (together comprising all 64 trials); the order of the blocks was randomized across participants. The number of dots displayed on each trial varied between 10 and 90 dots, inclusive. To determine the numerosities shown to a given subject, 16 numbers were chosen randomly from within that range. The same 16 numbers were shown to the participant in each 4 time conditions, with presentation order randomized across the conditions. The median range size across participants was 71 (minimum 54, maximum 79). The median lowest number shown was 14 and the median highest number shown was 86.

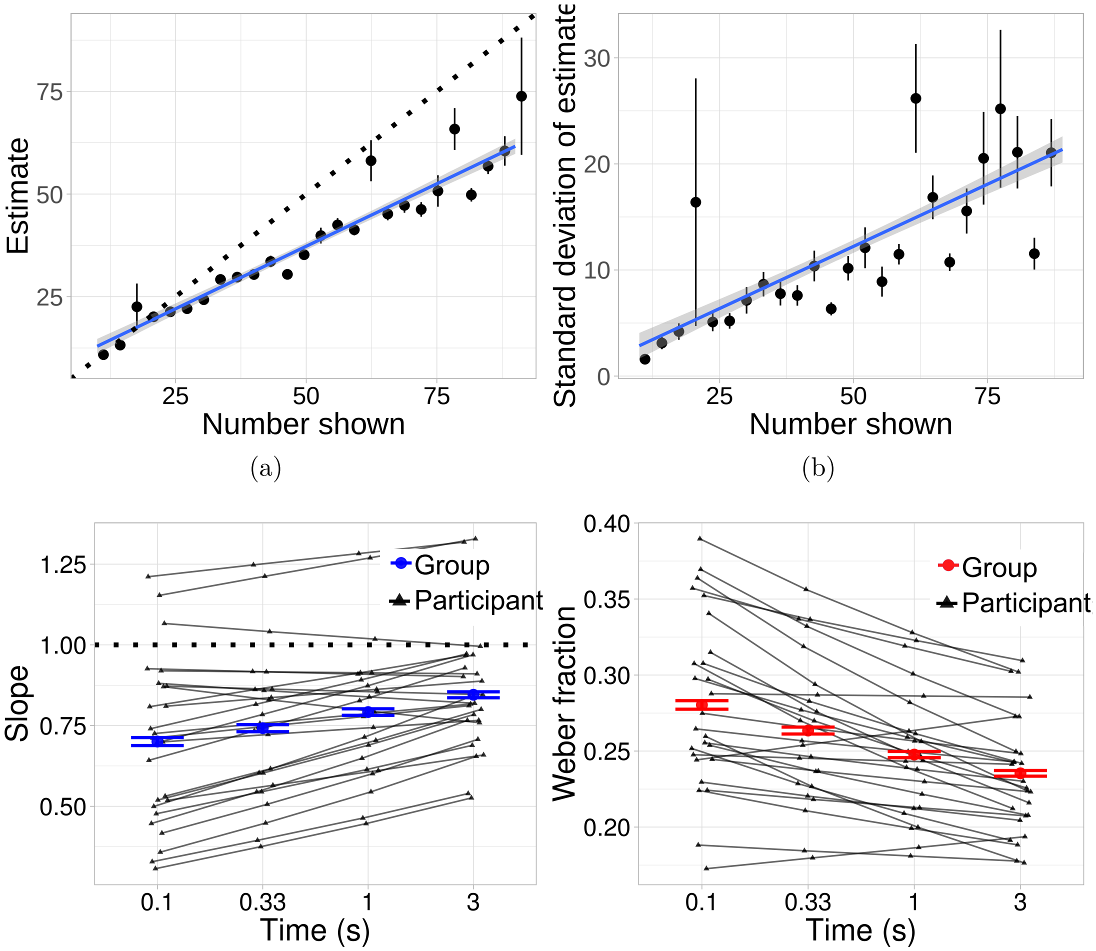
Procedure

All study procedures were approved by the University of Rochester IRB. After providing consent, participants were placed directly in front of a computer, with the eye-tracker mounted on top. The screen sat on an adjustable desk, which was vertically re-aligned for each participant to ensure that that the center of the screen was level with their eyes. The participants were fixed to a distance such that their eyes were 26 inches away from the screen, which was ensured by measurement with a yardstick. On each trial, dots were displayed, followed by a noise mask. Subjects then typed their responses into a text box using a keyboard and pressed the enter key to move onto the next trial.

Results

Replication of basic number psychophysics

Figure 2.2a shows how the mean estimate (y-axis) varied as a function of the quantity displayed (x-axis), collapsing over all time conditions. There are two aspects of this graph worth highlighting: first, mean estimates vary approximately linearly as a function of quantity, exactly as should be found in Weber models of the number system. Second, this shows a strong tendency to increasingly underestimate larger numbers, as shown by the fact that the *slope* of the line is less than 1, which would have corresponded to perfectly veridical estimation (assuming an intercept of 0). Both effects have been found robustly in the literature previously (e.g. Izard & Dehaene, 2008). Figure 2.2b shows that Experiment 1



(a)

(b)

(c)

(d)

Figure 2.2: Panel (a): Estimates as a function of the number of dots presented, collapsing across time conditions. Points are binned means, with errors representing bootstrapped 95% confidence interval. Panel (b): The standard deviation of participant's estimates as a function of the number of dots displayed, collapsing across time conditions. Panel (c): Participant (black) and group-level (blue) slopes in each time condition of the estimation task are shown. Slopes represent the way the mean estimate scales as a function of quantity shown. Panel (d): Participant (black) and group-level (red) Weber fractions in each time condition of the estimation task are shown.

replicates the second traditional property of ANS estimation: *scalar variability*, wherein the error in estimation increases linearly in magnitude.

More time improves estimation mean and variance

To evaluate whether timing influenced participants' ANS, we ran a hierarchical regression to estimate the effect of time on both the mean estimate and Weber fraction, including participant- and group-level regression effects fit jointly. The model assumes that each individual's mean estimate and standard deviation about that estimate vary linearly as a function of the quantity displayed and logarithmically as a function of time. We will call participants' baseline (independent of time) slopes and Weber fractions β_0 and w_0; we will denote time t; and we will call the effect of time on slopes β_t and Weber fractions w_t. In order to keep slopes and Weber fractions positive, we assume an exponential linking function between slope and the predictors. Specifically, Equations 2.1 and 2.2 show how the slope and Weber fractions for each participant are calculated:

$$\beta = e^{\beta_0 + \beta_t \cdot log(t)} \tag{2.1}$$

$$w = e^{w_0 + w_t \cdot log(t)} \tag{2.2}$$

Then, each participant's mean estimate is drawn from a Gaussian centered around $\beta \cdot n$ with standard deviation $w \cdot \beta \cdot n$.

Figure 2.2c shows the mean slope and Figure 2.2d shows the mean Weber fraction in each time condition extracted from this model. The group-level means are shown in blue and each participant is shown by a line in black. If participants' estimates were unbiased (e.g veridical estimation as opposed to under-estimation) then the group mean slopes would be 1 (black dotted line) and if time did not have an effect, the group mean slopes and Weber fractions (y-axis) would remain constant across time (x-axis). In contrast, Figure 2.2c shows that subjects consistently underestimate with slopes less than 1, but that this underestimation effect decreases with increasing time. Participants' average slope increases by about 17% (0.71 to 0.83) from the shortest to the longest time condition. This is what would be expected by quantity accumulation over time: more time increases reported quantities. Additionally, their average Weber fraction decreases by about

21% (0.28 to 0.22). Correspondingly, Figure 2.2d shows that Weber fractions improve (decrease) with more time.

Var	**Value**	**2.5%**	**97.5%**
β_0	-0.24	-0.28	-0.19
β_t	0.05	0.03	0.09
w_0	-1.42	-1.59	-1.21
w_t	-0.11	-0.12	-0.09

Table 2.1: Group-level regression weights and their 95% credible intervals for each condition. For the mean slope, the inferred weights include the intercept (β_0), the effect of time (β_t). For Weber fractions, the inferred values are analogous, with the intercept (w_0), the effect of time (w_t).

Table 2.1 provides the inferred group-level regression weights and the uncertainty of the estimate. The fact that the intercept is negative ($\beta_0 = -0.24; CI = [-0.28, -0.19]$) indicates a baseline tendency to underestimate. Most significantly, the effect of time on both mean slopes and Weber fractions is significantly different than 0: time increases the group mean slope ($\beta_t = 0.05$; $CI = [0.03, 0.09]$)[1]; and it decreases the group mean Weber fraction ($w_t = -0.11$; $CI = [-0.12, -0.09]$). Participants' average slope increases by about 17% (0.71 to 0.83) from the shortest to the longest time condition; and their average Weber fraction decreases by about 21% (0.28 to 0.22).

Foveation, not time, is what matters for estimation

If ANS estimation is driven by accumulation of quantity across saccades, we should first expect that mean estimates increase with foveation. We should also expect that time has *no* effect when jointly considering foveation — i.e., that time simply allows for more saccades and nothing more. To evaluate this, we summed the number of dots that fell within 5° (often called the "para-foveal region") of the center of participants' fixation paths for more than 50ms on a trial.[2] We denote the dots that are seen for at least this amount of time as "foveated."

[1] CI here indicates the *credible interval*, not the *confidence interval*.

[2] We also tested 16ms, 100ms as possible thresholds; and 2° and 10° as possible visual degrees thresholds. These differences did not affect the qualitative pattern of results.

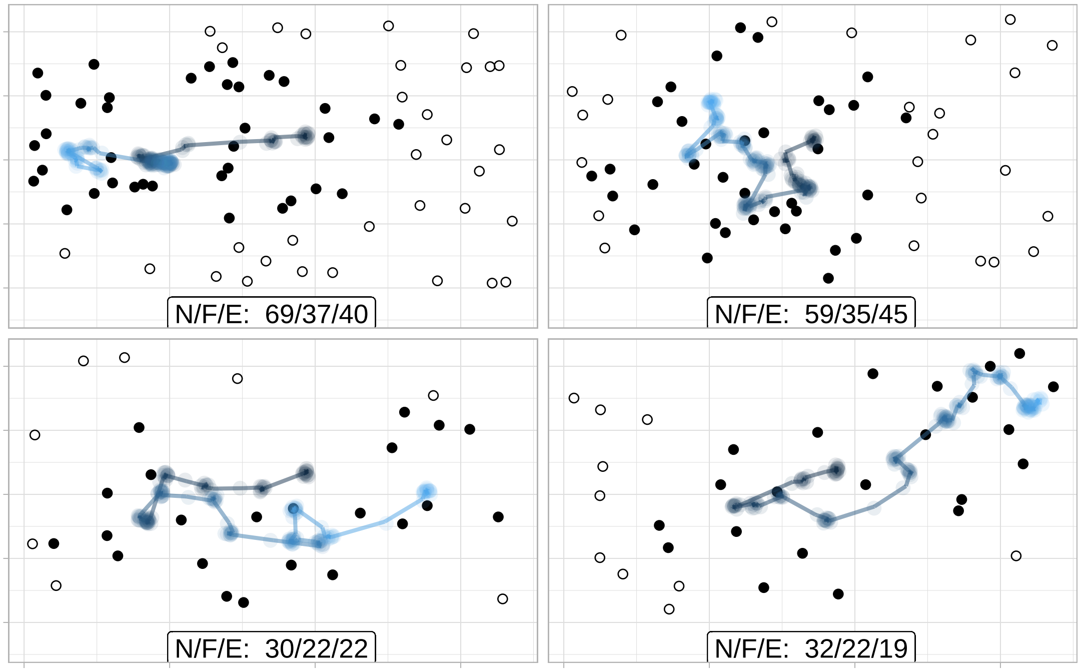
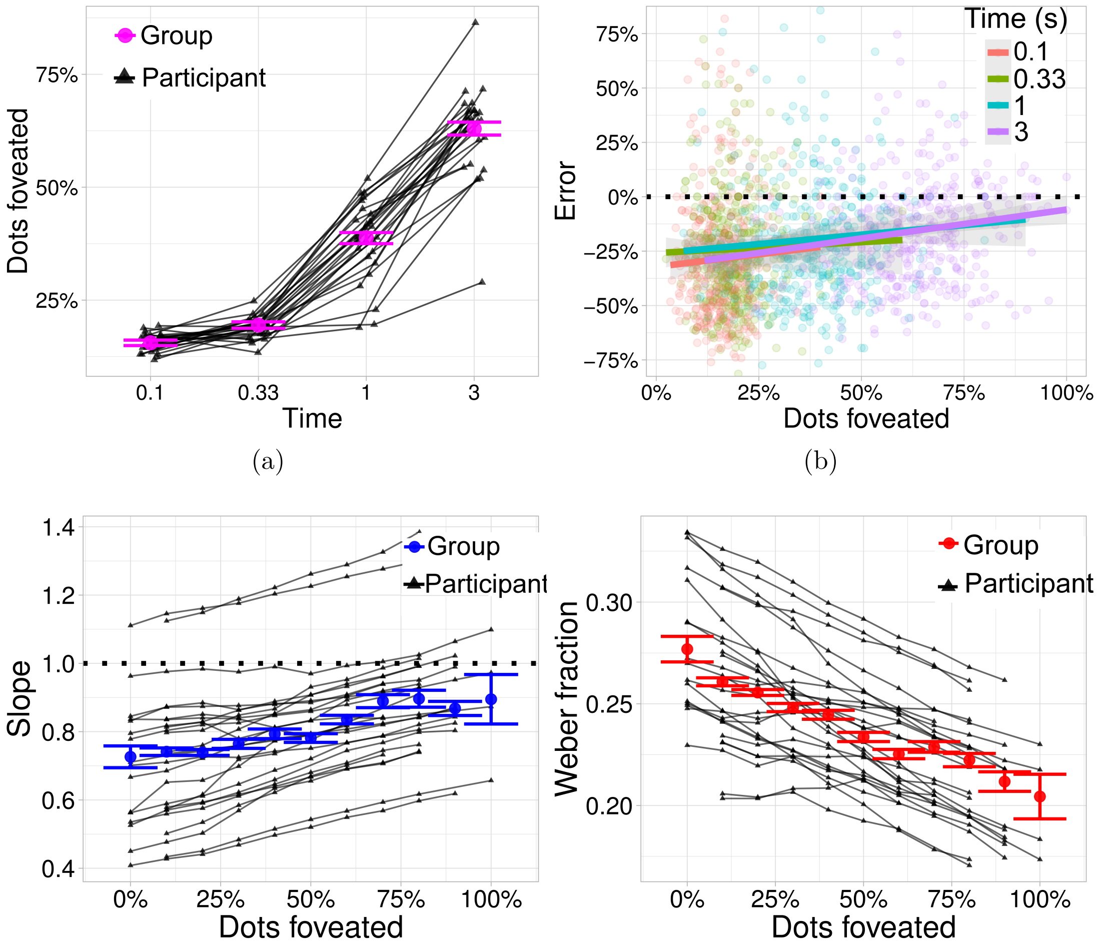


Figure 2.3: Example fixation paths of one subject in the 3-second time condition, with each panel representing a single trial. The points represent the dots displayed on their screen, where filled dots represent the ones that were foveated. At the bottom of each panel, a label N/F/E shows how many dots were shown (N), how many were foveated (F) and what quantity the participant actually estimated (E).

Figure 2.3 provides four example trials, depicting a participant's gaze path across the screen while the stimulus is being shown. The filled points represent "foveated" dots and the unfilled points represent those that were not[3]. At the bottom of each display, the number of dots shown, the number foveated, and the participant's estimate is shown. We provide a more rigorous formalization and test of this idea in Section 2.4.

Figure 2.4a shows the percent of dots that are foveated for each time condition. As should be expected, more dots are foveated with longer exposure duration. The average proportion of dots foveated more than triples from the shortest to longest time condition (18% to 64%). Consistent with the hypothesis that effects

[3] This is for illustrative purposes only — stimuli were entirely static during a trial.



(a) (b)

(c) (d)

Figure 2.4: Panel (a): The proportion of dots foveated (y-axis) as a function of time (x-axis), at the group-level (red) and for each participant (black). Panel (b): The percent deviation of estimates from the true number of dots (y-axis) as a function of the percent of dots foveated (x-axis). Each time condition is grouped by color. Panel (c): The slope of participants' mean estimates (y-axis) as a function of the percent of dots foveated (x-axis). Panel (d): Weber fractions (y-axis) as a function of the percent of dots foveated.

of time are due to accumulation of foveated dots, the effects of time on estimation disappear when the effect proportion of dots foveated is jointly taken into account. Figure 2.4b shows the percent deviation of estimates from the true quantity as a function of dots foveated, colored by time. That the lines overlap suggests that there is no effect of time when both foveation and time are taken into account.

To quantify whether the effect of time was explained by eye movement patterns, we ran another Bayesian regression that was identical to the one reported above, with the addition of random variables for the effect of the *proportion of dots foveated* on the mean and variance of each participant's estimate. That is, we used the same terms as in equations 2.1 and 2.2, but added terms β_s and w_s to the way slope and Weber fraction are computed — each term is multiplied by s, representing the proportion of dots foveated. Equations 2.3 and 2.4 show the calculation of β and w in full.

$$\beta = e^{\beta_0 + \beta_t \cdot log(t) + \beta_s \cdot s} \tag{2.3}$$

$$w = e^{w_0 + w_t \cdot log(t) + w_s \cdot s} \tag{2.4}$$

Table 2.2 shows the results of this analysis in full. Two findings are worth highlighting. First, the proportion of dots foveated significantly affects the mean and variance of participants' estimates. Second, time no longer has a significant effect on either. The effect of foveation on the mean can be seen in Figure 2.4c: as the proportion of dots foveated increases, so do participants' mean estimates.[4] Congruently, the regression shows that the group-level mean β_s is significantly above 0 ($\beta_s = 0.43; CI = [0.26, 0.59]$). There is also an effect of foveation on the variance of estimates. Figure 2.4d shows that as foveation increases, Weber fractions tend to decrease. The group-level regression revealed that the effect of foveation is significant ($w_s = -0.67, CI = [-0.95, -0.27]$). Finally, consistent with our hypothesis, the effects of time were no longer significantly different than 0 when accounting for the visual samples. The lack of a time effect on the mean (when conditioning on percent of dots foveated) can be seen clearly in Figure 2.4b which shows the average deviation as a function of dots foveated

[4] Note that the lines are non-linear in Figures 2.4c and 2.4d because the y-axis measures are collapsed over other predictors.

and colored by time: if there were a significant effect of time over-and-above the differences driven by fixations, the regression lines for each time condition would be non-overlapping.

Var	**Value**	**2.5%**	**97.5%**
β_0	-0.74	-0.87	-0.65
β_t	-0.00	-0.02	0.03
β_s	0.43	0.26	0.59
w_0	-0.98	-0.74	-1.21
w_t	0.05	-0.03	0.12
w_s	-0.67	-0.95	-0.27

Table 2.2: Group-level regression weights and their 95% confidence intervals for each condition. For the mean slope, the inferred weights include the intercept (β_0), the effect of time (β_t), and the effect of the percent of dots seen (β_s). For Weber fractions, the inferred values are analogous, with the intercept (w_0), the effect of time (w_t), and the effect of the percent of dots seen (w_s).

Thus, these results provide an alternative account of prior findings of (i) underestimation and (ii) effects of time. Indeed, both are unified into an account where serial accumulation of foveated dots drives numerical quantity estimates. This finding calls into question the construct validity of Weber ratios as a measure of an individual's innate "number sense," since numerical estimates depend on how many dots happen to be foveated, a capacity which is non-numerical.

Exploratory analyses

Foveation analysis

It will be important for future research to better determine how saccades are programmed, since this may explain some influence of the properties of visual displays on numerical estimation (Burr & Ross, 2008; Dakin et al., 2011), including biases introduced by object clustering (Im et al., 2016). Research has shown that a pre-saccadic selection process takes place over competing regions in the visual periphery (Fischer & Weber, 1993). Indeed, in our tasks, there is likely a non-random nature to participants' saccades — this is suggested by the fixation paths in Figure 2.3. After starting at center fixation, participants tended to fixate regions of the screen that had higher density; and their gaze remained in higher density regions for longer. We computed the mean x- and y- coordinates of the displayed dots for each trial in the 3-second time condition (where participants could saccade freely). We found that the mean x-coordinate of the dots significantly correlated with the participant's mean x-coordinate gaze ($r = 0.27, p < 0.001$); this is likewise true for the y-coordinates ($r = 0.34, p < 0.001$). This is consistent with previous results showing that people tend to look towards the greater of two quantities first and for longer in a quantity discrimination task (Odic & Halberda, 2015).

Analysis using cortical magnification factor

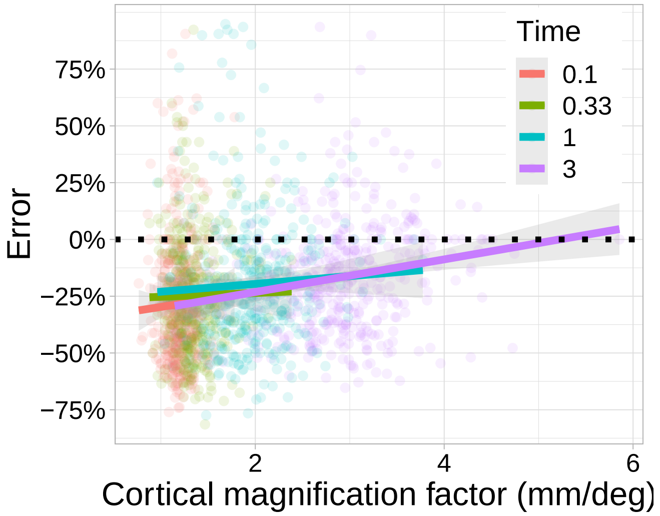


Figure 2.5: The percent deviation of estimates from the true number of dots (y-axis) as a function of the percent of the average cortical magnification (x-axis) of each dot displayed in a trial. Each time condition is grouped by color.

In our primary analysis, we considered a dot "foveated" if it fell within a 5° window around someone's gaze for more than 50ms. While this measure has the benefit of simplicity, the exact values are somewhat arbitrary. A somewhat more complicated, but less ad hoc approach, would be to use the *cortical magnification* *factor*, which is known to predict visual acuity (e.g. Cowey & Rolls, 1974). The cortical magnification factor (CMF) is inversely proportional to the eccentricity of an object from someone's gaze. For each trial, we calculated the CMF for each dot based on the minimum distance between each dot and the participants' gaze (in terms of visual degrees). We took the mean CMF over all dots and used that as a predictor of the mean and variance of estimates. The results revealed no substantial differences between this metric and the one used in the main text. Figure 2.5 shows, for example, the relationship between the CMF and participants' errors, which follows the same qualitative pattern as Figure 2.4b.

Analysis using convex hull

There is an ongoing debate about the importance of continuous variables such as area, density, and convex hull on number estimation (e.g. Ferrigno et al., 2017; Gebuis & Reynvoet, 2012a, 2012b; Starr et al., 2017). Our experimental design was not suited to testing whether people were using these types of heuristics (nor was it intended to). In particular, total area was perfectly correlated with number since the size of the dots was constant across trials. Convex hull was also strongly correlated with the total number of dots displayed, however there was enough random variation to allow us to test its influence on numerical estimation. Given the dependence of estimates on eye movements, one might expect greater underestimation from displays with greater convex hull. To evaluate this, we ran a regression to predict percent estimation error (signed, so not absolute error) from the number of dots shown, the proportion of dots foveated, and the convex hull of the dots. There was a significant effect of the number of dots on estimation error, such that participants' bias to under-estimate increased with the number of dots ($\beta = -0.24; t = -6.43; p < 0.001$), even controlling for convex hull. The effect of convex hull trended in the same direction, but 1/4th the size and was only marginally significant ($\beta = -0.06; t = -1.74; p = 0.08$). Consistent with previous analyses, foveating a greater proportion of the dots has the opposite effect, pushing estimates higher ($\beta = 0.14; t = 6.14; p < 0.001$). These results therefore indicate that there are strong effects of number over and above convex hull, and weak-to-nonexistent effects of convex hull controlling for number and foveation. Interestingly, the influence of convex hull (to the extent it is present) is in the opposite direction as has been found previously (Gebuis & Reynvoet, 2012b).

2.3 Experiment 2

Because there is evidence that Weber fractions may differ between estimation and discrimination tasks (Guillaume & Gevers, 2016), it is important to replicate these patterns in a discrimination task. We ran a second experiment with the same participants as Experiment 1, again recording participants' gaze. Participants saw two stimuli of dot arrays (as in Figure 2.1) sequentially and were then asked to indicate which had a greater quantity. The properties of the stimuli were identical to those in Experiment 1. We manipulated timing in four conditions, which determined whether the first or second array of dots was visible for longer. Specifically, we crossed long and short durations to give presentation times of 100:100ms, 1000:100ms, 100:1000ms, and 1000:1000ms for the two displays. We predicted that, if ANS estimation relied on foveal accumulation in this task as well, timing would bias participants towards whichever display was presented for longer.

Results

Replication of basic psychophysics

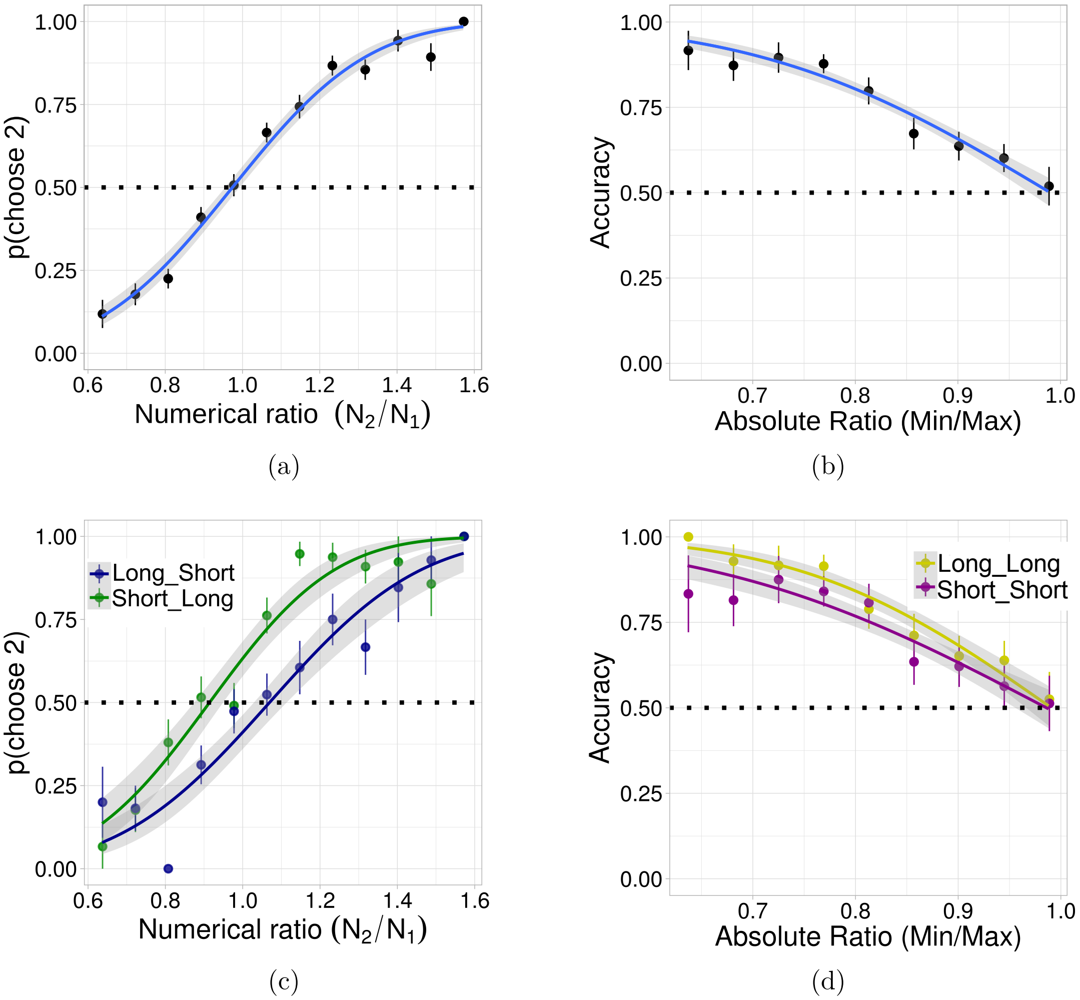
Participants' responses as a function of ratio collapsed across time conditions can be seen in Figures 2.6a and 2.6b. Figure 2.6a shows the proportion of participants who responded that the second display had more dots than the first as a function of the ratio of dots in the second display relative to the first. The proportion participants who responded that the second display was more numerous increased monotonically with the ratio. Participants reported that the second display was more numerous on average (56% of the time), possibly suggesting an effect of memory. This is consistent with studies finding effects of recency in non-symbolic magnitude comparison (Van den Berg et al., 2017). Figure 2.6b shows participants' accuracy as a function of the absolute magnitude ratio, or the minimum magnitude over the maximum. Participants were able to discriminate ratios of 5 : 6 with roughly 75% accuracy. This again replicates known psychophysics of large-number discrimination.

Effects of time

Figure 2.6c shows response curves for the critical conditions where the first and second displays were shown for different amounts of time but the total presentation time is controlled (*Long-Short* versus *Short-Long*). The difference between the curves indicates a bias to choose the second display when it was long compared to when it was short, as predicted. Figure 2.6d shows response curves for the conditions where the first and second displays are shown for the same amount of time but overall presentation time differs (*Short-Short* versus *Long-Long*). The observed difference between the response curves in Figure 2.6d indicates that responses in the *Long-Long* condition were more accurate than those in the *Short-Short* condition. Collapsing across ratios, participants chose the second display 62% of the time in the *Short-Long* condition and 45% of the time in the *Long-Short* condition, as predicted. Participants chose the second display at intermediate (though above-chance) rates in the *Short-Short* (56%) and *Long-Long* (57%) conditions.

Var	**Value**	**2.5%**	**97.5%**
β_0	-1.22	-1.53	-0.99
β_t	0.05	0.04	0.07
w_0	-1.73	-1.14	-2.31
w_t	-0.07	-0.12	-0.02

Table 2.3: Group-level regression weights and their 95% credible intervals for each condition in the discrimination task using only time as a predictor for the mean and variance. For the mean slope, the inferred weights include the intercept (β_0), the effect of time (β_t). For Weber fractions, the inferred values are analogous, with the intercept (w_0), the effect of time (w_t).



(a)

(b)

(c)

(d)

Figure 2.6: Panel (a): The probability that participants responded that the second display had more dots as a function of the ratio N_1/N_2, where N_1 and N_2 are the number of dots in the first and second display, respectively, collapsed across conditions. The fit curve (as well as all other fits in this display) is from a probit regression. Panel (b): Accuracy as a function of the absolute ratio $(Min(N_1, N_2)/Max(N_1, N_2)$. Panel (c): The probability participants responded that the second display had more dots in the Long-Short (blue) and Short-Long (green) conditions as a function of ratio. Panel (d): Accuracy as a function of the absolute ratio in the Long-Long (Yellow) and Short-Short (Red) conditions.

Recency bias cannot explain effects of time

An alternative account of our finding that people respond that the second stimulus is more numerous than the first in the Short-Long condition than the Long-Short condition is that it is a mere effect of *recency* (Van den Berg et al., 2017). We can test this possibility in our design by comparing responses across the conditions where the second stimulus' duration is held constant. That is, comparing responses in the Short-Long to the Long-Long condition; and comparing responses in the Long-Short condition to the Short-Short condition. In both cases, recency could not explain differences in perceived numerosity. If longer duration does, in fact, increase perceived numerosity, then participants should rate the first stimulus in the Short-Long as less numerous than the second relative to the first stimulus in the Long-Long condition.

We ran two logistic regressions to determine the effect of increasing the first stimulus' duration. The first regression was run on only the conditions where the second stimulus was 0.1s (short) and the second regression was run on only the conditions where the second stimulus was 1s (long). The ratio of the numerosities presented was also entered as a predictor. Responding that the first stimulus was greater numerosity than the second was coded as *0* and responding that the second was greater than the first was coded as *1*.

The results of this analysis revealed effects of stimulus duration in the predicted direction. The first regression — looking at the conditions when the second stimulus was short — showed a significant effect of the first stimulus' duration ($\beta = -0.32$; $z = -4.04$; $p < 0.001$), such that increasing the presentation duration of the first stimulus increased participants' likelihood of reporting that the first stimulus was of greater numerosity. Likewise, the same effect was revealed in the second regression ($\beta = -0.19$; $z = -2.22$ $p = 0.03$), looking at the conditions when the second stimulus was long. Taken together, these findings effectively rule out the possibility that a *recency bias* explains the patterns in our data; instead, longer duration of presentation itself seems to increase perceived numerosity.

Effects of foveation

To determine whether participants' visual samples mediate the observed effect of time, we ran a regression analogous to the one used for the estimation task, but predicting the probability of guessing the second array given the participant's mean slope (β_0), mean Weber fraction (w_0), and the effect of the duration of the stimuli on the mean (β_t) and the Weber fraction (w_t). If n_1 is the number of dots in the first display and n_2 is the number of dots in the second display, then the ratio is defined as $\frac{(n_1-n_2)}{\sqrt{(n_1^2+n_2^2)}}$. If t_1 and t_2 are the display times of the first and second stimulus, then the probability a participant chooses the second screen is given below in Equation 2.5.

$$p(choose\ 2) = \frac{1}{2} + \frac{1}{2}erf\left(\frac{n_2 \cdot e^{\beta_0+\beta_t \cdot log(t_2)} - n_1 \cdot e^{\beta_0+\beta_t \cdot log(t_1)}}{\sqrt{2}e^{w_0+w_t \cdot log(t_1)+w_t \cdot log(t_2)}\sqrt{(n_1^2+n_2^2)}}\right) \tag{2.5}$$

Var	Value	2.5%	97.5%
β_0	-0.29	-0.94	0.41
β_t	-0.01	-0.07	0.03
β_s	2.93	0.24	4.01
w_0	-0.96	-0.61	-1.36
w_t	-0.07	-0.14	-0.02
w_s	-0.14	-0.21	-0.05

Table 2.4: Group-level regression weights and their 95% credible intervals for each condition in the discrimination task. For the mean slope, the inferred weights include the intercept (β_0), the effect of time (β_t), and the effect of the percent of dots seen (β_s). For Weber fractions, the inferred values are analogous, with the intercept (w_0), the effect of time (w_t), and the effect of the percent of dots seen (w_s).

The results of the regression are shown in Table 2.4. Most importantly, the effect of the percent of dots foveated on the mean is positive ($\beta_s = 2.93; CI = [0.24, 4.01]$); and the effect on the Weber fraction is negative ($w_s = -0.14; CI = [-0.21, -0.05]$). The effect of time on the slope is negligible ($\beta_t = -0.01; CI = [-0.07, 0.03]$); but there is still an effect of time on the Weber fraction ($w_t =$

$-0.07; CI = [-0.14, -0.02]$), indicating that the percent of dots foveated probably does not entirely mediate the effect of time on accuracy.

2.4 The mechanics of ANS estimation

We next developed a statistical model that allowed us to use people's behavioral data in order to quantify how different components of visual input contributed to numerical estimates. This model was parameterized in a way that allowed us to test a variety of a priori plausible hypotheses about how ANS accumulation might relate to visual behavior. Primarily, this allowed us to test separable contributions of several behaviorally-measured factors to an estimated quantity μ. The weight of each factor was inferred by the model. Figure 2.7a shows all of these terms in the full equation for μ, with the fit parameters in color and the behaviorally-measured variables on each trial in black.

The model assumed that there were five components that contributed to μ. First, the number of dots foveated (N_{foveal}) and the number of dots not foveated ($N_{peripheral}$), which were each weighted by their corresponding regression parameters (β_{foveal} and $\beta_{peripheral}$). In addition, we tested the contribution of dots that were fixated more than once after first saccading away (N_{double} weighted by the parameter β_{double}). Finally, the proportion of *area* that has been foveated ($A_{foveated}$)—which we measured as percent of the screen within the 5^o window used above—and the area not foveated ($A_{peripheral}$) were allowed as scaling factors.

The full model is given below. The subscripts for each variable denote whether that variable applies to *foveal* (F) or *peripheral* (P) dots; and whether that variable is at the *group*-level (G) or at the *subject*-level (S). All variables that have only one subscript apply to both foveal and peripheral dots, so the subscript denotes only whether it is a group- or subject-level variable. The mean of a participant's estimate, μ, is a function of five quantities: the number of foveated (N_F) and peripheral (N_P) dots; the proportion of screen area foveated (A_F) and peripheral (A_P); and the total number of times all the dots were re-fixated (N_D). Each of these parameters has a corresponding inferred weight in the regression.

$$\beta_{F,G},\ \beta_{P,G},\ \beta_{N,G},\ \beta_{D,G} \sim \mathcal{N}(0,100)$$

$$\sigma_{F,G},\ \sigma_{P,G},\ \sigma_{N,G}, \sigma_{D,G} \sim |\mathcal{N}(0,100)|$$

$$\gamma_{F,G},\ \gamma_{P,G} \sim Beta(1,1)$$

$$\lambda_G \sim Exp(1)$$

$$\beta_{F,S} \sim \mathcal{N}(\beta_{F,G}, \sigma^2_{F,G})$$

$$\beta_{P,S} \sim \mathcal{N}(\beta_{P,G}, \sigma^2_{P,G})$$

$$\beta_{N,S} \sim \mathcal{N}(\beta_{N,G}, \sigma^2_{N,G})$$

$$\beta_{D,S} \sim \mathcal{N}(\beta_{D,G}, \sigma^2_{P,G})$$

$$\gamma_{F,S} \sim Beta(\lambda\gamma_{F,G},\ \lambda(1-\gamma_{F,G}))$$

$$\gamma_{P,S} \sim Beta(\lambda\gamma_{P,G},\ \lambda(1-\gamma_{P,G}))$$

$$\mu_S = (1 + \beta_{D,S} \cdot N_D)\ (\beta_{F,S} \cdot N_F \cdot (1/A_F)^{\gamma_{F,S}}) \quad + \quad \beta_{P,S} \cdot N_F \cdot (1/A_P)^{\gamma_{P,S}}$$

$$\sigma_S = \beta_{N,S} \cdot \mu_S$$

$$\hat{n} \sim \mathcal{N}(\mu_S, \sigma^2_S)$$

To fit this model to behavioral data, we again used a hierarchical Bayesian model which allowed partial pooling of parameters. Examination of the inferred parameters allows us to characterize the mechanisms of ANS estimation in three critical ways: *first*, comparison of $\beta_{peripheral}$ and β_{foveal} will show if the accumulation mechanism relies more, less, or equally on foveal and peripherally observed dots. This, in turn, tells us whether the ANS is primarily parallel or whether foveated dots contribute more to the observed estimates. *Second*, examination of β_{double} will tell us whether participants "double count" dots that are re-foveated ($\beta_{double} \approx 1$) or not ($\beta_{double} \approx 0$). This will answer a basic question about ANS accumulation: is it based on mere retinal input or on a spatially-based picture of the world that is built up across saccades (e.g. Farah et al., 1988). *Third*, do participants re-scale their input by the area they have foveated ($\gamma_{foveated} \approx 1$) in order to correct for their limited visual sample? Or, is estimation a more simple *accumulator* ($\gamma_{foveated} \approx 0$) that does not take into account how much of the scene has been viewed? Note that our formalization does *not* test whether area, density, convex-hull or some other continuous quantity is the *basis* of numerical estimation (Anobile et al., 2014; Odic & Halberda, 2015; Starr et al., 2017). Rather, this tests if the ANS relies preferentially on foveated objects and whether it adjusts for the proportion of screen area that has been foveated.

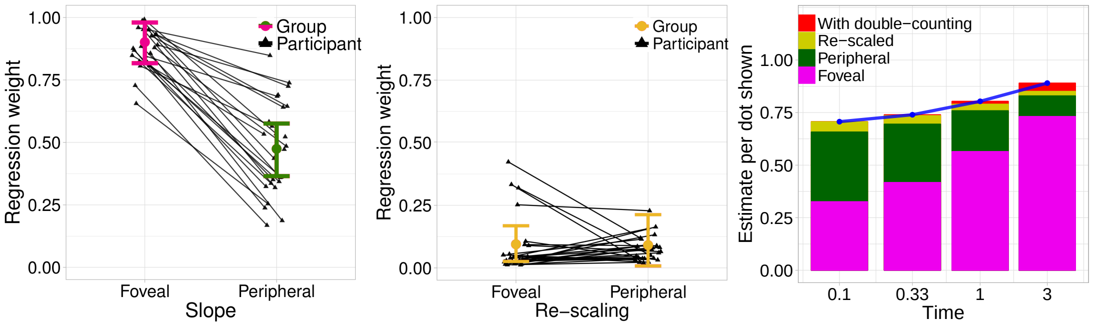
Figure 2.7b shows the inferred group-level and subject-level means for β_{foveal} (x-axis) and $\beta_{peripheral}$ (y-axis). This shows that foveated dots contribute about twice as much as peripheral dots to estimates. Moreover, the value of β_{foveal} is approximately 1, meaning that people veridically count one foveated dot as one more in their estimate[5]. Interestingly, however, the peripheral dots do provide a non-zero contribution, explaining why ANS estimation is possible with very fast presentation times, albeit with a lower precision (Inglis & Gilmore, 2013). Figure 2.7c shows that both γ_{foveal} and $\gamma_{peripheral}$ are near zero, indicating little area re-normalization. This finding supports our primary claim that the estimation is based on accumulation rather than inference using the density of dots observed in part of the scene. Finally, β_{double} is near 0 for all participants, indicating there is almost no effect of seeing the same dot multiple times in the same display. This would happen, for instance, if people build up a mental image of the dot array that is fed to the accumulator.

Figure 2.7d visualizes the relative contribution of each factor to mean estimates (y-axis) across time conditions (x-axis), as inferred by the model. The color of each bar corresponds to the factor it represents in Figure 2.7a. At 0.1 seconds, peripheral and foveated dots contribute roughly equal amounts to estimates, accounting for the significant degree of underestimation given such a short exposure. However, as the exposure time increases foveated dots contribute increasing amounts to the estimate, such that peripheral dots barely play a role in

[5] This does not mean that they were actually counting, as the short display times precluded that. Rather, it means that if all dots in a scene were foveated, estimates would be *un-biased in expectation*, though not error-free.

$$\mu = \overbrace{\beta_{foveal}\cdot(N_{foveal}+\beta_{double}\cdot N_{double})}^{\text{foveal accumulation}} \cdot \underbrace{\left(\frac{1}{A_{foveal}}\right)^{\gamma_{foveal}}}_{\text{re-scaling by foveal area}} + \overbrace{\beta_{peripheral}\cdot N_{peripheral}}^{\text{peripheral accumulation}} \cdot \underbrace{\left(\frac{1}{A_{peripheral}}\right)^{\gamma_{peripheral}}}_{\text{re-scaling by peripheral area}}$$

(a)



(b) (c) (d)

Figure 2.7: Panel (a): The mean estimate, μ, given as a function of the number of dots foveated, N_{foveal}; the number of dots not foveated, $N_{peripheral}$; the percent of screen area foveated, A_{foveal}; the percent of screen area not foveated, $A_{peripheral}$; and the number of dots foveated more than once, N_{double}. Each of these has a corresponding parameter quantifying its contribution to the estimate μ. Panel (b): Parameters β_{foveal} and $\beta_{peripheral}$ capture the foveal and peripheral contribution to the accumulated count. Panel (c): Parameters γ_{foveal} and $\gamma_{peripheral}$ capture the degree to which the accumulated count is normalized by the percent of screen for area foveated (A_{foveal}) or peripheral ($A_{peripheral}$). Panel (d): A visualization of how each factor contributes to μ over time. As exposure time increases, the average proportion of dots foveated increases, leading to differences in the expected contribution of each factor to the mean estimate.

estimation at 3 seconds. Re-scaling and double-counting play almost no role at any amount of time. The important group-level parameters are given in Table 2.5.

2.5 Discussion

ANS estimation is typically thought to operate rapidly and in parallel. There is evidence to support this view. For instance, people can discriminate quantities at above-chance levels given only $16ms$ of exposure (Inglis & Gilmore, 2013). Studies have also demonstrated that reaction times are roughly constant across numerosities in humans and monkeys performing approximate numerical estima-

Var	**Value**	**2.5%**	**97.5%**
β_{foveal}	0.88	0.79	0.96
$\beta_{peripheral}$	0.47	0.36	0.55
γ_{foveal}	0.11	0.04	0.18
$\gamma_{peripheral}$	0.11	0.01	0.23
β_{double}	0.01	0.00	0.05

Table 2.5: Group-level regression weights and their 95% credible intervals for the effect of dots foveated (β_{foveal}) and not foveated ($\beta_{peripheral}$) on the mean estimate; for the effect of re-foveating on the mean (β_{double}) and for the foveal re-scaling factor (γ_{foveal}) and the peripheral re-scaling factor ($\gamma_{peripheral}$).

tion and discrimination tasks Mandler and Shebo, 1982; Nieder et al., 2002. The latency of number-sensitive neurons tends to be independent of numerosity in monkeys performing numerical discrimination tasks as well (Nieder et al., 2006; Nieder et al., 2002). However, the results and analysis we present support an alternative theory: that ANS estimation relies on a serial accumulation mechanism that integrates information—either numerical quantity itself or lower-level visual content—across eye fixations.

Our experiments first replicate two prior behavioral findings: an underestimation bias (Izard & Dehaene, 2008) and a dependence of ANS acuity on time (Inglis & Gilmore, 2013). We then showed that the underestimation bias decreases with time, such that participants estimated higher numbers as the stimulus' duration increased. Such an influence of time is predicted by an accumulation model, but not by prior accounts that attribute underestimation to miscalibration of response scales (Izard & Dehaene, 2008). Finally, we showed that the effect of time is almost entirely mediated by visual fixations, suggesting that time matters *because* with more time, subjects are able to fixate more of the display. Freely fit parameters from our model indicate that foveated points contribute twice as much to a numerical estimate as peripheral ones. This analysis also revealed that the accumulation likely does not adjust for area nor does it double-count re-fixated dots. Together, these results suggest that a primarily foveal, serial accumulation mechanism is at the heart of ANS estimation rather than the rapid, parallel mechanism previously proposed and commonly imagined.

A serial accumulator is similar to ANS models that perform temporal integration of, for instance, sequences of clicks (Meck & Church, 1983), as well as an approximate version of counting logic observed in sequential presentation of quantities to primates (Cantlon et al., 2015). Thus, visual ANS estimation may share resources and processes with non-visual quantity estimation, as experiments on cross-modal matching would suggest (Starkey et al., 1990). Specifically, visual fixations may be a proxy for attention, which would be consistent with the finding that the numerosity of auditory and tactile stimuli are increasingly underestimated as their presentation rate increases (Forsyth & Chapanis, 1958; Lechelt, 1975). Still, it is surprising to see such serial effects in visual displays since vision could in principle support parallel processes (as in, e.g., “pop out” (A. M. Treisman & Gelade, 1980)).

We note that our proposed accumulation mechanism is interestingly different from a statistical *sampler* which has been used, for example, to model incremental changes-of-mind in ANS finger pointing tasks (Alonso-Diaz et al., 2018; Dotan & Dehaene, 2016). Those models assumed that subjects accumulated samples from a Gaussian in order to estimate a mean number of dots, and used this information in an optimal motor plan. However, they did not tie the samples to visual fixations, which appear to be the key mechanism at play in our experiments. The finding that mean estimates increased with greater foveation — even for the two participants who tended to overestimate — is not necessarily expected from a sampling account. In addition, many natural ways of formalizing sampling accounts would adjust for the amount of area sampled to correct the underestimation bias, but our analysis shows that people are probably not adjusting for area.

One limitation of the current work is that our results do not address the specificity of the accumulation mechanism. In particular, our results are consistent with at least two possibilities: either numerical quantities themselves are being integrated across visual fixations; or people build up an increasingly precise image of the visual scene as they saccade, from which numerical information is later extracted. In either case, our results do show that performance in ANS tasks is largely determined by the serial component of this process — in particular how many dots are foveated.

Regardless of the ultimate mechanisms, our results raise an important methodological point for both basic cognitive research on the ANS and applied education research which relies on it. In light of our findings it is difficult to interpret results from studies that compare participants' performance across ANS tasks which use different display sizes or stimulus exposure durations (e.g. Guillaume & Gevers, 2016; Halberda & Feigenson, 2008; Revkin et al., 2008). More broadly, our results suggest that the nearly universal use of ANS tasks to index a "pure" sense of number may be misguided. A full picture of ANS estimation will require integrating aspects of visual cognition such as attention, occular-motor control, and saccade selection in order to understand the cognitive mechanisms that translate visual scenes into abstract numerosities.

3 A shared functional origin of subitizing and estimation

3.1 Introduction

People estimate small numerosities much more rapidly and accurately than large numerosities (Jevons, 1871; Mandler & Shebo, 1982; Revkin et al., 2008), suggesting that we posses two separate representational systems (Dehaene, 2011; Feigenson et al., 2004): a precise small number system, which allows for rapid identification of quantities up to around four objects with little error (Feigenson et al., 2004; Jevons, 1871; Kaufman et al., 1949; Mandler & Shebo, 1982; Revkin et al., 2008); and an imprecise large number system where the standard deviation of estimates increases linearly with numerosity (Burr et al., 2010; Dehaene, 2011; Gallistel & Gelman, 1992; Pica et al., 2004; Xu & Spelke, 2000). This hallmark of large number estimation is known as scalar variability, and can be found in many species across the animal kingdom (Cantlon, 2012; Cantlon & Brannon, 2007; Gallistel, 1990; McComb et al., 1994; Meck & Church, 1983; Piantadosi & Cantlon, 2017; Platt & Johnson, 1971; Uller et al., 2003; Xu & Spelke, 2000; Yang & Chiao, 2016). However, the reason why two qualitatively different patterns of representation would arise in evolution remains obscure. Here we show that the distinct behavior on small and large numerosities is actually expected from a single system which optimally represents quantity under a resource constraint.

Building on recent information-theoretic approaches to visual perception (Brady et al., 2016; Brady & Tenenbaum, 2013; Sims, 2016; Sims et al., 2012) and studies showing the adaptation of perceptual systems to environmental statistics (Geisler, 2011; Olshausen & Field, 1996, 2004; Simoncelli & Olshausen, 2001), we assume that the goal of a numerical processing system is to minimize estimation error. We further assume that there is a time-dependent constraint on the numerical system's ability to process information. Under these assumptions, we present a derivation that recovers the core properties of number psychophysics, including (i) nearly exact representations for small sets (Burr et al., 2010; Choo & Franconeri, 2014; Feigenson et al., 2004; Revkin et al., 2008); (ii) scalar variability in estimation for larger numbers (Dehaene, 2011; Xu & Spelke, 2000); (iii) an underestimation bias (Izard & Dehaene, 2008; Mandler & Shebo, 1982) that diminishes with exposure time (see Chapter 2); (iv) large number estimation acuity that is modulated by time (Inglis & Gilmore, 2013) and display contrast; (v) a subitizing range that is moderated by time (Mandler & Shebo, 1982) and contrast (Melcher & Piazza, 2011); and (vi) roughly normally-shaped response distributions for estimation (Nieder & Dehaene, 2009; Pica et al., 2004). Beyond these general properties, we test the quantitative predictions of the model about how subitizing range, estimation acuity, and response distribution shape should change as a function of the amount of information perceptually available. Our results show a close agreement between human participants and bounded-optimal numerosity perception.

3.2 Model setup and assumptions

The consensus view among cognitive psychologists is that at least two different systems support numerical cognition, giving rise to veridical representations of small numerosities and approximate representations of large numerosities. However, an alternative possibility is that different performance characteristics on large and small numbers is the result of a single psychophysical function which itself reflects a trade-off between the benefits of veridical perception and the cost of processing sensory input. To intuitively understand this alternative, note first that most decisions which depend on numerosity involve only a small number of objects. In fact, the "need probability" (Anderson & Schooler, 1991) of number — how often a numerosity n is encountered and represented — robustly follows a $P(n) \propto 1/n^2$ law.

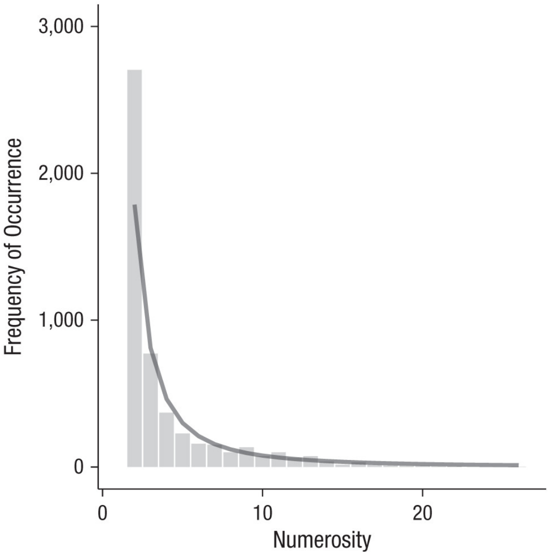
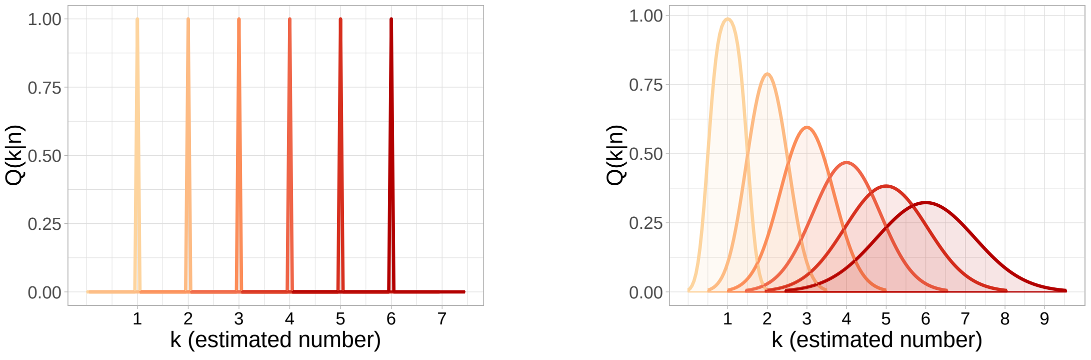


Figure 3.1: The frequency that baboons encountered subgroups of each numerosity in the wild, with a best-fitting power-law of $n^{-1.95}$. Figure from Piantadosi and Cantlon (2017), with baboon troop data collected by Strandburg-Peshkin et al. (2015).

Empirically, the need probability is reflected both in the frequency of number words (Dehaene & Mehler, 1992; Piantadosi, 2016) and how often numerosities are encountered and used for decision making in the wild (Piantadosi & Cantlon, 2017; Strandburg-Peshkin et al., 2015) (see Figure 3.1). This means, for instance, that we should expect that organisms need to represent seven about $1/7^2 = 1/49$th as often as they need to represent one. Efficient representational systems will take advantage of this non-uniformity and be better at representing the more frequently encountered numerosities. Second, universally in information theory, rare events require more bits of information to represent or communicate (Shannon, 1948; Stone, 2018), meaning that high and low numbers will naturally place differing information processing demands in virtue of their different probabilities. Third, any organism will have a finite amount of information processing capability. This is a physical necessity and also a consequence of limited perceptual systems: the amount of internal precision reserved for representations should not in general exceed the amount of information provided by perception (Gallistel, 2018).

Taken together, these facts mean that we should expect different behavior from high and low numbers since they differ in probability; and moreover, we might expect a relatively sharp behavioral discontinuity between them if we assume a hard bound on information processing ability, with low numbers operating below the bound and high numbers operating above (and indeed, what is considered “low” vs. “high” is determined by the information processing bound). We formalize these intuitions by applying standard measures from information theory and analytically computing the optimal representation given an information processing bound. These standard assumptions give rise to the details of number psychophysics as previously determined in behavioral experiments. As we show, the representation that minimizes mean squared error subject to a bounded information capacity transitions from exactness to approximation above and below the capacity bound, even though what is being optimized is a single objective function, itself representing a single “system.”



(a)

(b)

Figure 3.2: Response distributions for two possible forms of Q are shown, with probabilities (y-axis) of estimates (x-axis) shown for numerosities 1-6 (colors). Panel (a) shows the form of a precise estimation system and panel (b) shows the form of a scale variable estimation system.

Consider a psychophysical function Q that maps from an observed quantity to a subjective estimate. Specifically, let $Q(k \mid n)$ give the probability that an observed numerosity n is represented internally with quantity k. Thus, maximally precise, veridical representations have the form,

$$Q(k \mid n) = \begin{cases} 1, & \text{if } k = n \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

In general, any Q that puts high probability on k close to n will have low error rates. Models of large-number estimation typically assume that estimates are drawn from a Gaussian, i.e.,

$$Q(k \mid n) = \mathcal{N}(k \mid n, w \cdot n), \tag{3.2}$$

for some constant w, corresponding to scalar variability (linear increase in the standard deviation of Q with n). Response distributions for numerosities 1-6 under these two possible forms of Q are shown in Figure 3.2.

In principle, many forms of Q are logically possible including, for example, agents who precisely represent numbers in some intermediate range, or who fail completely above a given cardinality. However, we will show that the optimal Q transitions from exact solutions (as in Figure 3.2a) to scalar variability (as in Figure 3.2b) under some basic assumptions. First, we assume that $Q(k \mid n)$ is chosen to minimize the expected squared error between an input n and its representation k,

$$\mathbb{E}\left[(n-k)^2\right] = \sum_n P(n) \sum_k Q(k \mid n) \cdot (n-k)^2. \tag{3.3}$$

Here, $P(n)$ denotes the need probability of number which follows a $P(n) \propto 1/n^2$ power law. Note, however, that this particular power law is not necessary to recover the key properties of the model — other need distributions exhibit similar behavior (see Figure 3.8). Note that here we using P to denote the true need frequency and Q to denote the organism's beliefs.

If an organism had unlimited neural resources at their disposal, then the optimal Q would be given in (4.3) — i.e., they would perfectly encode the numerosity of every set. But neural resources are not unlimited. Just as scientists do not usually attain measurements to more than a few digits of precision, an organism's information processing systems cannot extract arbitrary amounts of information from the world. We can formalize this constraint using a fundamental information-theoretic measure called Kullback-Leibler divergence (KL-divergence) (Cover & Thomas, 2012). KL-divergence intuitively measures how far one distribution differs from another in terms of bits of information. For instance, two overlapping distributions will have small KL-divergence, and two distributions that put most of their probability mass on different outcomes will have high KL-divergence.

For our purposes, KL-divergence quantifies how many bits of information it takes to represent the distribution $Q(\cdot \mid n)$ starting with the distribution $Q(\cdot)$, or equivalently how much information processing an organism must do to change its beliefs from $Q(\cdot)$ to $Q(\cdot \mid n)$[1]. It is natural, therefore, to assume that organisms with limited information processing ability will only be able to form $Q(\cdot \mid n)$ that are boundedly far away from $Q(\,\cdot\,)$ as measured by KL-divergence. In general, this bound should depend on the amount of time that an organism has to process a stimulus since perceptual systems provide a limited bandwidth. Specifically, we assume that perception extracts information linearly in time at rate R until an overall capacity bound B is reached. Using $D_{KL}\left[Q(\,\cdot\,|n) \parallel Q(\,\cdot\,)\right]$ to denote the KL-divergence between $Q(\cdot)$ and any hypothetical $Q(\,\cdot \mid n)$, the definition of KL-divergence therefore yields the bound,

$$D_{KL}\left[Q(\,\cdot\,|n) \parallel Q(\,\cdot\,)\right] = \sum_k Q(k \mid n) \cdot \log \frac{Q(k \mid n)}{Q(k)} \leq \min(B, R \cdot t) \qquad \forall n. \quad (3.4)$$

Since Q is a distribution, we also have a constraints that $\sum_k Q(k \mid n) = 1$ for all n.

To summarize, we are seeking a function $Q(k \mid n)$ which gives the probability that an organism represents n with an internal quantity k. Equation (4.2) defines an objective function saying how accurate any hypothesized Q is in terms of representing the world. Equation (3.4) says how costly any hypothesized Q is in terms of information processing. To apply the method of Lagrange multipliers,

[1] Note here that we are assuming that the organism's prior matches the true frequency, i.e, for all n, $Q(n) = P(n)$.

we encode the objective function and constraints into a single equation,

$$\begin{aligned}\mathcal{F}[Q(k \mid n)] = &\sum_{n=1}^{N} P(n) \sum_{k=1}^{N} Q(k \mid n)(n-k)^2 \\ &+ \sum_{n=1}^{N} \lambda_n \cdot \left(\min(B, R \cdot t) - \sum_{k=1}^{N} Q(k \mid n) \log \frac{Q(k \mid n)}{Q(k)} \right) \\ &+ \sum_{n=1}^{N} \gamma_i \cdot \left(1 - \sum_{k=1}^{N} Q(k \mid n) \right). \end{aligned} \tag{3.5}$$

We then solve for the of the zeroes of the derivative of $\mathcal{F}$ with respect to $Q(k \mid n)$ (i.e. treating "$Q(k \mid n)$" as a separate variable for each n and k). These zeros occur when

$$P(n) \cdot (n-k)^2 + \lambda_n \cdot \left(1 + \log \frac{Q(k \mid n)}{Q(k)} \right) + \gamma_n = 0 \tag{3.6}$$

or

$$Q(k \mid n) \propto Q(k) \cdot \exp\left(-\frac{P(n)}{\lambda_n} \cdot (n-k)^2 \right). \tag{3.7}$$

for λ_n chosen to satisfy the bound in (3.4). This solution has a form of a weighted Gaussian with variance $\lambda_n / 2P(n)$, though in our formulation this distribution is discretized.[2] We solve for λ_n using numerical methods. Specifically, given a bound, rate, and time, we use a Monte-Carlo search algorithm on λ_n to find the maximum $D_{KL}\left[Q(\cdot|n) \parallel P(\cdot)\right]$ that satisfies the constraints. This optimizer is run for 5,000 steps for each λ_n for all numbers up to 100, which was sufficient to find KL-divergences within 0.0001 bits of the bound.

Figure 3.3 shows the value of $Q(\,\cdot \mid n)$ (y-axis) across possible numerical estimates k (x-axis) and the presented numerosity n (color), for various information capacity bounds B (faceted). The derived equation captures properties commonly reported in the literature on the psychophysics of number, including (i) estimation error is almost zero for small sets because they are high probability in $P(n)$ and thus require little information to specify exactly; (ii) large sets exhibit scalar variability since the Gaussian component of (3.7) has a standard deviation proportional to $1/\sqrt{P(n)} \propto n$ for need distribution $P(n) \propto 1/n^2$;

[2] The Euler-Lagrange equations of the calculus of variations can derive an analogous equation for continuous Q.

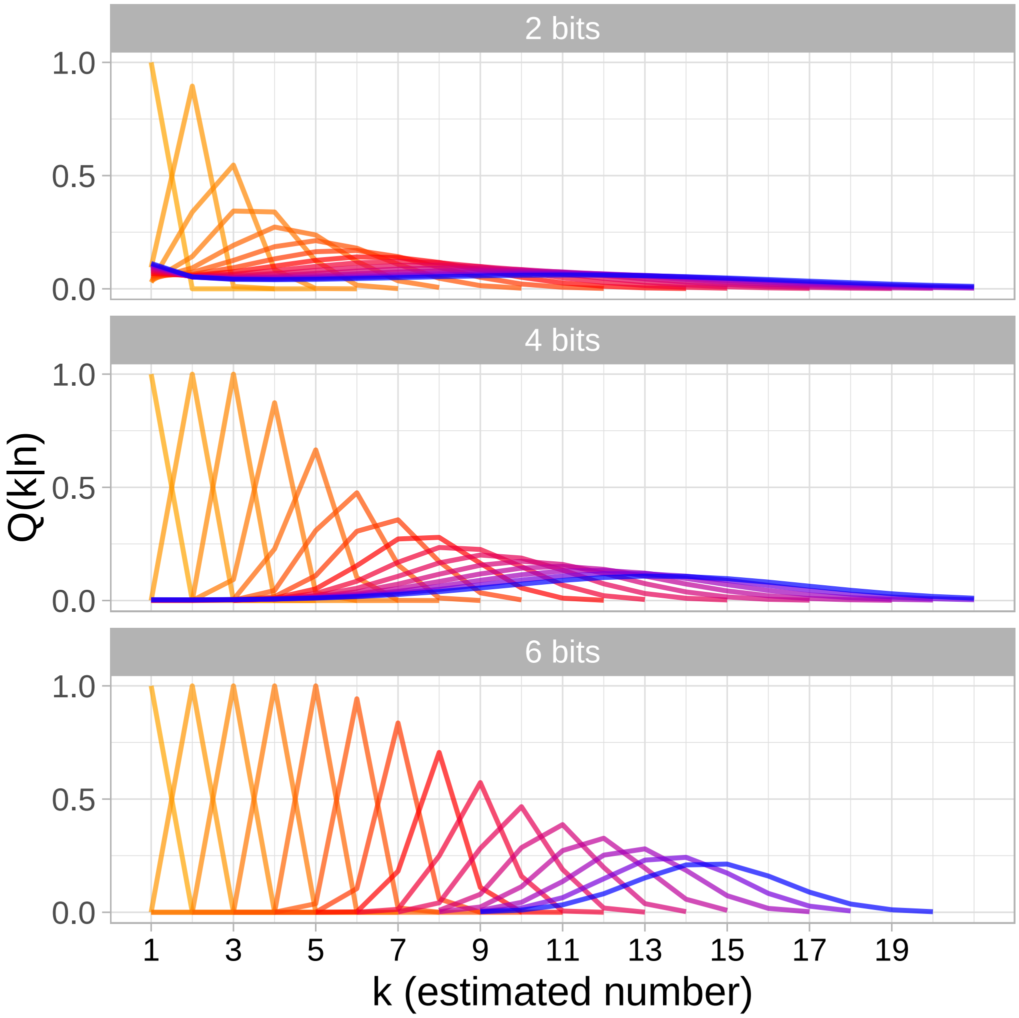


Figure 3.3: The model's posterior probability (y-axis) over numerosities (x-axis), when shown 1 to 15 objects (color). The top panel shows predictions when it has 2 bits of information; the middle panel shows predictions when it has 4 bits; and the bottom panel shows predictions when it has 6 bits.

(iii) an underestimation bias at low information bounds (e.g. 2 bits) due to the skew caused by the $Q(k)$ term; (iv) estimation acuity (the standard deviation of $Q(k \mid n)$) varies with the information bound and thus presentation time; (v) subitizing range varies with the information bound; and (vi) response distributions for large numerosities are roughly normally-distributed, as a result of the form of (3.7).

It is important to emphasize that the roughly Gaussian tuning curves, exact representations for small sets, and scalar variability are not "built in" as representational assumptions, but rather arise solely as a solution to the above optimization problem. The model does not even assume that $Q(k \mid n)$ is centered on n and, in fact, this property only approximately holds. Note, though, that while this model shares many properties with existing psychophysical theories, (3.7) is neither an exact system nor merely an implementation of Weber's law. Instead, this equation recovers the expected behavior of both systems in specific regimes.

3.3 Experiments

The model makes testable predictions about how estimation acuity, subitizing range, and underestimation bias should depend on the amount of information available to participants. We evaluated these predictions against human behavior in four online numerical estimation experiments[3] ($N = 100$ per experiment), which reflect different ways of manipulating available information (variable exposure time versus display contrast) and different ways of controlling non-numerical properties of the stimuli (the average dot size, surface area, or density of the dots). We first varied the presentation time of the dot arrays (Mandler & Shebo, 1982), holding the mean dot size constant (Experiment 1). Varying the exposure time affects the time t in (3.4)—longer presentation times allow more information to be gathered, until the bound B is met. The dots were presented for either $40ms$, $80ms$, $160ms$, $320ms$, or $640ms$. We then ran three replications (Experiments

[3] We pre-registered the experiment and analysis with the Open Science Foundation, which can be found at https://osf.io/svcy5/.

2-4), two of which also varied the exposure time but held the density and area of the dots constant, and one which varied the color contrast of the dots rather than the exposure time as a means of varying the amount of available information. Table 3.1 lists the pairs of these variables and controls that comprise the four experiments.

	Variable	**Controlled**
Experiment 1	Duration	Dot size
Experiment 2	Duration	Surface area
Experiment 3	Duration	Density
Experiment 4	Contrast	Dot size

Table 3.1: Each row shows the manipulated variable (duration/contrast) and the way the stimuli were controlled (size/area/density) for each of the four experiments.

3.4 Experiment 1

Methods

Participants

We recruited 110 US adults from Amazon Mechanical Turk, who were paid $2.50 to participate. We only allowed Mechanical Turk users who had above 95% acceptance rates for their work to participate. Following our pre-registration plan, we removed the 10 subjects in each experiment whose mean absolute error was highest, leaving $N = 100$.

Design

On each trial, an array of between 1-15 dots were presented for either $40ms$, $80ms$, $160ms$, $320ms$, or $640ms$. Participants saw each cardinality in this range twice within each of the five exposure times or contrasts (depending on the experiment). This means that, in total, participants each completed 150 trials total. The order of the stimuli was randomized over number-duration (or number-contrast in Experiment 4) pairs.

Materials

The background was gray (hex value $\#B4B4B4$). The dots were darker gray, with constant Weber contrast of 200%. The experimental window was fixed to 500 x 500 pixels in any browser. However, because this was an online experiment, there were likely a range of monitor sizes and screen resolutions. We had access only to data on any browser size changes in pixels, and so we can only confirm that all browsers allowed participants to see the full experiment, but not the physical size of the display. There was a range of window sizes, from 820 x 524 pixels at the smallest end and 2,560 x 1349 at the largest end. The median width was 1,280 pixels and the median height was 768 pixels. The dots were presented in a 200 pixel radius around the center of the screen.

Procedure

After providing consent and reading the instructions, participants were taken to the main experiment. On every trial in each experiment, a fixation cross was displayed for 750 milliseconds, after which a number of dots were flashed on the screen within a radius of up to 2 inches around the center of the screen. A noise mask was then applied to the screen for 250 milliseconds and subjects were presented with a text box in which they typed their guess of how many dots were displayed. No feedback was given. Participants were given the opportunity to take a break every 10 trials.

Model fitting

We ran a hierarchical regression to infer participants' information rates R, bounds B, and guessing (inattention) rates G based on their performance. We assume that on some proportion G of trials, participants guessed randomly, here meaning that they sampled from their prior. So each participant's probability of making an estimate k given a number n was modeled as $(1 - G) \cdot Q(k \mid n) + G \cdot P(k)$.

Instead of fitting G directly, we fit a transformed variable G', where $G = 1/(1 + exp(-G'))$. Parameter estimates were partially pooled (A. Gelman & Hill, 2006), meaning inference was run jointly on the subject- and group-level rates, bounds, and guessing rates. We note also that while the model fitting here assumes a prior $Q(k) \propto 1/k^{\alpha}$ for $\alpha = 2$, the model's qualitative behavior is robust to changes in α and can be fit with subject effects to yield similar results (see Section 3.7).

We used flat priors[4] for the group-level mean, μ_g and variance σ_g^2 of R, B, and G'. Bounds and rates for individual subjects were then drawn from $Gaussian(\mu_g, \sigma_g^2)$. We used the Metropolis-Hastings algorithm to jointly infer posterior distributions over each parameter at the group- and subject-level. Parameter estimates were averaged over runs from two chains, each with 50,000 steps and 5,000 steps of burn-in.

Results

Figure 3.4a shows the inferred subject bit capacity at different times according to (3.4), in the experiment with variable duration and size-controlled stimuli. The model infers that people's information accumulation saturates at around 100-150ms. The maximum amount of information most people take in is just over 4 bits, which is close to previous, independent estimates of information capacity on similar tasks (Sims, 2016; Verghese & Pelli, 1992). Figure 3.4b shows the probability the model assigns to each possible response k as n varies versus the proportion of humans who gave that response. This overall summary illustrates that over 92% of the variability in human average responses are explained by the model.

Figure 3.5a-d shows model posterior predictive fits including subject effects (left) and human data (right) for absolute estimation error (top), mean estimates (middle), and the shape of the response distributions (bottom). Critically, Figure 3.5a shows that the model predicts that the error of $Q(\,\cdot \mid n)$ should also vary with presentation time, an effect found in human behavior in Figure 3.5b.

[4] This choice has little influence on the results. To ensure robustness, several other priors were tested, including a Normal distribution for the means and an inverse gamma for the variances, using a wide-ranging set of parameters. There was almost no effect on the posterior, which was expected given the large amount of data.

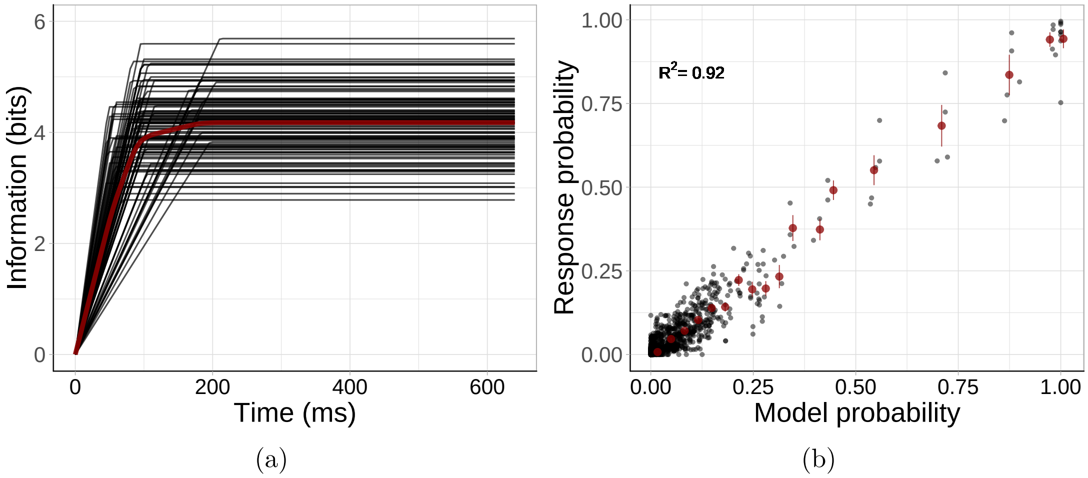
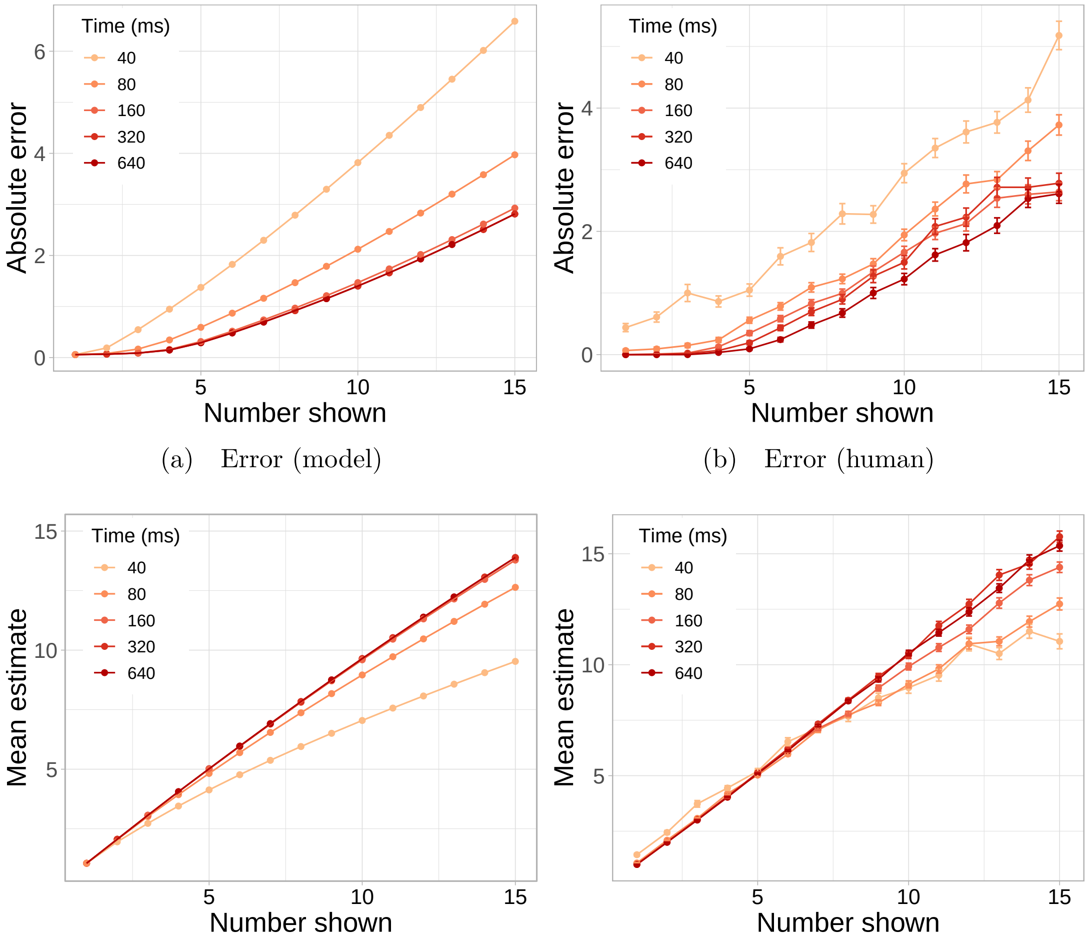


Figure 3.4: The left panel (a) shows the amount of information (y-axis) available to each participants (black) and on average (red) across presentation times (x-axis), as inferred by the model. The right panel (b) shows correlation between model probabilities of responses (x-axis) and average human responses (y-axis), where each point represents responses to one number/estimate pair.

Near-zero estimation error is found for low numbers—"subitizing"—in both the model and human subjects at long display times. However, error increases even for small quantities at short presentation times both for the model and for human subjects, reverting instead to scalar variability (linear relationship) when the amount of available information is low. Intuitively this is because less information in the input reduces the allowable KL-divergence in (3.4), which forces the model to begin to approximate lower numerosities — even those in the typical subitizing range. Thus, in both people and the model, subitizing is not driven by a fixed object capacity, but rather flexibly responds to the amount of information that is visually available.

Figure 3.5c shows that the model predicts an underestimation bias in mean responses that diminishes at longer exposures, which is also found in human behavior in Figure 3.5d. Note that even at the shortest durations, estimates are not random — mean estimates still monotonically increase with the number shown in both the model and people. As predicted by the model, participants'



(a) Error (model)

(b) Error (human)

(c) Mean estimates (model).

(d) Mean estimates (human).

Figure 3.5: The top two panels (a and b) show model predictions (a) and human data (b) for the absolute error of estimates (y-axis) as a function of the number displayed (x-axis). The middle two panels (c and d) show model predictions (c) and human data (d) for mean estimates (y-axis) as a function of the number of dots displayed (x-axis) and time (color) for the experiment with variable duration and size-controlled stimuli. All errorbars represent standard error.

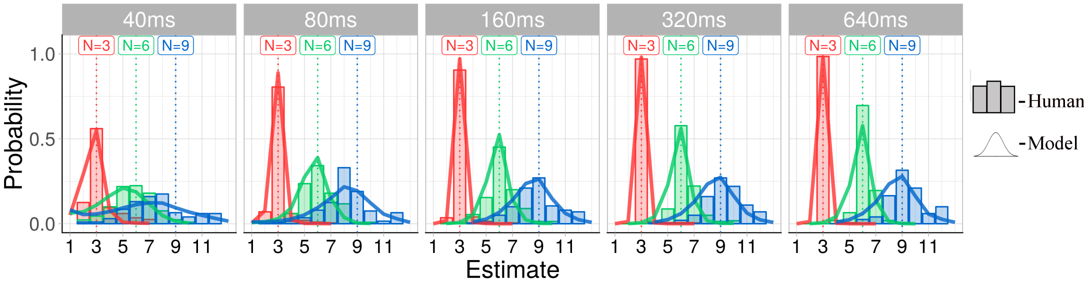
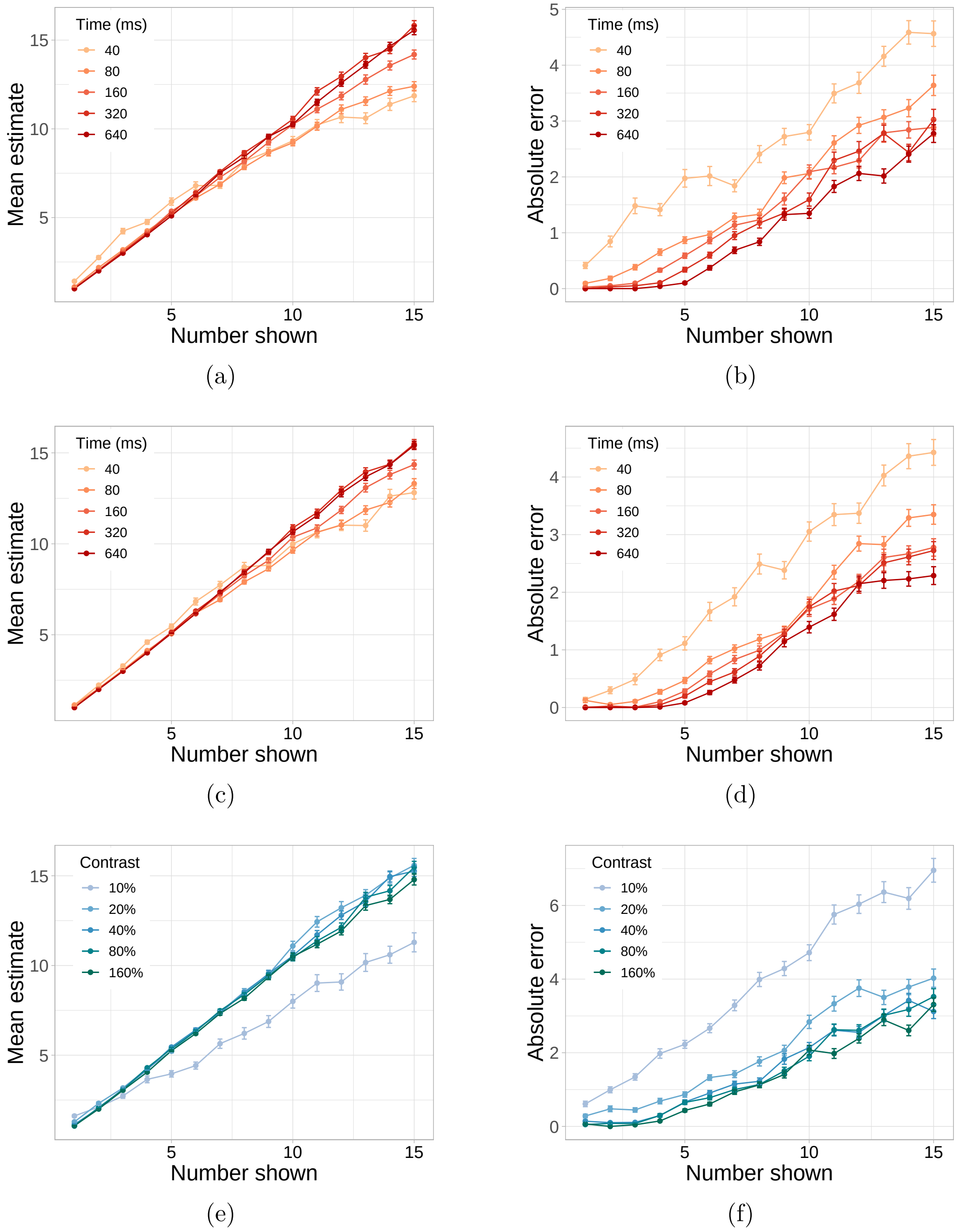


Figure 3.6: The probability (y-axis) of numeric responses (x-axis) over presentation times (faceted) for N=3, N=6, and N=9. Bars are shown for the human data and lines are shown for the model predictions.

mean estimates become increasingly unbiased at longer durations, such that the average estimate converges on the veridical number after around 160ms. This plot shows that the model is less gradiently sensitive to time than people, and this is likely due to our assumption of strictly linear accumulation in (3.4). Figure 3.6 shows the shape of the model (line) and human (bar) response distributions for $N = 3, 6, 9$. These make it clear that it is not just the means and standard deviations which match closely, but rather the shape of the entire distribution derived in Equation 3.7.

3.5 Replications

To ensure that participants are actually using number rather than a correlated dimension, we had two groups of subjects perform the same task as above but with either the total surface area or the average density of the dots controlled. Second, because other manipulations of information should have similar effects as time, we varied the display contrast (Melcher & Piazza, 2011) of the dot arrays, which affects the rate R at which information about numerosity could be extracted from the scene. In the variable-contrast experiment, the color of the dots varied between the background (gray) and pitch black, by Weber contrasts of 10%, 20%, 40%, 80%, and 160%, at a constant presentation time of 200ms. All other aspects of the design, procedure, materials, and number of participants



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.7: Mean estimates and absolute error of estimates as a function of number shown in the three replication experiments (compare to the model in Figure 2a,c).

were identical to Experiment 1.

Results

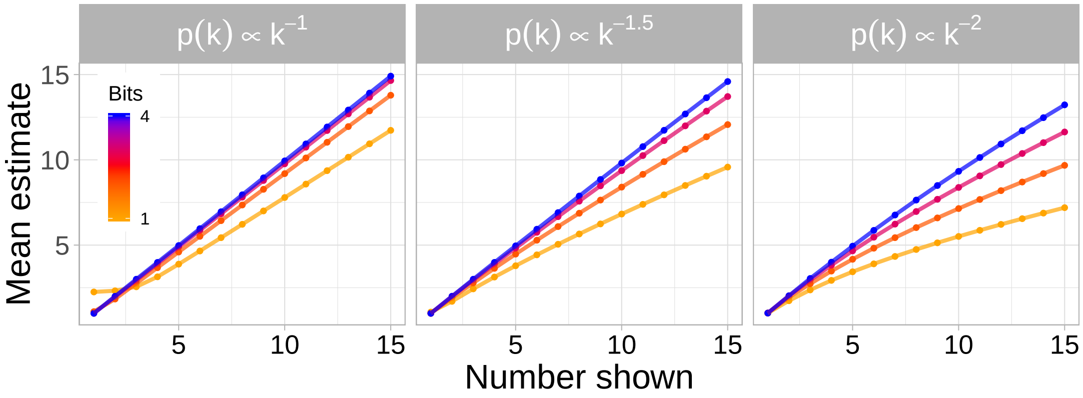
The inferred group-level rates and bounds were similar to first experiment, similarly corresponding to an average subitizing range of about 4. As shown in the left panels of Figure 3.7 (a,c,e), participants tended to underestimate larger numbers for short exposure times and low levels of contrast, matching predictions of the model (e.g. Figure 3.5a). Likewise, the panels on the right of Figure 3.7 (b,d,f) show that in each experiment, absolute error is scale-variable at low levels of information and then becomes precise for small numbers at higher levels of information.

3.6 Model comparisons

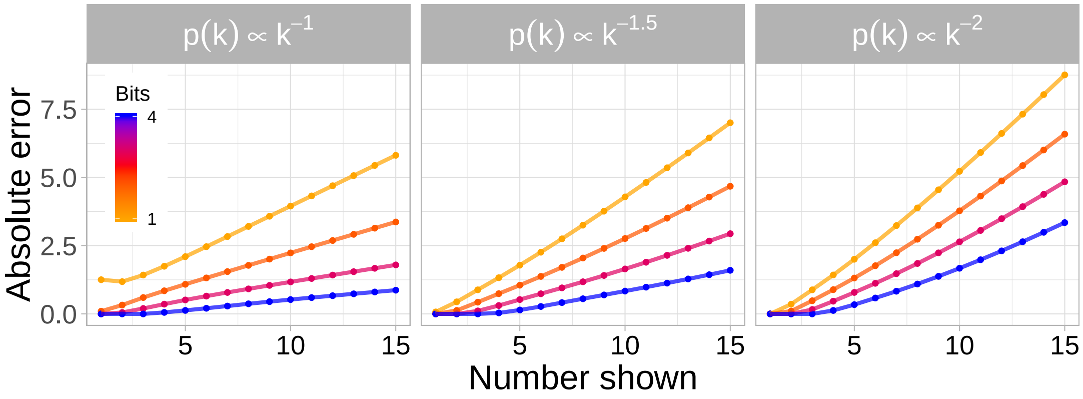
One popular alternative to a two-systems theory is that number representations are scale variable even throughout the "subitizing range" (Barnard et al., 2013; Gallistel & Gelman, 1991, 1992; Piazza et al., 2011; Trick & Pylyshyn, 1994): the error in this range under scalar variability may be small enough to yield essentially perfect accuracy. We first compared the performance of the model to an implementation of this theory, which assumes that a subject's estimate of a number n is drawn from $Gaussian(n, w \cdot n)$, where w is a constant fit for each subject. To compare models, we use the Akaike Information Criterion (AIC), which gives better (lower) scores for models that fit data well and have few free parameters. Using maximum likelihood fits for each model, the difference in AIC scores was over $3,000$ in each experiment (duration/size difference: $3,076$; duration/density difference: $4,902$; duration/area difference: $3,454$; contrast/size condition difference: $4,014$) in support of our model.

Second, we fit a Weber model with time or contrast effects to each experiment, assuming estimates of a number n are drawn from a $Gaussian(n, e^{w_0+w_t \cdot t}) \cdot n)$. With this model, there were AIC differences of over 750 in favor of our model (duration/size difference: 768; duration/density difference: $2,838$; duration/area difference: $1,084$; contrast/size difference: 924). Together, these results indicate that human behavior cannot be explained by assuming only scalar variability, nor with ad hoc modifications to scalar variability that allow acuity to vary with time and contrast.

3.7 Inferring prior distributions



(a)



(b)

Figure 3.8: Each facet shows model predictions about mean estimates (a) and absolute errors (b) for different prior distributions, when different amounts of information is available (color). The facets on the left show predictions for $Q(k) \propto 1/k$; the facets in the middle show predictions for $Q(k) \propto 1/k^{1.5}$ and the facets on the right show predictions for $Q(k) \propto 1/k^2$.

While the primary analyses assume that people's prior distribution over numbers reflects the natural frequency of number words, this is likely an over-simplication and is not necessary to recover the model's primary qualitative properties. In fact, because we sampled numbers uniformly in the range 1-15 in the experiments, people's number perception systems might have adapted to the statistics of the experiment. So, following the pre-registration plan, we re-ran the model to jointly participants' priors along with the other variables (rate, bound, and guessing rate). More specifically, we assume priors have the form $Q(k) \propto 1/k^{\alpha}$, where α is a random variable. Figure 3.8 shows predictions about mean estimates (top) and absolute errors (bottom) for different α (1, 1.5, and 2). In every case, the model predicts underestimation of larger numbers when less information is available, which diminishes with increasing information. The model also predicts roughly scale variable error across numbers given low information and, given more information, zero error for small quantities and less error (but still scale variable) for large quantities. This illustrates that the key effects of our model are robust to α.

By inferring α, we found that subjects' value tended to be closer to 1 than 2 in each experiment. In the experiment with variable duration and with size-controlled stimuli, the MAP inferred group-mean α was 1.30. In the experiment with variable duration and density-controlled stimuli, the inferred group-mean α was 1.37. In the experiment with variable duration and area-controlled stimuli, the inferred group-mean α was 1.29. In the experiment with variable contrast condition and size-controlled stimuli, the inferred group-mean α was 1.11. Moreover, the model with inferred priors provides significantly better fits to the data in each experiment than the model with the fixed $1/k^2$ prior (AIC differences each over 1000). Generally, this shows that the results do not hinge critically on assuming a particular value of the exponent in the need distribution $Q(k)$.

3.8 Discussion

Empirical studies dating back more than a century have charted many robust characteristics of numerosity perception in humans and other animals. However, most of these properties are treated as separate phenomena without a common explanation. For instance, the finding that people are able to exactly represent small sets (Burr et al., 2010; Choo & Franconeri, 2014; Feigenson et al., 2004; Jevons, 1871; Revkin et al., 2008) and show scalar variability in estimation for larger sets (Dehaene, 2011; Xu & Spelke, 2000) has been explained in terms of two different representational systems (Dehaene, 2011; Feigenson et al., 2004). The tendency to underestimate larger quantities (Jevons, 1871; Mandler & Shebo, 1982) has been explained in terms of a miscalibration of response scales (Izard & Dehaene, 2008). The sensitivity of numerical acuity to display time (Cheyette & Piantadosi, 2019; Inglis & Gilmore, 2013; Mandler & Shebo, 1982) seemingly requires ad hoc modifications to processing theories. Our derivation, however, shows that these phenomena—underestimation, distinctive behavior on large and small sets, sensitivity to timing and contrast, and even the shape of response distributions—can be explained as natural consequences of optimal representation under a resource constraint.

The sensitivity of numerosity judgments to certain non-numeric properties of the visual scene, such as object spacing (Atkinson et al., 1976) and arrangement (Ginsburg, 1976; Mandler & Shebo, 1982), also fit naturally in this framework if they are considered as manipulations of information in the visual scene. For instance, regularly spaced objects appear more numerous than randomly spaced objects (Ginsburg, 1976). Likewise, objects with similar orientations appear more numerous than objects with randomly distributed orientations (DeWind et al., 2020). These effects are predicted under our model since regularities should decrease the information processing demands on the visual system.

An information-theoretic approach connects number psychophysics to the broader emerging picture of visual working memory. Contrary to a once dominant conception of visual working memory as discrete and "slot-like" (Awh et al., 2007; Luck & Vogel, 1997), recent behavioral and neural evidence suggests instead that visual memory flexibly allocates limited resources in a continuous manner (Brady et al., 2016; Brady & Tenenbaum, 2013; Keshvari et al., 2013; Ma et al., 2014; Van den Berg et al., 2012). Like such accounts, our model assumes that bits of information are the common currency that limit numerosity perception (Gallistel, 2018). While others have hypothesized that subitizing is driven by a capacity limit (Trick & Pylyshyn, 1994), no work has formally derived how such a limit gives rise to the psychophysics of both subitizing and estimation.

Prior accounts of numerosity perception have also not explained why infants (Starr et al., 2013a), some primates (Barnard et al., 2013; Piantadosi & Cantlon, 2017), and other animals (Agrillo et al., 2014; Petrazzini et al., n.d.), may have a smaller subitizing range than human adults. A two-systems theory would require a separate small number system to suddenly arise either in evolution or development. However, the model we describe suggests a simple alternative: infants and many animals may have a lower visual memory capacity (Elmore et al., 2011), leading the model to predict numerical approximation and scalar variability even throughout the small number range. Conversely, chimpanzees may have a subitizing range up to 4 or 5 (Tomonaga & Matsuzawa, 2002), exceeding that of humans, because they have a greater visual memory capacity (Inoue & Matsuzawa, 2007). Similarly, an information-theoretic perspective predicts that the point at which a person transitions from subitizing to estimation should depend their visual memory capacity, which it does (Green & Bavelier, 2003, 2006; Piazza et al., 2011).

More generally, this work highlights that behavioral discontinuities are not always good markers of distinct systems. Discontinuities often arise in biology when single systems face constraints—for instance, when an animal's gait varies discontinuously with its speed (Alexander, 1984) or a neuron spikes when its input exceeds a threshold. Our results illustrate that optimization of a single objective function may in fact show starkly different behavior above and below a capacity bound, thus providing a resource-rational (Griffiths et al., 2015) account of qualitatively different patterns of numerical perception.

In sum, the theory we present relies on combining an a priori biological consideration (bounded informational capacity) with an environmental input distribution $P(n)$ and analytically computing the optimal internal representation. The resulting representational system replicates all of the standard properties of number psychophysics and explains them with a simple, resource-rational model. Our experiment has also shown that human numerical cognition quantitatively tracks this bounded optimal solution as the amount of information available varies, a fact not explainable in existing psychophysical theories. Together, these results suggest that the core properties of numerical cognition arose as efficient solutions to the problem of representing the world with finite cognitive and neural resources.

4 Limitations in spatial memory explain number psychophysics

4.1 Introduction

A key unresolved question is whether the behavioral patterns found in the domain of number result from numerical processing itself or from some of the perceptual processes that feed into numerical perception. In the first case, people may posses a "number system" that itself is the origin of phenomena seen in behavioral tasks involving number, such as Weber's law and underestimation. For instance, the noise and bias observed in numerical estimation might arise from a sampling process in which numerical information is extracted from visual representations, rather than from noise inherent to visual representations themselves (Dehaene & Changeux, 1993; Heng et al., 2020; Woodford, 2020). Alternatively, such phenomena may emerge as a consequence of more general visual processes which precede numerical estimation and indeed feed into it (Anobile et al., 2020; Stoianov & Zorzi, 2012; Testolin, Dolfi, et al., 2020; Trick & Enns, 1997; Zorzi & Testolin, 2018). Under this latter hypothesis, the psychophysics of estimation in vision could result from constraints inherent to visuospatial memory, and then we would expect that people's behavior in visual tasks not involving number to show equivalent hallmarks to those seen in estimation.

The model presented in Chapter 3 demonstrates that principles of efficient representations can explain many features of number psychophysics, including the discontinuity from exactness to scalar variability. The key idea there was that an efficient encoding of number, using at maximum some number of bits of information, will prioritize representations of small numbers at the expense of large numbers because people tend to need to represent small numbers more frequently (Dehaene & Mehler, 1992; Piantadosi & Cantlon, 2017). That work therefore derived exactness for small numbers (e.g. subitizing) and approximation for large numbers by solving a single, unifying optimization. However, the model did not explain the key step of how numerosities are actually computed from visual input, and therefore does not explain *where* noise in representations of numerosities comes from. Furthermore, that model made the unrealistic assumption that, all else being equal, small and large numerosities are equally easy to perceive—their differing behavioral signatures being solely a matter of frequency of use.

Our goal in this chapter is to formalize and test the relationship between number perception and visuospatial memory in order to determine whether the properties observed in the number literature—including subitizing, approximation, and sensitivity to presentation time—result from more general mechanisms of the visual system. If a model of basic visual processing fit to a non-numerical task still shows the hallmarks observed in number psychophysics, this would suggest that features of number perception should really be considered artefacts of basic vision rather than number itself. Conversely, if features of numerical perception are not latent in a non-numerical visual memory task, they have to be the result of specifically numerical processes.

We develop a computational model of bandwidth-limited scene memory which forms beliefs about where individual objects exist in space; these beliefs can then be straightforwardly converted into beliefs about the number of objects in that scene. This approach builds on recent neural network models that exhibit approximate numerical representations as a consequence of imperfectly representing a scene (Kim et al., 2021; Stoianov & Zorzi, 2012; Testolin, Dolfi, et al., 2020; Testolin, Zou, et al., 2020; Zorzi & Testolin, 2018). We show that even though the model is explicitly optimized only to detect and remember the presence of objects in various locations, the resulting probability distributions over numerosities closely match known properties of number psychophysics, including both subitizing and Weber's law. Notably, although the model represents a probability distribution over discrete individuals, it behaves like an "analog magnitude system" (Carey, 2009; Gebuis & Reynvoet, 2012a; Lourenco & Longo, 2010) when its information capacity is exceeded.

The model makes predictions about the psychophysics of spatial memory and numerosity perception, and how they should co-vary over time. Specifically, the model predicts that people's ability to remember the locations of objects in space will be near-perfect for small groups — since smaller groups of objects are less informationally demanding to represent — but that spatial memory will degrade proportionally with the number of objects in the scene for larger groups. The model additionally predicts that at shorter exposure times, people will become increasingly unable to precisely remember the locations of even small groups of objects. We can likewise derive predictions about the psychophysics of numerical estimation as a function of cardinality and exposure time, when the output of this bandwidth limited system is used as input for numerosity estimation.

To test the model's predictions about both spatial memory and numerical perception, we ran two pre-registered experiments[1]: a change-localization task to probe participants' memory for the locations of objects (Experiment 1); and a numerical estimation task (Experiment 2). In both experiments, we manipulated the amount of information available to participants by varying the exposure time of the presented objects. We find that participants' ability to remember the locations of objects — both for different exposure times and for different numbers of objects present — predicts the observed psychophysics of number under analogous conditions. In other words, the patterns of bias and noise observed in numerical estimation precisely matches the amount of uncertainty observed in visual representations of scenes, indicating that the psychophysics of number are governed by a domain general, rather than number-specific, information bottleneck.

[1] The pre-registration of the model and both experiments can be found at https://osf.io/vgm65/.

4.2 Model

The model aims to capture how an idealized, information-limited perceptual system would perform if its only aim was to accurately store the presence or absence of objects in various locations. Although this formalizes the idea of object memory—not specifically numerical estimation—its output nonetheless yields psychophysical properties seen in number. For a given, observed scene containing objects s, we will consider the probability distribution $Q(s' \mid s)$, giving the system's belief that s' was observed instead of s. We analytically derive an optimal form of Q, by specifying three components: (i) a prior distribution representing how likely the model is to encounter a given scene *a priori*, (ii) a loss function representing how good or bad a given representation of the scene is, and (iii) an information capacity bound, representing the maximum allowable information processing. These three elements define a constrained optimization problem, which can be solved to determine the optimal psychophysical distribution $Q(\cdot \mid s)$, corresponding to the optimal perceptual system. This process is not identical with, but is somewhat analogous to, Bayesian inference that begins with a prior distribution and combines it with evidence to produce a "posterior" distribution; the key difference is that the shape of the "posterior" $Q(\cdot \mid s)$ is not derived from Bayes rule, but rather from minimizing the loss function (ii) subject to an information bound (iii).

Figure 4.1 illustrates the basic setup, assuming for the sake of clarity that there are only 4 possible object locations (or pixels). When a person sees a particular scene, they encode a probability distribution over each possible arrangement of objects, which is a weighted combination of a prior for small numbers and how well the representation matches their observation (akin to a likelihood). This probability distribution in turn can be converted into a probability distribution over numerosities by summing the probabilities of each scene with a given number of objects. One key simplifying assumption we make in modeling this setup is that spatial memory encodes the presence or absence of objects in various locations as a discrete matrix. In other words, we assume that visuospatial memory represents a matrix with M black and white pixels to specify where there are objects (and

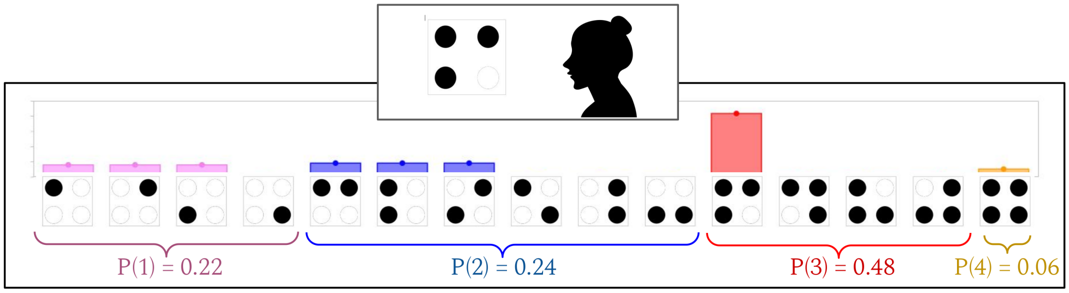


Figure 4.1: This figure conceptually illustrates how the model works, simplifying it to assume that there are only 4 pixels for clarity. In this example, a person sees a scene with 3 objects, which is represented as a probability distribution over all possibilities of what she saw. Possible arrangements of objects are grouped by numerosity, shown as different colors. To get the probability of a numerosity k, the model simply sums the probability of all possible scenes with numerosity k, highlighted at the bottom.

where there aren't).

We further assume a prior on binary matrices where the number of 1's in a matrix matches the naturalistic frequency of a given number. Specifically, the naturalistic frequency of a number n follows a $\frac{1}{n^2}$ law, where n represents cardinality (Dehaene & Mehler, 1992; Piantadosi & Cantlon, 2017). There are $\binom{M}{n}$ matrices with cardinality n, so a given matrix s with cardinality n has prior $P(s) \propto 1/\left(n^2 \cdot \binom{M}{n}\right)$. We emphasize that this choice of prior does *not* play a determining role in governing the model's psychophysics, unlike in the model presented in the previous chapter. It is a reasonable choice, however, as a $\frac{1}{n^2}$ need frequency for number occurs in naturalistic settings (Piantadosi & Cantlon, 2017) as well as in word counts (Dehaene & Mehler, 1992).

Derivation

Given a matrix s, the goal is to represent s with as high fidelity as possible, remembering whether an object was present at each row i and column j, s_{ij}. We define a loss function $L(s, s')$ specifying how closely a matrix s' matches s, or how costly it would be to represent s with s'. We will assume that the loss function $\mathcal{L}(s, s')$ is proportional to some (perhaps unequal) combination of the proportion of false negatives, $P(s'_{ij} = 0 \mid s_{ij} = 1)$, and the proportion of false positives, $P(s'_{ij} = 1 \mid s_{ij} = 0)$:

$$\begin{aligned} \mathcal{L}(s, s') = \quad & \alpha \cdot P(s'_{ij} = 0 \mid s_{ij} = 1) \quad + \\ & (1 - \alpha) \cdot P(s'_{ij} = 1 \mid s_{ij} = 0), \end{aligned} \tag{4.1}$$

with α as a weighting parameter, where $0 \leq \alpha \leq 1$. The reason we separate the contribution of false negatives and false positives here is simply that it is natural to think that the visual system might care about one more than the other. There are, of course, other plausible loss functions, which in fact give qualitatively similar results (see SI) — though we note that the form of this loss function was pre-registered.

Given a loss function and prior, we now seek a function $Q(\,\cdot \mid s)$ that minimizes the expected loss between possible inputs s and representations s', corresponding to the "best" representation possible. If the set of possible scenes is S, the expected loss is

$$\mathbb{E}\left[\mathcal{L}(s, s')\right] = \sum_{s \in S} P(s) \sum_{s' \in S} Q(s' \mid s) \cdot \mathcal{L}(s, s'). \tag{4.2}$$

Unconstrained, the function $Q(\,\cdot \mid s)$ that minimizes the expected loss would simply correctly encode the scene,

$$Q(s' \mid s) = \begin{cases} 1, & \text{if } s' = s \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

However, cognitive systems are constrained by the amount of information they can process over a given span of time. We incorporate this constraint into the model as a bound on the maximum allowable Kullback-Leibler divergence (KL-divergence) between the prior distribution $P(\cdot)$ and resultant distribution $Q(\,\cdot \mid s)$ over displays. The KL-divergence here represents the amount of information in bits needed to represent the resultant distribution $Q(\,\cdot \mid s)$ starting with the distribution $P(\cdot)$, which is equivalent to the total amount of information processing required. Given a information bound B we then have the constraint on the KL-divergence from $P(\,\cdot\,)$ to $Q(\,\cdot \mid s)$, often notated $D_{KL}\left[Q(\,\cdot \mid s) \parallel P(\,\cdot\,)\right]$,

$$\sum_{s' \in R} Q(s' \mid s) \cdot \log \frac{Q(s' \mid s)}{P(s')} \leq B \quad \forall s \in R. \tag{4.4}$$

To apply the method of Lagrange multipliers, we encode the objective function and constraints into a single equation,

$$\begin{aligned}\mathcal{F}[Q(s' \mid s)] = \sum_{s \in R} P(s) \sum_{s' \in R} Q(s' \mid s) \cdot L(s, s') \\ + \sum_{s \in R} \lambda_s \cdot \left(B - \sum_{s' \in R} Q(s' \mid s) \log \frac{Q(s' \mid s)}{P(s')} \right) \\ + \sum_{s \in R} \gamma_s \cdot \left(1 - \sum_{s' \in R} Q(s' \mid s) \right) .\end{aligned}$$

We then solve for the of the zeroes of the derivative of $\mathcal{F}$ with respect to $Q(s' \mid s)$ (i.e. treating "$Q(s' \mid s)$" as a separate variable for each s and s'). These zeros occur when

$$P(s) \cdot L(s, s') + \lambda_s \cdot \left(1 + \log \frac{Q(s' \mid s)}{P(s')} \right) + \gamma_s = 0 \tag{4.5}$$

or

$$Q(s' \mid s) \propto P(s') \cdot \exp\left(-\frac{P(s)}{\lambda_s} \cdot L(s, s') \right) . \tag{4.6}$$

Here, λ_s is chosen to satisfy the bound in (4.4).

Finding λ_s using numerical approximation

We solve for λ_s using numerical methods. Specifically, given a bound, we use gradient descent to find λ_s that allows the maximum $D_{KL}\left[Q(\,\cdot\,|s) \parallel P(\,\cdot\,)\right]$ that satisfies the constraint. This optimizer was run for 5,000 steps for each λ_s, which is sufficient to find KL-divergences within a millionth of a bit of the bound.

One complication is that the representational space in our experiments was very large — there are 49 grid cells so there are 2^{49} possible grid states ($\approx 10^{15}$). Memory and runtime constraints therefore make it impossible to represent the prior and posterior of each possible grid state independently. Luckily, for every scene s, there are many representations that are "equivalent" in that they have equal prior probabilities and losses. For a given representation s', we define the loss as a function of the number (or proportion) of false positives and false negatives between s and s'. To get the number of false negatives $f_n(s' \mid s)$ and false positives $f_p(s' \mid s)$, we can write,

$$f_n(s' \mid s) = \sum_i \sum_j s_{ij} \cdot (1 - s'_{ij}), \tag{4.7}$$

and

$$f_p(s' \mid s) = \sum_i \sum_j (1 - s_{ij}) \cdot s'_{ij}, \tag{4.8}$$

where i and j are the rows and columns of the grid.

We can count the number of representations that have $f_n(\cdot \mid s) = r_n$ and $f_p(\cdot \mid s) = r_p$. This is the product of all the ways to make $n - r_p$ true positives in given that s in n on cells and $k - r_n$ true negatives in $M - n$ off cells, where n is again the cardinality of the scene s, k is the cardinality of the representation s', and M is the total number of grid cells. So we therefore can write the total number of equivalent states S as,

$$S = \binom{n}{n - r_p}\binom{k - r_n}{M - n}. \tag{4.9}$$

In this way, we can collapse the representational space into only individual instances of each equivalence class and when calculating the KL-divergence multiply each term by S.

Modeling change-localization

We assume that subjects choose in the change-localization task proportionally to their belief that a cell has changed. In disappearing trials, subjects are only allowed to respond with a zero cells, and in this case the probability that the cell changed is the belief that the cell was initially 1. This means that the probability of responding ij out of only the other zeros is,

$$P(\text{choose } ij) \propto \sum_{s' \in R} Q(s' \mid s) \cdot \mathbf{1}_{s'_{ij}=1}. \tag{4.10}$$

To compute the probability that subjects answer accurately, $P(\text{choose } ij)$ is computed for the correct disappearing cell relative to all of the other zero cells in the final display. Appearing trials are defined analogously.

Modeling numerical estimation

To compute the probability the model believes that the scene contained k' objects, we can sum across the model's posterior for all scenes containing k' objects. More formally,

$$p(k = k' \mid s) = \sum_{s' \in R} Q(s' \mid s) \cdot \mathbf{1}_{|s'|=k'} \tag{4.11}$$

where $|s'|$ represents the cardinality of representation s' (i.e. the number of objects in s'), and $\mathbf{1}_{x=y}$ is 1 when $x = y$ and 0 otherwise.

To illustrate the model, we generated predictions assuming a 7x7 grid of possible object locations, as will be used in the eventual experiments. We first simulated the model's predicted performance on a change localization task in which the model has to guess which location on the grid has changed — with an object either appearing or disappearing — between two subsequent presentations (see Experiment 1). Figure 4.2 shows the model's predicted accuracy (y-axis) on this task as a function of the number of objects in the scene (x-axis), at different information bounds (color). At each information bound, performance decreases as a function of the number of objects, reflecting both the decreasing prior over numerosities and the fact that there are more ways to arrange more numerous sets in the range shown. Also apparent is that as the information bound increases, the model saturates in performance for small numbers, meaning it can veridically recall the scene it viewed when there are a few objects.

Critically, the model's probability distribution over possible object arrangements s' can be converted into a probability distribution over the total number of objects. Figure 4.3 shows the *implicit* distribution (y-axis) of numerical estimates (x-axis) for each number 1-15 (lines), at the same information bounds given in Figure 4.2. This visual memory model exhibits several key properties of number psychophysics, most notably a transition from exactness to scalar variability. The precise point of transition, as well as the acuity of estimation, are determined by the information bound, as in the model of Chapter 3. We show in Section 4.2 that the model transitions from subitizing to Weber's law specifically.

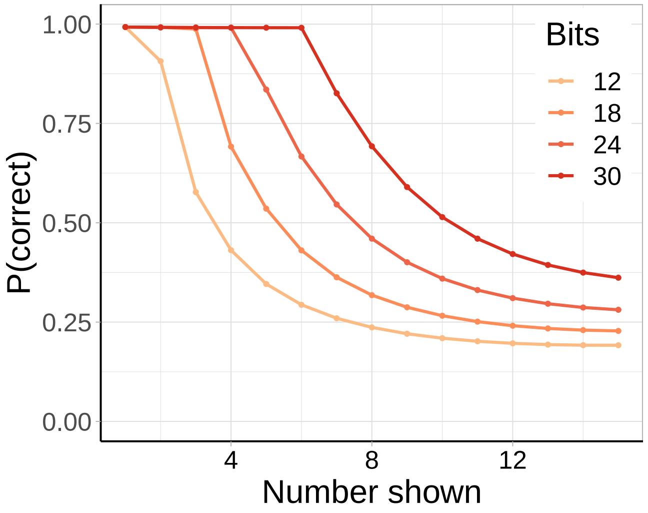


Figure 4.2: The model's predicted accuracy in a change localization task at information bounds ranging from 12-30 bits, assuming a 7x7 grid size and loss function parameter $\alpha = 1/3$ (as derived from model fitting).

Alternative loss functions

In the analyses of the experiments below, we use a loss function that combined a weighted proportion of false negatives and false positives relative to the number of locations with objects and locations without objects respectively. We had pre-registered this choice, however, it is not the only plausible loss function. One alternative choice would be the total number of places the representation s' differs from the scene a; another would be a possibly weighted combination of the *number* rather than *proportion* of false negatives and false positives. Here we show that while the choice of loss function somewhat influences the form of the resulting psychophysics, the outcomes are qualitatively very similar and preserve the core properties of the model in the paper.

For a given scene s and representation s' we will define a function for the number of false negatives $f_n(s' \mid s)$ and false positives $f_p(s' \mid s)$. We can write,

$$f_n(s' \mid s) = \sum_i \sum_j s_{ij} \cdot (1 - s'_{ij}), \tag{4.12}$$

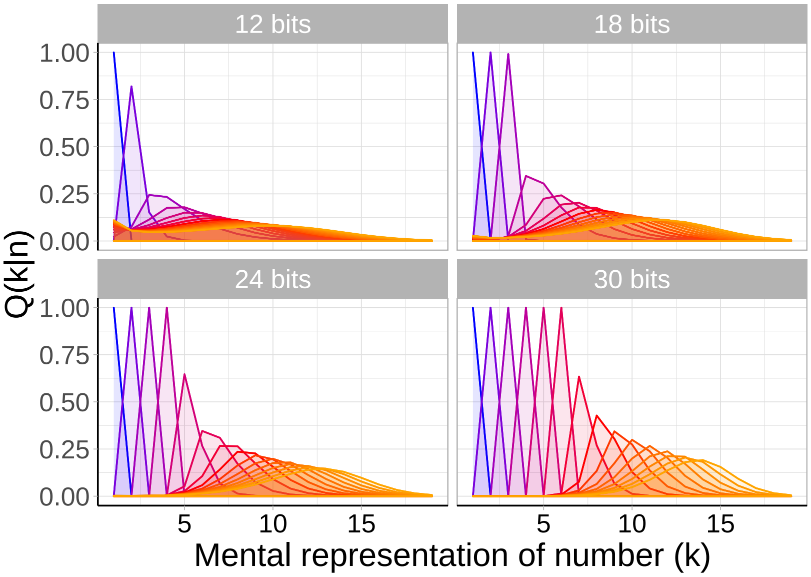


Figure 4.3: The implied psychophysics of number from the model of spatial memory, assuming a 7x7 grid size and loss function parameter $\alpha = 1/3$ (as derived from model fitting). Each line shows beliefs ($Q(k|n)$) over estimates (k) given numbers $n = 1...15$. Each facet shows the results of the optimization at various information bounds.

and

$$f_p(s' \mid s) = \sum_i \sum_j (1 - s_{ij}) \cdot s'_{ij}. \tag{4.13}$$

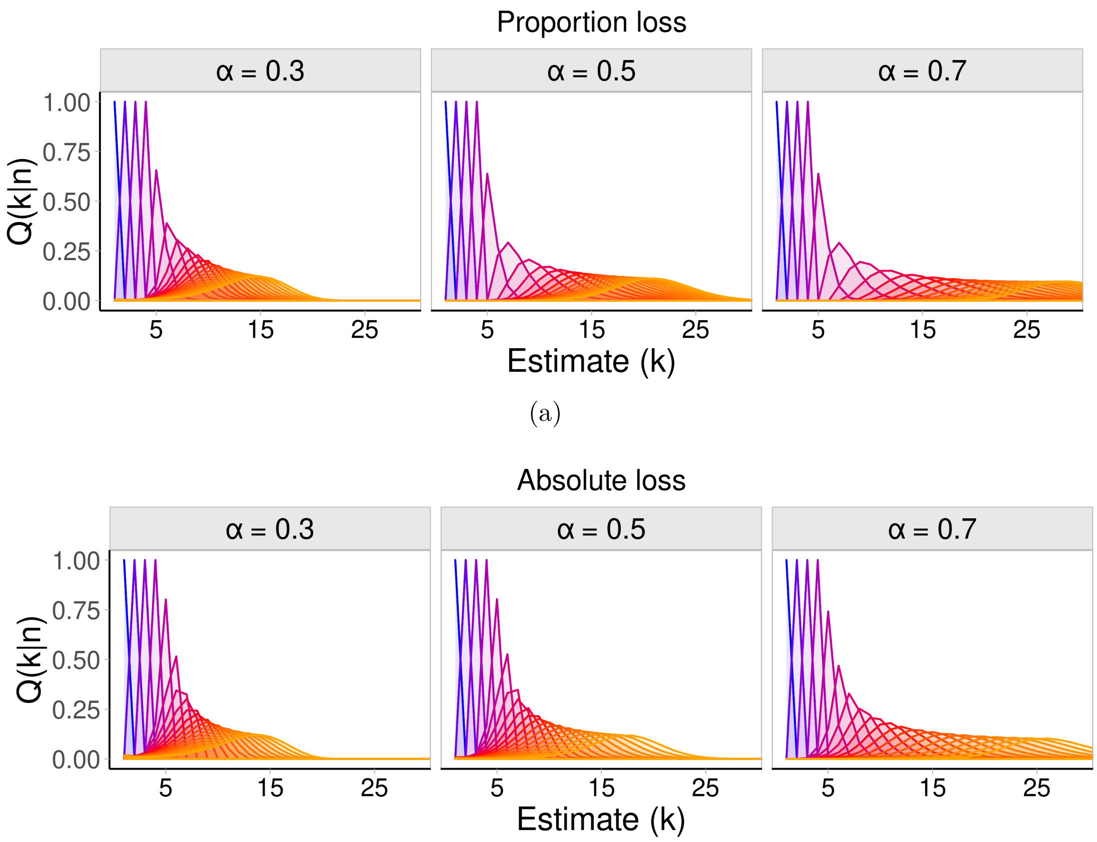
Then we can write the loss function assumed in the paper (using proportions) as,

$$\mathcal{L}_{proportion}(s, s') = \alpha \cdot \frac{f_n(s' \mid s)}{k} + (1 - \alpha) \cdot \frac{f_p(s' \mid s)}{n - k}. \tag{4.14}$$

The loss function that is a weighted combination of the number, rather than proportion, can be written as,

$$\mathcal{L}_{absolute}(s, s') = \alpha \cdot f_n(s' \mid s) + (1 - \alpha) \cdot f_p(s' \mid s). \tag{4.15}$$

Figure 4.4 shows predicted number psychophysics using both loss functions under different values of α, with Figure 4.4a showing the proportional loss function used in the primary analyses and Figure 4.4b showing the absolute numeric



(a)

(b)

Figure 4.4: These two plots illustrate the number psychophysics produced by various formulations of the loss function. Each line shows the estimates produced for a different number $n = 1 ... 20$. We assume an information bound of 25 bits. (a) These panels illustrate the psychophysics produced by different parameterizations of the loss function assumed in the main analyses, weighting the proportion of false negatives out of true positives by alpha and weighting the proportion of false positives out of the true negatives by one minus alpha. Each panel shows a different possible weighting, with $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$. (b) These panels illustrate the psychophysics assuming a loss function that is an analogous weighted combination of the *number* rather than *proportion* of false negatives and false positives.

loss function. At $\alpha = 0.5$ (middle panels), the weighting of both false negatives and false positives (either by proportion or absolute value) is equal; false negatives are under-weighted on the left panels and over-weighted on the right panels. Comparing the loss functions at each value of α, the psychophysics look very similar, particularly for low values of α. At higher values of α, the proportional loss function over-weights false negatives more strongly than the numeric counterparts for large numbers, and so ends up over-estimating.

The effect of the prior

In the model presented in the previous chapter, the decreasing prior over numerosities plays the central role in determining the noise and bias of estimates as a function of magnitude. That model would therefore predict that if a large number, say 75, happened to be high in the prior, people should be able to accurately represent sets of 75 items. But this seems perceptually implausible — could people really represent 75 items with higher fidelity than 2 items? One possible way of understanding the intuition that large groups of objects are intrinsically more difficult to represent precisely than smaller groups is that there is a lot more *spatial* information to represent about large groups.

If we take the simple method used here of dividing the world up into a grid with M possible locations, then there are $\binom{M}{n}$ ways to represent n objects in space. There are M places to put a single object, meaning it takes only $\log M$ bits to represent scenes when $n = 1$. However, there are many more ways to place n items when n grows larger (as it approaches its zenith at $\frac{M}{2}$). Using Stirling's approximation of the Binomial, it takes about $\log \frac{4^n}{\sqrt{\pi n}}$ bits to represent $\frac{M}{2}$ objects. To put this in perspective, if $M = 50$, it would take $\log 50 \approx 5.6$ bits to represent $n = 1$ object's location but about $\log \frac{4^{25}}{\sqrt{25\pi}} \approx 47$ bits to represent the location of $n = 25$ objects.

Unlike in the model we presented in the previous chapter, the model here accords with the intuition that more numerous sets are intrinsically more difficult to process perceptually. Even if there were a uniform prior over numerosities, small numerosities would be represented with significantly higher fidelity. In fact, the shape of the prior has much less of an impact on either mean estimates or the standard deviation of estimates relative to the loss function. We demonstrate this property in Figures 4.5-4.7.

Suppose the prior on a scene s with n objects is given by the function,

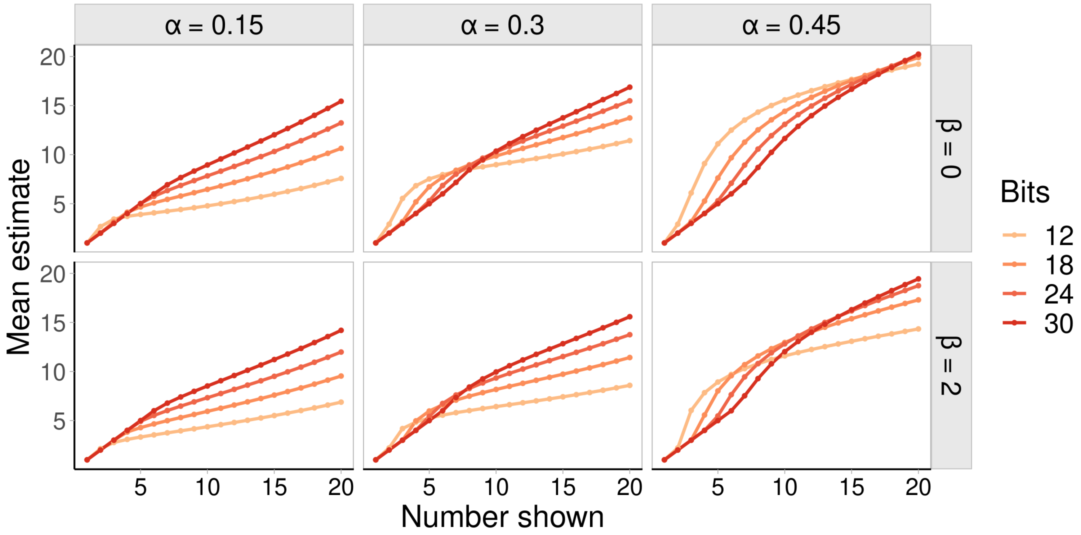


Figure 4.5: Predicted mean estimates as a function of the number of objects shown (x-axis) and the information bound (color). The columns give predictions under different loss function parameters (α) and the rows show predictions for a uniform prior distribution ($\beta = 0$) and naturalistic need frequency ($\beta = 2$) used in the main text.

$$P(s \mid |s| = n) \propto \frac{1}{n^{\beta} \cdot \binom{M}{n}}, \tag{4.16}$$

where β is a free parameter controlling the numerical bias. So $\beta = 2$ here is the naturalistic need frequency of number used in the main analyses ($P(n) \propto 1/n^2$) and $\beta = 0$ corresponds to a uniform prior over numerosities. Figures 4.5-4.7 give the model's predictions for mean estimates and standard deviations under these two distributions ($\beta = 0$ and $\beta = 2$), at different values of the loss function parameter α (controlling how much the model cares about false positives versus false negatives).

Figure 4.5 demonstrates that the bias in the model's mean estimates is affected much more strongly by α than by β — i.e., the loss function, rather than the prior, mostly determines the patterns of under- or over-estimation. Figure 4.6 shows, analogously, the model's predictions for the coefficient of variation (CV) as a function of numerosity ($CV = \frac{\sigma}{\mu}$). Crucially, Figure 4.6 illustrates that even with a uniform prior ($\beta = 0$), the model precisely represents small numerosities

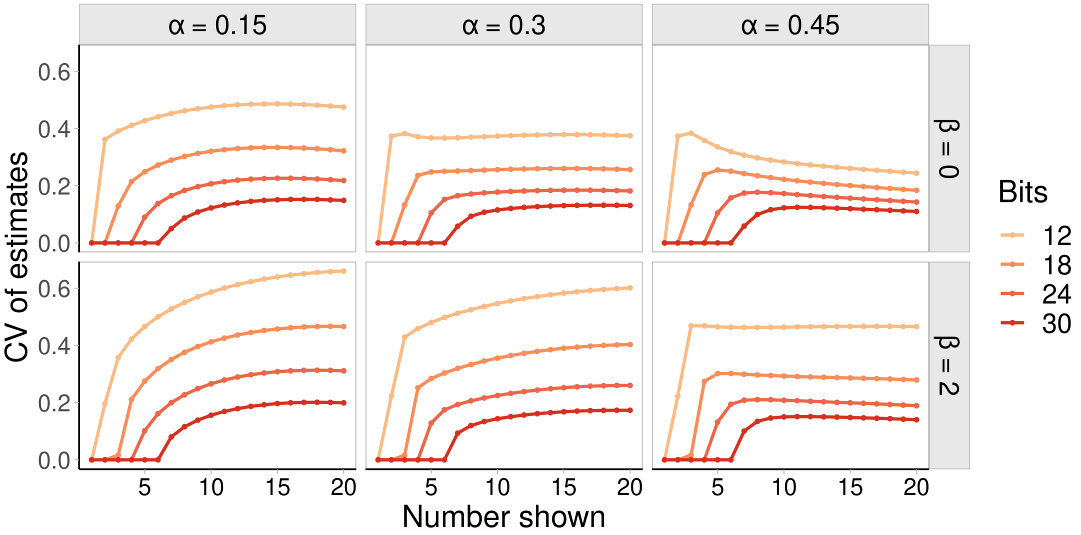


Figure 4.6: Predicted coefficient of variation ($CV = \frac{\sigma}{\mu}$) as a function of the number of objects shown (x-axis) and the information bound (color). The columns give predictions under different loss function parameters (α) and the rows show predictions for a uniform prior distribution ($\beta = 0$) and naturalistic need frequency ($\beta = 2$) used in the primary analyses.

but not larger ones. In fact, the point of transition from subitizing to estimation is essentially entirely determined by the information bound, with α and β only having any significant influence on the standard deviation of estimates beyond the subitizing range.

Finally, Figure 4.7 demonstrates that the change in CV converges to 0 for larger numerosities, across different choices of the prior and loss function. This indicates that the model recovers Weber's law in estimation — which predicts a constant CV across numerosities above the subtizing range — without requiring fine-tuning of any parameters. A further demonstration that the model recovers Weber's law in estimation is given in the section below.

Weber's law

In addition to an estimation task, the model can be extended to a numerical discrimination task. For two numbers n_1 and n_2, we make model predictions for n_1 and n_2 independently and subsequently compute the probability that the

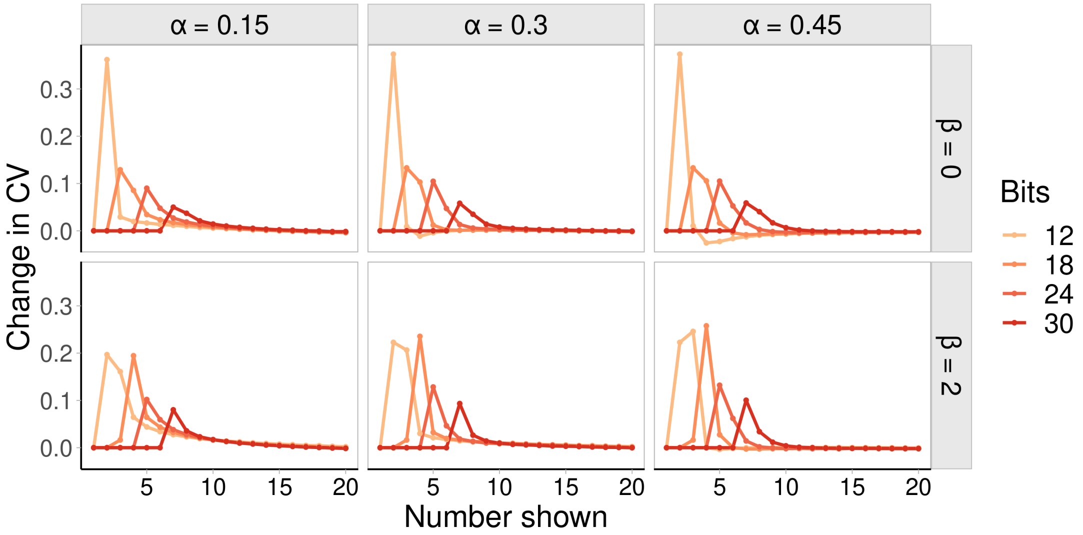


Figure 4.7: Predicted change in the coefficient of variation ($\delta CV = CV_n - CV_{n-1}$) as a function of the number of objects shown (x-axis) and the information bound (color). The columns give predictions under different loss function parameters (α) and the rows show predictions for a uniform prior distribution ($\beta = 0$) and naturalistic need frequency ($\beta = 2$) used in the primary analyses.

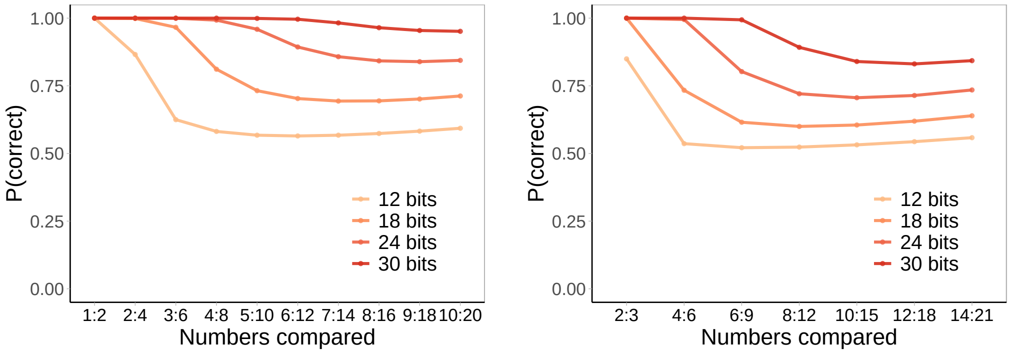
model believes that n_2 was greater in magnitude than n_1,

$$P(n_2 > n_1) = \sum_{k=1}^{M-1} \sum_{j=k+1}^{M} P(k \mid n_1) \cdot P(j \mid n_2). \tag{4.17}$$

Figure 4.8 shows model predictions for discrimination performance on 1:2 ratios for numerosities 1:2 through 10:20 (a) and 2:3 ratios for numerosities 2:3 through 14:21 (b) across information capacity bounds. Weber's law implies that performance should be constant across ratios, which is true for the model somewhat beyond the subitizing range.

The relationship between subitizing and estimation

Previous work has shown that the relationship between subitizing and estimation is not straightforward. For instance, subitizing seems to be more greatly affected by attentional load than estimation (Burr et al., 2010); other studies have found



(a) (b)

Figure 4.8: Model predictions for numerical discrimination on (a) 1:2 ratios and (b) 2:3 ratios. The model was parameterized with $\alpha = 1/3$ and the prior used in the primary analyses ($P(n) \propto 1/n^2$).

little or no correlation between one's subitizing range and their estimation acuity (Revkin et al., 2008). One possible explanation afforded by the model is that some small changes in capacity can lead to sharp changes in the subitizing range. Conversely, other changes in capacity can lead to no changes in the subitizing range whatsoever. This could lead to puzzling results — subitizing and estimation will sometimes seem related but sometimes not. But, as we show, the model actually predicts that large changes in capacity are necessary for the relationship to become apparent.

We modeled the relationship between estimation acuity and subitizing range with the range of numerosities (1-8) tested in the studies cited above $(2, 64)$. The subitizing range was calculated as the largest number with $\epsilon < 0.001$ squared estimation error; and the estimation acuity was calculated as the average coefficient of variation of numerosities beyond the subitizing range. Figure 4.9 shows the results of this simulation, with the subitizing range on the x-axis, estimation acuity acuity on the y-axis, and each point representing the model's prediction at a given information capacity. There are sudden changes in the subitizing range as the information capacity increases; conversely, there are small, less dramatic effects on estimation acuity.

Because the subitizing range can change dramatically without requiring es-

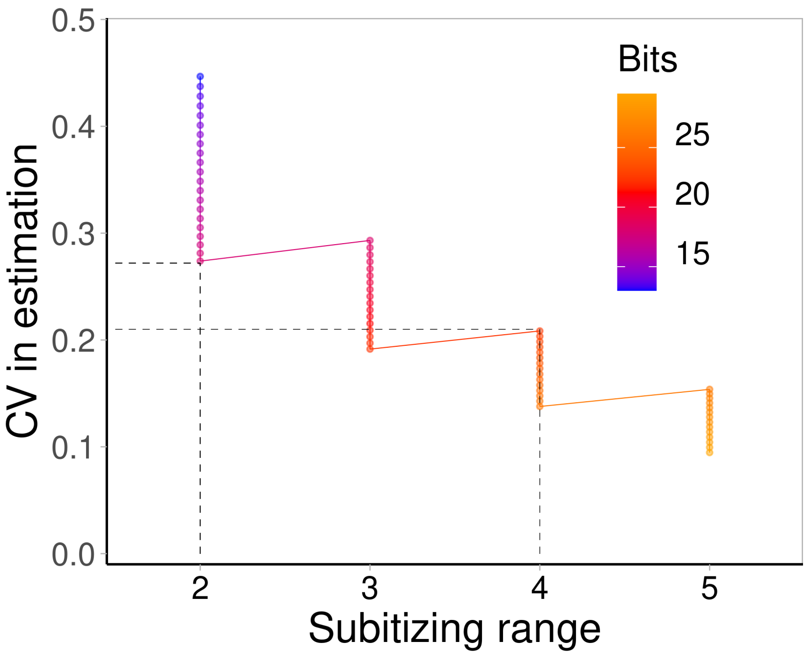


Figure 4.9: The relationship between subitizing range (x-axis) and estimation acuity (y-axis) across information capacities (colors). Changes in capacity always change the observed estimation acuity but only sometimes dramatically change the subitizing range.

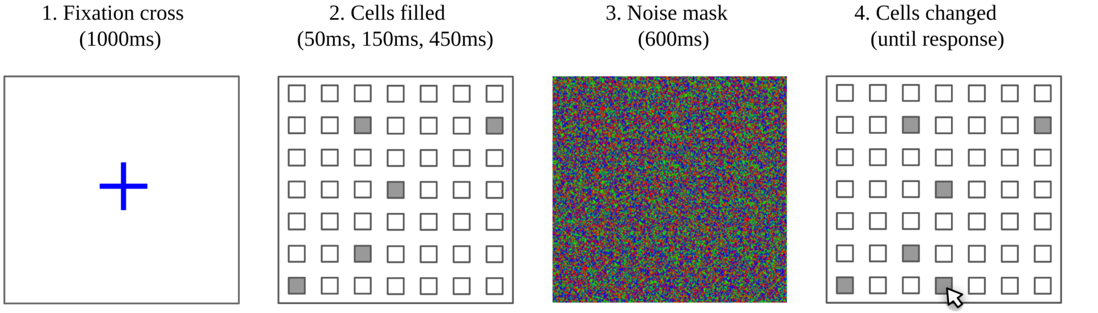
sentially any change in estimation acuity[2], it may not be altogether surprising that some studies have found that the subitizing range is affected by an attentional manipulation when estimation acuity is not. The relationship between the subitizing range and estimation acuity should only become apparent with substantial changes in capacity — and even then, estimation acuity need not change by a substantial margin. For instance, to increase the subitizing range from 2 to 4 would only require a decrease of the coefficient of variation in estimation from 0.27 to 0.21 (highlighted in Figure 4.9 by the dashed lines). This level of change seems insubstantial relative to the change in subitizing range — and may even be hard to detect without high statistical power — but does not imply that the

[2] One curious thing to note is that when the subitizing capacity changes, the observed estimation acuity actually very slightly *decreases*. This is because numerosities very near the subitizing range tend to have slightly higher acuity than larger numerosities, but when the subitizing range increases to encompass that numerosity, it is no longer counted towards the average estimation acuity.

two phenomena are unrelated.

4.3 Experiment 1

In order to understand how visuospatial perception in humans is modulated by processing time and the number of objects in a scene, we ran a change-localization task in which items flashed on a screen, disappeared, and then re-appeared with a single modification (illustrated in Figure 4.10). Our visual model predicts that, given sufficient processing time, participants should be able to remember the locations of small groups of objects with high fidelity but become increasingly inaccurate for larger numerosities, which accords with basic intuition and previous findings (Alvarez & Franconeri, 2007; Vul et al., 2009). With only limited processing time, however, participants should become increasingly incapable of remembering the locations of even a small number of objects, and the disparity in performance between smaller and larger groups should decrease, per Figure 4.2. In addition to testing whether localization is well explained by the model, by fitting the information bound to *non*-numerical human spatial memory, we can test whether the inferred parameters are consistent with the psychophysics of number.



Please click the cell that turned from **white to gray.**

Figure 4.10: An illustration depicting each step of a trial in Experiment 1. Participants were first shown a fixation cross, followed by a 7x7 grid with some of the cells (1-15) filled in gray. A noise mask then appeared after a short time (50ms, 150ms, or 450ms). In the final step, participants were shown a display identical to the one shown previously except for a single cell — one of the previously gray cells either turned white ("disappearance") or one of the previously white cells turned gray ("appearance"). Participants tried to guess which cell had changed.

Method

Participants

We recruited 110 registered users of Prolific, an online psychology experiment platform. Participants were 18 years old or older, fluent English speakers, and physically present in the United States based on pre-screening questions. Each participant who completed the task received compensation of $3. Both experiments were approved by the University's Institutional Review Board and complies with all relevant ethical regulations. Informed consent was obtained from all participants before beginning the study. Following the pre-registration, we removed the 10 participants with the highest error rate from our analyses. Based on pilot studies and previous work, we believed the sample size included for analysis (100 participants x 90 trials per participant = 9000 data points) would be sufficient to determine model parameters within a small interval.

Materials

The experiment was designed in JavaScript using the psiTurk framework (Gureckis et al., 2016). There were 49 grid cells (7 x 7), with each grid cell $35px^2$ and an equal margin separating the cells. Unfilled grid cells were white and filled grid cells were gray with hex color #A0A0A0. When a cell was clicked in the task, its border was bolded and turned red. The noise mask was multicolored static and had a size of $455px^2$ to cover the entire grid.

Design

There were four within-subject variables manipulated in the study: the number of cells filled (1-15); the exposure time of the displayed pattern (50ms, 150ms, 450ms); and the direction of the changed cell from the first to second presentation (white-to-gray or gray-to-white). Each three-tuple of number, time, and direction was shown exactly once, for a total of 15 x 3 x 2 = 90 trials. The initial direction of changed cell was randomly chosen and then remained constant for the first 45 trials, with the last 45 trials assigned to the opposite direction. Within that constraint, the order of the trials was randomized, i.e. number-time pairs were assigned randomly within each direction of change. The positions of the filled cells were chosen randomly on each trial. If the direction of change was white-to-gray, a random white cell from the initial exposure would turn gray on the second presentation; conversely, if the direction of change was gray-to-white, a random gray cell would turn white.

Procedure

After providing consent and reading instructions, participants began the first section of the experiment. Both halves of experiment — the white-to-gray section and gray-to-white section — started with 3 practice trials. Participants were informed in both the practice trials and main task whether a cell would be changing from white to gray or vice-versa. Each trial started with a fixation cross displayed on the center for 1000 ms, followed by the grid with some cells filled in (50-450ms) and then a noise mask for 600 ms. Then, the grid reappeared, with one modified cell. Subjects then clicked the cell they thought changed color and proceeded to the next trial. The basic setup is illustrated in Figure 4.2.

Model fitting

For both experiments, we used a Markov Chain Monte Carlo (MCMC) algorithm to fit four parameters to the data: a) power law functions for how the information capacity changes over time, of the form $a \cdot t^k$, with a and k as free parameters and t representing time in seconds; b) the loss function parameter α, which weights the cost of false negatives and false positives; and c) a guessing parameter p_g which captured the rate of choosing randomly. Because α and p_g represented probabilities and thus were constrained to be between 0 and 1, we parameterized these through transformations $\alpha = logit(\alpha')$ and $p_g = logit(p'_g)$. We fit these parameters in a hierarchical Bayesian network, with partial pooling of parameter estimates across participants. We used uninformative group-level priors for the means of each parameter, which we believed would not exert a strong influence in any case given the large amount of data collected. We drew group-level standard deviations from $HalfNormal(\sigma = 10)$. Subjects' parameters were drawn from the distributions,

$$a_s \sim Normal(\mu_{a,g}, \sigma_{a,g}), \tag{4.18}$$

$$k_s \sim Normal(\mu_{k,g}, \sigma_{k,g}), \tag{4.19}$$

$$\alpha'_s \sim Normal(\mu_{\alpha',g}, \sigma_{\alpha',g}), \tag{4.20}$$

$$p'_{g_s} \sim Normal(\mu_{p'_g,g}, \sigma_{p'_g,g}), \tag{4.21}$$

where group-level parameters are denoted $\mu_{.,g}$ and $\sigma_{.,g}$ and subject-level parameters are denoted with subscript s.

We used the Metroplolis-Hastings algorithm to jointly fit the posterior distributions of each group-level and subject-level parameter. Because there is a high runtime cost to compute the model's posterior distribution, we rounded the information bounds given by samples of a and k to the nearest 0.1, and each α to the nearest 0.01, and cached the results. This can only have a negative impact on the fit of the model and so it could not impact (e.g.) model comparisons in our model's favor. We ran two chains of Metropolis-Hastings for 50,000 steps, with 10,000 steps of burn-in, storing every 10th value to avoid auto-correlation of samples. We checked for convergence of the chains using the Gelman-Rubin statistic (A. Gelman & Rubin, 1992), and found in both tasks that $\hat{r} < 1.05$ for all group-level parameters and $\hat{r} < 1.1$ for all subject-level parameters, indicating that the chains converged.

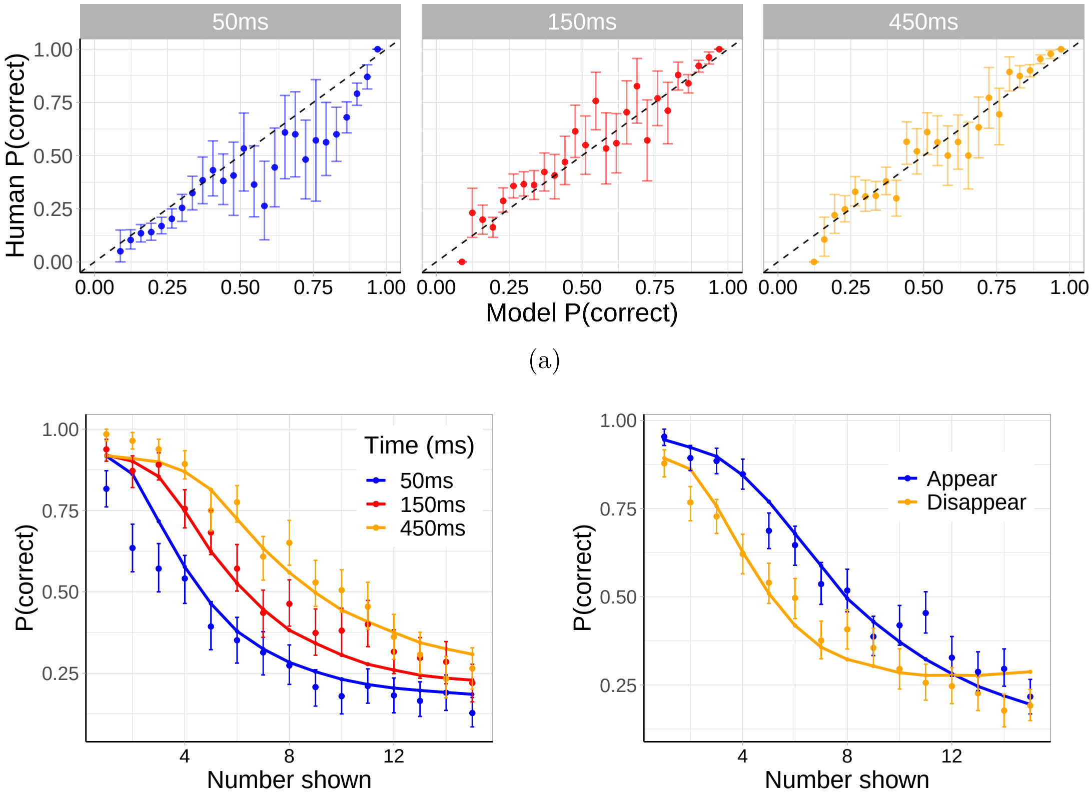
Results

To fit model parameters, we assumed that the information bound changes as a function of time according to a power law $B = a \cdot t^k$, where a and k are free parameters and t is exposure time in seconds. The other key parameter of the model is the weighting parameter in the loss function α, capturing the extent to which false negatives (high α) or false positives (low α) are more costly. To account for attention lapses and mis-presses, we also included a guessing-rate parameter, p_g, which captured the rate participants chose randomly from the set of valid alternatives (as opposed to via the model). We fit parameters under a hierarchical Bayesian model using MCMC, assuming partial pooling of parameter estimates across participants (see SI).

The Maximum A Posteriori (MAP) estimates for the group-level parameters were, $a = 33.5$ (CI=[32.2, 34.6]), $k = 0.21$ (CI=[0.20, 0.22]), $p_g = 0.16$ (CI=[0.12,0.19]), and $\alpha = 0.35$ (CI=[0.33,0.37]). This entails information bounds of 17.9, 22.5, and 28.3 bits at 50ms, 150ms, and 450ms, respectively. The relatively high inferred rate of guessing likely reflects the fact that the model does not account for spatial errors, treating each cell independently. Figure 4.11a shows binned model predictions for accuracy (x-axis) against human performance (y-axis) across all exposure durations (facets). Comparing the points to the dashed $y = x$ line reveals that the model's predictions tightly align with human accuracy across exposure durations, though the model is slightly biased to over-estimate human performance at short times (left facet). The correlation between model predictions and human data across trials grouped by numerosity and exposure duration was 0.96 ($R^2 = 0.93$), another indication that the model provides a good fit to the data.

Turning to the crucial question of how performance on the change-localization task was affected by the number of filled cells over time, the model predicts near-veridical memory for visual displays with small numbers of objects, at longer exposure durations, and sharply increasing noise for larger numbers of objects and shorter durations. Figure 4.11b shows human accuracy (points and error bars) and the model's predicted accuracy (lines) as a function of the total number of cells filled in, grouped by the exposure duration (colors). As predicted by the model, participants' performance saturates only for small numerosities at longer durations and quickly degrades as a function of number in each duration. The one notable discrepancy is that the model predicted better performance on small numerosities ($n < 4$) at $50ms$ than was actually observed. Figure 4.11c depicts accuracy grouped by whether a cell appeared or disappeared from the first to second display, and shows that participants performed substantially better on "appear" trials than "disappear" trials — a trend the model captures. The model would capture this trend even if α were fixed to 0.5, and in fact higher values of α exaggerate rather than reduce the gap between "appear" and "disappear" trials.

To be clear, the fact that human performance on the change-localization task is strongly affected by numerosity is not an indication that the visual system is representing or using number — the experiment was explicitly designed so



(a)

(b)

(c)

Figure 4.11: (a) Binned (25 bins/facet) model predictions (x-axis) and human data (y-axis) of performance on the change-localization task. Each facet shows predictions at different exposure durations. In (b) and (c) model predictions are shown as lines and human data from the change-localization task are shown as points with bootstrapped 95% confidence intervals. (b) Accuracy (y-axis) in the change-localization task as a function of the number of grid cells filled (x-axis) at each exposure duration. (c) Accuracy (y-axis) as a function of number (x-axis) grouped by whether or not a cell appeared or disappeared from first-to-second presentation.

that number cannot be used as a heuristic. Instead, the effect of numerosity on performance is an indication that spatial memory is making use of limited information in an efficient way, combining a prior expectation that there will be fewer gray pixels than white pixels with evidence gathered by observing the scene. Additionally, the inability to precisely remember scenes with more filled cells is a reflection of the fact that there are more ways to arrange scenes with more filled cells (up to half the number of grid cells), meaning that it takes more information to represent any one of them precisely.

4.4 Experiment 2

While Experiment 1 showed that the model is able to account for effects of number and exposure duration in spatial memory, it does not answer the question of whether human numerical estimation abilities arise from this same system. If the patterns of noise and bias in estimation derive from limitations in spatial memory, then the model of spatial memory should be able to explain the psychophysics of estimation over time; moreover, we should be able to recover similar parameter values from the model fit to a numerical estimation task as from the model fit to a spatial memory task. To test this, we ran a number estimation task with a design matched to Experiment 1.

Method

The procedure and display was identical to Experiment 1 up to the noise mask. After the noise mask, however, participants were asked to estimate the number of cells that were filled. 110 adult participants from Prolific again completed 90 trials, with each number (1-15) paired with duration (50ms, 150ms, 450ms) displayed twice. Following the pre-registration, we removed the 10 participants with the highest mean absolute error in estimation from our analyses and windsorized estimates to the 95% interval for each numerosity.

Results

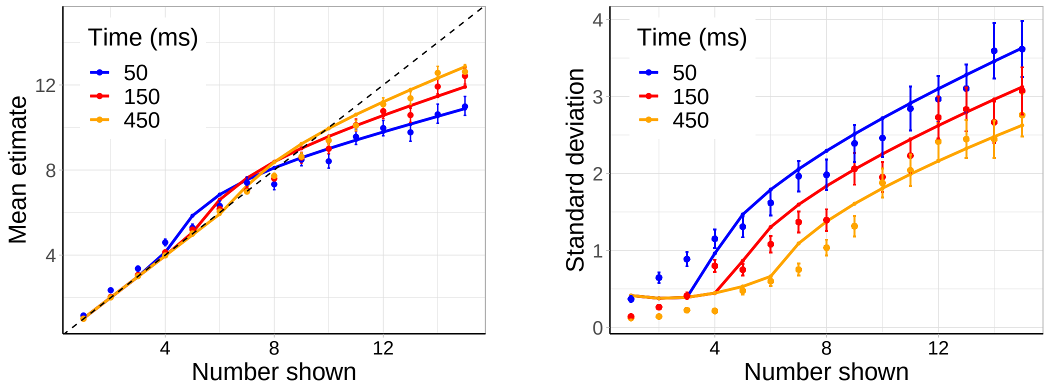
We fit the same parameters in the model with the estimation data as with the change localization task. The MAP group-level parameters were $a = 32.9$ (CI=[30.9,33.8]), $k = 0.18$ (CI=[0.17,0.20]), $p_g = 0.03$ (CI=[0.02,0.03]) and $\alpha = 0.31$ (CI=[0.29,0.32]). The implied average information bounds are therefore 19.2, 23.4, and 28.5 bits at 50ms, 150ms, and 450ms respectively. This is slightly higher than the estimates derived from the change-localization task data, but the differences at each exposure duration are small ($< 10\%$) at each exposure duration. Table 4.1 provides a side-by-side comparison of the inferred MAP parameters from both experiments. A notable difference between the inferred parameters between the two tasks is the guessing rate, which is much lower than in the change-localization task. As noted previously, however, the relatively high guessing rate in the change-localization task is likely due to the fact that the model does not account for spatial errors or mis-presses (only completely random guessing) — this would increase the inferred rate of guessing in the change-localization task but not the estimation task.

MAP parameters from both experiments				
Experiment	a	k	α	p_g
Localization (E1)	33.5	0.21	0.35	0.16
Estimation (E2)	32.9	0.18	0.31	0.03

Table 4.1

The resulting psychophysical curves from the model (lines), along with the data from the experiment (points and error-bars), are shown in Figure 4.12. The model captures the key psychophysical trends observed in the data: an underestimation bias that diminishes with exposure time; a subitizing range that increases with exposure time; scalar variability in estimation; and acuity in estimation that increases with exposure time. The non-zero but flat standard deviation for small numerosities in 4.12*b* reflects influence of guessing — without the guessing parameter it would show zero variability. The model predictions diverge somewhat from human performance on small numerosities ($n < 4$) at $50ms$ — the model predicts better performance than is actually observed. An analogous discrepancy was observed in the change-localization task (also for $n < 4$ at $50ms$), which makes this deviation less concerning to the validity of our proposal that the two abilities are intimately related (in fact, it may even bolster this claim).

Following the pre-registration, we compared the model's Maximum Likelihood Estimate (MLE) parameters for each subject to a standard psychophysical model of numerical estimation as well as a modified one that accounts for the effects of time. The overall log likelihood of the model using MLE estimates of participants'



(a) (b)

Figure 4.12: Model predictions (lines) and data from the estimation experiment (points and 95% CI). (a) Mean estimates as a function of numerosity, grouped by exposure duration. (b) Standard deviations as a function of numerosity, grouped by exposure duration.

parameters was -14,129. In the first comparison model, we assume that participants' estimates are drawn from a Gaussian centered around the number shown, n, with mean n and standard deviation $w \cdot n$, where w is a free parameter (their "Weber fraction"). We also fit a version of this where the standard deviation could vary as a function of time, such that $w = e^{w_0 + w_t \cdot t}$, where w_0 and w_t are fit and t is time in seconds. The median MLE w fit in the static (non-time-varying) version was 0.24, with log likelihood -16,166. In the time-varying version, the median MLE w_0 was -1.15 and w_t was -1.75, giving w's of 0.29, 0.24, and 0.15 at 50ms, 150ms, and 450ms respectively, and had log likelihood -15,428. The Weber models thus did not fit nearly as well as our model, with AIC differences of 3,974 and 2,498 (we pre-registered AIC differences of 10 as "significant").

4.5 Discussion

This chapter presented a novel model of visuospatial memory that captures both human performance in both a spatial memory task and in a number estimation task. Crucially, we are able to recover the key properties of numerical cognition in an entirely non-numerical visual task using a visual model; moreover, the patterns of noise and bias in estimation align precisely with the noise inherent to spatial memory, indicating that the psychophysics of number are attributable to perceptual uncertainty rather than number-specific processing. Our results show that the defining features of numerical cognition can be understood as downstream consequences of basic visual processing, posing a challenge to theories that assume the psychophysics observed in estimation are the result of number-specific processing via one or more "number systems." While there must exist some number-specific processing—quantity must be extracted from visual memory—our findings indicate that Weber's law, subitizing, under-estimation and other effects observed in numerical estimation are not the direct *result* of that processing.

It is worth noting that some studies have found a strong relationship between object-tracking ability, visual memory capacity, and estimation acuity outside the subitizing range, as predicted by our model (Bugden & Ansari, 2016; Green & Bavelier, 2003, 2006; Passolunghi et al., 2015). However, other studies have found a stronger link between an individual's visual working memory capacity and their subitizing range than with their estimation acuity (Piazza et al., 2011; Revkin et al., 2008), which might seem to contradict predictions of our theory. Importantly, though, while the model does link both subitizing range and estimation to visuospatial information capacity, differences in information capacity do not necessarily cause equally large changes to the subitizing range and estimation acuity. Specifically, modulating the information bound tends to affect the subitizing range substantially more than the (implicit) Weber fraction.

More speculatively, another issue our model may help address is why, despite our subjectively rich experience of the world, people have such limited ability to estimate quantities. Models of numerosity discrimination in neural networks have essentially had to impose strict information bottlenecks — such as using unsupervised learning and limiting the number of hidden units — to roughly reach parity with human levels of performance (Stoianov & Zorzi, 2012; Testolin, Dolfi, et al., 2020; Testolin, Zou, et al., 2020). It is almost certainly the case that more powerful networks, like those that have reached near human-level performance on object recognition tasks, would significantly out-perform humans on numerical estimation and discrimination tasks with enough supervised learning.

One resolution to this issue suggested by our approach is that people's ability to extract numerical information about even large quantities is necessarily tied to their ability to track objects. This is in a way similar to the implicit solution provided by the neural network models, which only allow linear classifiers to train on frozen hidden layer representations of the visual scene.

Finally, it is worth highlighting two important limitations of our model and experiments that leave room for future work. First, the model and experiments were only designed to capture numerical perception in the domain of vision. However, innate numerical abilities have been documented in audition, touch, and across modalities (Barth et al., 2003; Meck & Church, 1983; Plaisier et al., 2009). Though the model we presented here was designed to deal with spatial rather than temporal integration, we believe similar principles of information processing apply and hence the methods used in this chapter could be adapted to capture (e.g.) the processing of auditory sequences. The other main limitation is our use of simplifying assumptions to model spatial memory — specifically, in discretizing the space so coarsely and in assuming objects to be identical. The model would thus likely need to be extended to capture, for instance, the influences of continuous visual features such as surface area, convex hull, and density on numerosity perception (e.g. Gebuis et al., 2016; Gebuis & Reynvoet, 2012a; Lourenco, 2015; Lourenco & Longo, 2010, 2011; Sokolowski et al., 2017). In fact, the methods we employed here may be useful to understanding some of these effects: since continuous features like surface area are correlated with numerosity in the real world, principles of efficient information compression dictate that their representations will not be independent.

5 Conclusion

To conclude, I will briefly summarize each chapter and then highlight the main takeaways of our work, along with some speculation about the broader implications of our findings. In Chapter 2, we found that numerical estimates are driven by a serial accumulation process operating over saccades: as participants fixate on more points, their mean estimates increase and the variance of their estimates decreases. In Chapter 3, we found that the discontinuous psychophysics of small- and large-number estimation can be understood as an optimal representation given an information capacity limit, and that varying the amount of visually available information modulates key properties of a person's number psychophysics like their subitizing range in a predictable way. In Chapter 4, we found that the psychophysics of number are largely driven by domain general perceptual processing — specifically, uncertainty about where objects are in space — and that subitizing, Weber's law, underestimation, and other effects can all be understood as consequences of a limited capacity to represent objects in space.

What's in a Weber fraction?

A key takeaway of our work is that numerosity perception cannot be viewed separately from general perceptual processing. Although "the approximate number system" has become common parlance, this term suggests the existence of an insular mechanism devoted to processing quantity, obfuscating the domain-general *perceptual* grounding of innate numerical abilities. Hundreds of papers in the literature describe participants' performance on numerical discrimination and estimation tasks as reflective of “the precision of their ANS” or “the acuity of their innate number sense.” However, as Chapter 2 shows, numerical estimates are driven in large part by where one happens to visually fixate, demonstrating that Weber fractions are neither a static measure nor a simple reflection of one’s innate number sense. Furthermore, as Chapter 4 shows, the widespread notion that Weber fractions index the noise inherent to one’s mental number line seems to be false: the uncertainty in numerical estimates seems to largely derive from upstream perceptual uncertainty regarding where objects are in space — i.e., not from noise added independently to representations of numerosity.

To be clear, our results do not suggest that people don’t represent number — in fact, they support the opposite — or even that people don’t have an innate “number sense.” Instead, they demonstrate that a person’s performance on an estimation task cannot be simply interpreted as a pure index of their innate number sense and that accounting for non-numerical factors, such as display time (Inglis & Gilmore, 2013), is critical to understanding performance differences between individuals within and between tasks. Yet, for instance, in a paper entitled “Developmental changes in number sense acuity,” a display time of 2,500ms was used for 3-year-olds, 1,200ms for 4-6 year-olds, and 750ms for adults (Halberda & Feigenson, 2008) — likely distorting observable changes in “number sense acuity.” The magnitude of such a distortion may be significant, as this study with 750ms display time for adults found much greater ANS precision (mean $w = 0.1$) than in another study using a 200ms display time (Halberda et al. (2008), mean $w = 0.3$).

Similarly, the findings presented in Chapter 2 demonstrating a strong link between visual fixations and numerical estimates suggest that basic visual acuity, oculomotor control, and attention may be just as important to performance on a number estimation task as one’s “innate number sense.” The point that performance on a number estimation task does not only reflect innate number abilities, while obvious sounding, is worth emphasizing because there are a number of poorly controlled studies aimed at investigating links between innate numerical abilities and mathematical achievement. However, the entire plausibility of a causal link between innate numerical abilities, as measured by a scalar value w, and mathematical achievement derives from the assumption that individual differences in w reflect differences in *number* representations rather than, say, oculomotor control or visuospatial memory capacity.

Does Weber's law require continuous, analog representations?

Discrimination of many magnitudes —including number, duration, length, and luminance, among others — follow Weber's law (or approximately do so). That is, the ease of discriminating between any two magnitudes depends on their ratio. It is widely accepted, to the point of being cited as mere fact in many papers and books, that this is a hallmark of *analog magnitude* representations. For instance, in her book *The Origin of Concepts*, Susan Carey (2009) writes,

> A psychophysical signature of analog magnitude representations is that the discriminability of any two magnitudes is a function of their ratio; that is, discriminability is in accordance with Weber's law.

It is true that, under some conditions, discriminability of sensory stimuli will vary according to Weber's law. For instance, if magnitudes are compressed onto a logarithmic scale with constant internal noise, or represented linearly with scalar noise, this will result in Weber's law. A log-compression encoding scheme that produces Weber's law is illustrated in Figure 5.1.

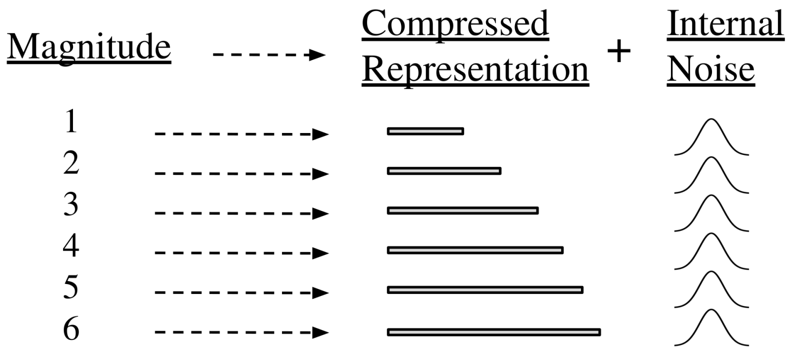


Figure 5.1: Log-compression of magnitudes with constant internal noise produces Weber's law.

However, there are two issues with what Carey wrote: 1) continuous, analog representations do not always produce Weber's law; and 2) Weber's law can arise from discrete, non-analog representations. On the first point, Weber's law is not a necessary result of analog encoding — it arises only from particular compression and noise schemes. For instance, the noise in representations of texture-density scale roughly with the square-root of their magnitude (e.g. Anobile et al., 2014, 2016), which does not give rise to ratio dependence in discrimination. Yet, human texture-density representations are still considered "analog."

The second point, which is more important and a departure from conventional wisdom, is that Weber's law does not necessarily result from analog or continuous representations. In the model presented in Chapter 4, probability distributions over the number of objects present are induced by uncertainty about objects' locations in space. In this model, the representations of both objects' locations and the number of objects present are discrete and non-analog[1]. Yet, as Figures 4.7 and 4.8 show, Weber's law results from this setup because the uncertainty in representations of quantity scales linearly with the number of objects. The model thus demonstrates how noisy beliefs over discrete representations can give rise to what appears to be analog behavior. To be clear, the model in Chapter 4 does not specify how numbers are represented, only that perceptual uncertainty about objects' spatial locations has the downstream effect of producing Weber's law in quantity discrimination. To conclude, then, Carey (and many others) have over-interpreted the implications of observing Weber's law in number discrimination: this does not imply that numerosity is represented as an analog magnitude on a continuous scale.

What number of number systems is the right number?

A broad theme of the work presented in this thesis is finding common functional and mechanistic origins of number psychophysics that have historically been treated as separate phenomena. For instance, people's exact representation of small sets (Burr et al., 2010; Choo & Franconeri, 2014; Feigenson et al., 2004; Jevons, 1871; Revkin et al., 2008) but increasingly imprecise representation of larger sets (Dehaene, 2011; Xu & Spelke, 2000) has been explained as arising

[1] The representations of numerosity *could* be analog but they are not necessarily — the point is that this doesn't matter.

from different representational systems (Dehaene, 2011; Feigenson et al., 2004). However, the model presented in Chapter 3 demonstrates that this discontinuity is actually an efficient representation of number given a limited information capacity, suggesting that the discontinuity need not reflect different representational systems. In addition to this theoretical argument, the empirical finding that the amount of visually available information (manipulated by exposure time or color contrast) alters both the subitizing range and the precision of large number representations in predictable ways seems hard to explain in a standard two-system account of numerosity perception.

However, the model in Chapter 3 is in essence a functional-level account explaining why number psychophysics looks the way it does, rather than a mechanistic one, and so it cannot be used to directly assess whether there are one or two (or ten or twenty) systems. On the other hand, the model and empirical results presented in Chapter 4 do have more direct implications for the number, and nature, of number systems. While the optimization model in Chapter 3 assumes that the visual system aims to minimize estimation error, the optimization in Chapter 4 accounts for the fact that numerosity itself cannot be directly optimized by perception, since the visual system is not directly presented with quantities — it is presented with objects in space, which must be represented before being transformed into a quantity estimate. Surprisingly, it turns out that optimizing memory to accurately remember objects' locations results in very similar predictions about number psychophysics as directly optimizing for numerical estimation accuracy. For instance, subitizing, Weber's law, underestimation, and the temporal dynamics of number psychophysics remain qualitatively the same in the two cases. But, importantly, the spatial encoding model additionally predicts that the capacity to remember the locations of objects should significantly influence (if not entirely determine) the capacity to estimate quantities.

We found that, indeed, the inferred capacity limit for remembering objects' spatial locations is nearly identical to the inferred capacity limit for numerical estimation. Moreover, the two capacity limits track very closely over time for numerosities in the range 1-15 and the model closely fits human data in both of these tasks. While this *could* be a (parsimoniously explained) coincidence, it seems quite unlikely, especially given how plausible the link between the two abilities is: representing a topographic map of objects in space would seem to be a prerequisite for estimating their numerosity. How else could numerosity be directly (rather than indirectly via continuous features) computed? What these results imply, then, is that number psychophysics in both the small- and large-number regimes have a common origin in spatial memory; and, while we cannot say with certainty if there are one or one-hundred number systems, it seems that the processing that takes place beyond spatial memory does not determine much, if any, of the core properties of number psychophysics.

To summarize, we are making two claims. First, the fact that a single optimization produces the discontinuous psychophysics of number estimation implies that observing a discontinuity cannot be used as evidence of two representational systems. Second, regardless of how many there are, the number systems themselves do not seem to be the basis of number psychophysics — those seem to arise due to a limited capacity to represent objects' locations in space. However, we acknowledge that analysis of behavior alone will probably not be sufficient to conclusively affirm or deny the existence of two representational systems, and a careful study of the neural systems involved in processing quantity will probably have a decisive role in the end. The results of neural studies looking at numerical representations to this point, unfortunately, are mixed (Cai et al., 2021; Ditz & Nieder, 2016; Hyde & Mou, 2016; Hyde & Spelke, 2011; Nieder & Merten, 2007). It is worth noting, though, that a recent study using a high-resolution fMRI scanner failed to find topographic differences in small- and large-number representations in the brain (Cai et al., 2021).

Bibliography

Agrillo, C., Petrazzini, M. E. M., & Bisazza, A. (2014). Numerical acuity of fish is improved in the presence of moving targets, but only in the subitizing range. *Animal cognition*, *17*(2), 307–316.

Alexander, R. M. (1984). The gaits of bipedal and quadrupedal animals. *The International Journal of Robotics Research*, *3*(2), 49–59.

Alonso-Diaz, S., Cantlon, J. F., & Piantadosi, S. T. (2018). A threshold-free model of numerosity comparisons. *PloS one*, *13*(4), e0195188.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision*, *7*(13), 14–14.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Anobile, G., Guerrini, G., Burr, D., Monti, M., Del Lucchese, B., & Cicchini, G. (2019). Spontaneous perception of numerosity in pre-school children. *Proceedings of the Royal Society B*, *286*(1906), 20191245.

Anobile, G., Burr, D. C., Iaia, M., Marinelli, C. V., Angelelli, P., & Turi, M. (2018). Independent adaptation mechanisms for numerosity and size perception provide evidence against a common sense of magnitude. *Scientific reports*, *8*(1), 1–12.

Anobile, G., Castaldi, E., Moscoso, P. A. M., Burr, D. C., & Arrighi, R. (2020). “groupitizing”: A strategy for numerosity estimation. *Scientific Reports*, *10*(1), 1–9.

Anobile, G., Cicchini, G. M., & Burr, D. C. (2014). Separate mechanisms for perception of numerosity and density. *Psychological science*, *25*(1), 265–270.

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*(1-2), 5–31.

Arrighi, R., Togoli, I., & Burr, D. C. (2014). A generalized sense of number. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1797), 20141791.

Atkinson, J., Campbell, F. W., & Francis, M. R. (1976). The magic number 4±0: A new look at visual numerosity judgements. *Perception*, *5*(3), 327–334.

Aulet, L. S., & Lourenco, S. F. (2021). Numerosity and cumulative surface area are perceived holistically as integral dimensions. *Journal of Experimental Psychology: General*, *150*(1), 145.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological science*, *18*(7), 622–628.

Barnard, A. M., Hughes, K. D., Gerhardt, R. R., DiVincenti Jr, L., Bovee, J. M., & Cantlon, J. F. (2013). Inherently analog quantity representations in olive baboons (papio anubis). *Frontiers in Psychology*, *4*, 253.

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

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological review*, *127*(5), 891.

Bonny, J. W., & Lourenco, S. F. (2013). The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of experimental child psychology*, *114*(3), 375–388.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85.

Bugden, S., & Ansari, D. (2016). Probing the nature of deficits in the 'approximate number system' in children with persistent developmental dyscalculia. *Developmental science*, *19*(5), 817–833.

Burr, D. C., & Ross, J. (2008). A visual sense of number. *Current biology*, *18*(6), 425–428.

Burr, D. C., Turi, M., & Anobile, G. (2010). Subitizing but not estimation of numerosity requires attentional resources. *Journal of Vision*, *10*(6), 20–20.

Cai, Y., Hofstetter, S., van Dijk, J., Zuiderbaan, W., van der Zwaag, W., Harvey, B. M., & Dumoulin, S. O. (2021). Topographic numerosity maps cover subitizing and estimation ranges. *Nature communications*, *12*(1), 1–10.

Cantlon, J. F. (2012). Math, monkeys, and the developing brain. *Proceedings of the National Academy of Sciences*, *109*(Supplement 1), 10725–10732.

Cantlon, J. F., & Brannon, E. M. (2007). Basic math in monkeys and college students. *PLoS biology*, *5*(12), e328.

Cantlon, J. F., Piantadosi, S. T., Ferrigno, S., Hughes, K. D., & Barnard, A. M. (2015). The origins of counting algorithms. *Psychological science*, *26*(6), 853–865.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Caviola, S., Colling, L. J., Mammarella, I. C., & Szűcs, D. (2020). Predictors of mathematics in primary school: Magnitude comparison, verbal and spatial working memory measures. *Developmental Science*, e12957.

Cheung, P., Rubenson, M., & Barner, D. (2017). To infinity and beyond: Children generalize the successor function to all possible numbers years after learning to count. *Cognitive psychology*, *92*, 22–36.

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences*, *116*(36), 17729–17734.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour*, *4*(12), 1265–1272.

Cheyette, S. J., Wu, S., & Piantadosi, S. (2021). The psychophysics of number arise from resource-limited spatial memory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Choo, H., & Franconeri, S. (2014). Enumeration of small collections violates weber's law. *Psychonomic bulletin & review*, *21*(1), 93–99.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2016). Spontaneous perception of numerosity in humans. *Nature communications*, *7*(1), 1–7.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Cowey, A., & Rolls, E. (1974). Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, *21*(5), 447–454.

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Morgan, M. J., et al. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences*, *108*(49), 19552–19557.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press, USA.

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

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience*, *5*(4), 390–407.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

DeWind, N. K., Bonner, M. F., & Brannon, E. M. (2020). Similarly oriented objects appear more numerous. *Journal of Vision*, *20*(4), 4–4.

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: Effects of feedback and training. *Frontiers in human neuroscience*, *6*, 68.

Ditz, H. M., & Nieder, A. (2016). Sensory and working memory representations of small and large numerosities in the crow endbrain. *Journal of Neuroscience*, *36*(47), 12044–12052.

Dotan, D., & Dehaene, S. (2016). On the origins of logarithmic number-to-position mapping. *Psychological review*, *123*(6), 637.

Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, *18*(18), R855–R856.

Elmore, L. C., Ma, W. J., Magnotti, J. F., Leising, K. J., Passaro, A. D., Katz, J. S., & Wright, A. A. (2011). Visual short-term memory compared in rhesus monkeys and humans. *Current Biology*, *21*(11), 975–979.

Farah, M. J., Hammond, K. M., Levine, D. N., & Calvanio, R. (1988). Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive psychology*, *20*(4), 439–462.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307–314.

Feigenson, L., Libertus, M. E., & Halberda, J. (2013). Links between the intuitive sense of number and formal mathematics ability. *Child development perspectives*, *7*(2), 74–79.

Ferrigno, S., Jara-Ettinger, J., Piantadosi, S. T., & Cantlon, J. F. (2017). Universal and uniquely human factors in spontaneous number perception. *Nature communications*, *8*, 13968.

Fischer, B., & Weber, H. (1993). Express saccades and visual attention. *Behavioral and Brain Sciences*, *16*(3), 553–567.

Fornaciai, M., Cicchini, G., & Burr, D. (2016). Adaptation to number operates on perceived rather than physical numerosity. *Cognition*, *151*, 63–67.

Forsyth, D. M., & Chapanis, A. (1958). Counting repeated light flashes as a function of their number, their rate of presentation, and retinal location stimulated. *Journal of Experimental Psychology*, *56*(5), 385.

Friendly, M. (2021). *Histdata: Data sets from the history of statistics and data visualization* [R package version 0.8-7]. https://CRAN.R-project.org/package=HistData

Fuhs, M. W., & McNeil, N. M. (2013). Ans acuity and mathematics ability in preschoolers from low-income homes: Contributions of inhibitory control. *Developmental science*, *16*(1), 136–148.

Gallistel, C. R. (1990). *The organization of learning.* The MIT Press.

Gallistel, C. R. (2018). Finding numbers in the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170119.

Gallistel, C. R., & Gelman, R. (1991). *Subitizing: The preverbal counting process.* Erlbaum Hillsdale, NJ.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1-2), 43–74.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59–65.

Gebuis, T., Kadosh, R. C., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta psychologica*, *171*, 17–35.

Gebuis, T., & Reynvoet, B. (2012a). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, *141*(4), 642.

Gebuis, T., & Reynvoet, B. (2012b). The role of visual information in numerosity estimation. *PloS one*, *7*(5), e37426.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771–781.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Harvard University Press.

Ginsburg, N. (1976). Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual and motor skills*, *43*(2), 663–668.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534.

Green, C. S., & Bavelier, D. (2006). Enumeration versus multiple object tracking: The case of action video game players. *Cognition*, *101*(1), 217–245.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

Gross, H. J., Pahl, M., Si, A., Zhu, H., Tautz, J., & Zhang, S. (2009). Number-based visual generalisation in the honeybee. *PloS one*, *4*(1), e4263.

Guillaume, M., & Gevers, W. (2016). Assessing the approximate number system: No relation between numerical comparison and estimation tasks. *Psychological research*, *80*(2), 248–258.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). Psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, *44*(5), 1457.

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., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665.

Hamilton, W. (1859). *Lectures on metaphysics and logic* (Vol. 1). Gould; Lincoln.

Heng, J. A., Woodford, M., & Polania, R. (2020). Efficient sampling and noisy decisions. *Elife*, *9*, e54962.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of experimental child psychology*, *103*(1), 17–29.

Hyde, D. C., Boas, D. A., Blair, C., & Carey, S. (2010). Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants. *Neuroimage*, *53*(2), 647–652.

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition*, *131*(1), 92–107.

Hyde, D. C., & Mou, Y. (2016). Neural and behavioral signatures of core numerical abilities and early symbolic number development. *Development of mathematical cognition* (pp. 51–77). Elsevier.

Hyde, D. C., & Spelke, E. S. (2011). Neural signatures of number processing in human infants: Evidence for two core systems underlying numerical cognition. *Developmental science*, *14*(2), 360–371.

Im, H. Y., Zhong, S.-h., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision research*, *126*, 291–307.

Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: How are approximate number system representations formed? *Cognition*, *129*(1), 63–69.

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica*, *145*, 147–155.

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, *17*(23), R1004–R1005.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221–1247.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature*, *3*, 281–282.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498–525.

Keshvari, S., Van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS computational biology*, *9*(2), e1002927.

Kim, G., Jang, J., Baek, S., Song, M., & Paik, S.-B. (2021). Visual number sense in untrained deep neural networks. *Science Advances*, *7*(1), eabd6127.

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and instruction*, *25*, 95–103.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438.

Lechelt, E. C. (1975). Temporal numerosity discrimination: Intermodal comparisons revisited. *British Journal of Psychology*, *66*(1), 101–108.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental science*, *14*(6), 1292–1300.

Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Is approximate number precision a stable predictor of math ability? *Learning and individual differences*, *25*, 126–133.

Lindskog, M., & Winman, A. (2016). No evidence of learning in non-symbolic numerical tasks–a comment on park and brannon (2014). *Cognition*, *150*, 243–247.

Lourenco, S. F. (2015). On the relation between numerical and non-numerical magnitudes: Evidence for a general magnitude system. *Mathematical cognition and learning* (pp. 145–174). Elsevier.

Lourenco, S. F., & Longo, M. R. (2010). General magnitude representation in human infants. *Psychological Science*, *21*(6), 873–881.

Lourenco, S. F., & Longo, M. R. (2011). *Origins and development of generalized magnitude representation*. Elsevier.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279.

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental science*, *17*(5), 714–726.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347.

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

Mazzocco, M. M., Feigenson, L., & Halberda, J. (2011). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLoS one*, *6*(9), e23749.

McComb, K., Packer, C., & Pusey, A. (1994). Roaring and numerical assessment in contests between groups of female lions, panthera leo. *Animal Behaviour*, *47*(2), 379–387.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological science*, *18*(8), 740–745.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 320.

Melcher, D., & Piazza, M. (2011). The role of attentional priority and saliency in determining capacity limits in enumeration and visual working memory. *PloS one*, *6*(12), e29296.

Mou, Y., Berteletti, I., & Hyde, D. C. (2018). What counts in preschool number knowledge? a bayes factor analytic approach toward theoretical model development. *Journal of experimental child psychology*, *166*, 116–133.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual review of neuroscience*, *32*, 185–208.

Nieder, A., Diester, I., & Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate parietal cortex. *Science*, *313*(5792), 1431–1435.

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, *297*(5587), 1708–1711.

Nieder, A., & Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *Journal of Neuroscience*, *27*(22), 5986–5993.

Odic, D., & Halberda, J. (2015). Eye movements reveal distinct encoding patterns for number and cumulative surface area in random dot arrays. *Journal of vision*, *15*(15), 5–5.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, *14*(4), 481–487.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological review*, *120*(2), 297.

O'Shaughnessy, D. M., Gibson, E., & Piantadosi, S. T. (2021). The cultural origins of symbolic number. *Psychological review*. https://www.proquest.

com / scholarly - journals / cultural - origins - symbolic - number / docview / 2541962021/se-2?accountid=14496

Park, J., Bermudez, V., Roberts, R. C., & Brannon, E. M. (2016). Non-symbolic approximate arithmetic training improves math performance in preschoolers. *Journal of Experimental Child Psychology*, *152*, 278–293.

Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological science*, *24*(10), 2013–2019.

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition*, *133*(1), 188–200.

Passolunghi, M. C., Lanfranchi, S., Altoé, G., & Sollazzo, N. (2015). Early numerical abilities and cognitive skills in kindergarten children. *Journal of Experimental Child Psychology*, *135*, 25–42.

Petrazzini, M. E. M., Mantese, F., & Prato-Previde, E. (n.d.). Food quantity discrimination in puppies (canis lupus familiaris). *Animal Cognition*, *23*(2).

Piantadosi, S. T. (2012). Kelpy: A free library for child experimentation in python.

Piantadosi, S. T. (2016). A rational analysis of the approximate number system. *Psychon Bull Rev*, *23*, 877–886. https://doi.org/10.3758/s13423-015-0963-8

Piantadosi, S. T., & Cantlon, J. F. (2017). True numerical cognition in the wild. *Psychological science*, *28*(4), 462–469.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147–153.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an amazonian indigene group. *Science*, *306*(5695), 499–503.

Plaisier, M. A., Tiest, W. M. B., & Kappers, A. M. (2009). One, two, three, many–subitizing in active touch. *Acta psychologica*, *131*(2), 163–170.

Platt, J. R., & Johnson, D. M. (1971). Localization of position within a homogeneous behavior chain: Effects of error contingencies. *Learning and Motivation*, *2*(4), 386–414.

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*(1), 50–57.

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological science*, *19*(6), 607–614.

Ross, J., & Burr, D. (2012). Number, texture and crowding. *Trends in Cognitive Sciences*, *16*(4), 196–197.

Ross, J., & Burr, D. C. (2010). Vision senses number directly. *Journal of vision*, *10*(2), 10–10.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662–674.

Schneider, R. M., Feiman, R., Mendes, M. A., & Barner, D. (2021). Pragmatic impacts on children's understanding of exact equality. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Schneider, R. M., Pankonin, A., Schachner, A., & Barner, D. (2021). Starting small: Exploring the origins of successor function knowledge. *Developmental Science*, e13091.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193–1216.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, *119*(4), 807.

Sokolowski, H. M., Fias, W., Ononye, C. B., & Ansari, D. (2017). Are numbers grounded in a general magnitude processing system? a functional neuroimaging meta-analysis. *Neuropsychologia*, *105*, 50–69.

Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, *36*(2), 97–127.

Starr, A., DeWind, N. K., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition*, *168*, 222–233.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013a). Infants show ratio-dependent number discrimination regardless of set size. *Infancy*, *18*(6), 927–941.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013b). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, *110*(45), 18116–18120.

Stoianov, I., & Zorzi, M. (2012). Emergence of a'visual number sense'in hierarchical generative models. *Nature neuroscience*, *15*(2), 194.

Stone, J. V. (2018). *Principles of neural information theory*. Sebtel Press.

Strandburg-Peshkin, A., Farine, D. R., Couzin, I. D., & Crofoot, M. C. (2015). Shared decision-making drives collective movement in wild baboons. *Science*, *348*(6241), 1358–1361.

Szkudlarek, E., Park, J., & Brannon, E. M. (2021). Failure to replicate the benefit of approximate arithmetic training for symbolic arithmetic fluency in adults. *Cognition*, *207*, 104521.

Szűcs, D., & Myers, T. (2017). A critical analysis of design, facts, bias and inference in the approximate number system training literature: A systematic review. *Trends in Neuroscience and Education*, *6*, 187–203.

Testolin, A., Dolfi, S., Rochus, M., & Zorzi, M. (2020). Visual sense of number vs. sense of magnitude in humans and machines. *Scientific reports*, *10*(1), 1–13.

Testolin, A., & McClelland, J. L. (2021). Do estimates of numerosity really adhere to weber's law? a reexamination of two case studies. *Psychonomic Bulletin & Review*, *28*(1), 158–168.

Testolin, A., Zou, W. Y., & McClelland, J. L. (2020). Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental science*, *23*(5), e12940.

Tomonaga, M., & Matsuzawa, T. (2002). Enumeration of briefly presented items by the chimpanzee (pan troglodytes) and humans (homo sapiens). *Animal Learning & Behavior*, *30*(2), 143–157.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Treisman, M. (1964). Noise and weber's law: The discrimination of brightness and other dimensions. *Psychological review*, *71*(4), 314.

Trick, L. M., & Enns, J. T. (1997). Clusters precede shapes in perceptual organization. *Psychological Science*, *8*(2), 124–129.

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*(1), 80.

Uller, C., Jaeger, R., Guidry, G., & Martin, C. (2003). Salamanders (plethodon cinereus) go for more: Rudiments of number in an amphibian. *Animal Cognition*, *6*(2), 105–112.

Van den Berg, R., Lindskog, M., Poom, L., & Winman, A. (2017). Recent is more: A negative time-order effect in nonsymbolic numerical judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1084.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785.

Verghese, P., & Pelli, D. G. (1992). The information capacity of visual attention. *Vision research*, *32*(5), 983–995.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of cognitive neuroscience*, *16*(9), 1493–1504.

Vul, E., Frank, M. C., Tenenbaum, J. B., & Alvarez, G. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 1955–1963.

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition*, *119*(1), 10–22.

Wang, J. J., Odic, D., Halberda, J., & Feigenson, L. (2016). Changing the precision of preschoolers' approximate number system representations changes their symbolic math performance. *Journal of Experimental Child Psychology*, *147*, 82–99.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae...* CF Koehler.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual review of psychology*, *69*, 105–129.

Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind, Brain, and Education*, *3*(4), 224–234.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601.

Wynn, K. (1992a). Addition and subtraction by human infants. *Nature*, *358*(6389), 749.

Wynn, K. (1992b). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220–251.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.

Yang, T.-I., & Chiao, C.-C. (2016). Number sense and state-dependent valuation in cuttlefish. *Proc. R. Soc. B*, *283*(1837), 20161379.

Yousif, S. R., & Keil, F. C. (2019). The additive-area heuristic: An efficient but illusory means of visual area approximation. *Psychological Science*, *30*(4), 495–503.

Zorzi, M., & Testolin, A. (2018). An emergentist perspective on the origin of number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170043.