A rational analysis of the approximate number system

Steven T. Piantadosi
Department of Brain and Cognitive Sciences, University of Rochester

Abstract

It is well-known in numerical cognition that higher numbers are represented with less absolute fidelity than lower numbers, often formalized as a logarithmic mapping. Previous derivations of this psychological law have worked by assuming that *relative* change in physical magnitude is the key psychologically-relevant measure (Fechner, 1860; Sun, Wang, Goyal, & Varshney, 2012; Portugal & Svaiter, 2011). Ideally, however, this property of psychological scales would be derived from more general, independent principles. This paper shows that a logarithmic number line is the one which minimizes the error between input and representation, relative to the probability that subjects would need to represent each number. This need probability is measured here through natural language and matches the form of need probabilities found in other literatures. The derivation does not presuppose anything like Weber's law, and makes minimal assumptions about both the nature of internal representations and the form of the mapping. The results prove in a general setting that the optimal psychological scale will change with the square root of the probability of each input. For stimuli that follow a power-law need distribution this approach recovers either a logarithmic or power-law psychophysical mapping (Stevens, 1957, 1961, 1975).

A fundamental challenge faced by cognitive agents in the world is that of mapping observable stimuli to internal representations. In human and animal cognition such mappings are mathematically regular, following a systematic relationship between stimulus and representation. This mapping is perhaps most studied in the case of the approximate number system (Dehaene, 1997), which is used to form non-exact representations of discrete quantities. Notably, this system of numerical representation is found across ontogeny (Xu & Spelke, 2000; Lipton & Spelke, 2003; Xu, Spelke, & Goddard, 2004; Xu & Arriaga, 2007; Feigenson, Dehaene, & Spelke, 2004; Halberda & Feigenson, 2008; Carey, 2009; Cantlon, Safford, & Brannon, 2010), age (Halberda, Ly, Wilmer, Naiman, & Germine, 2012), culture (Pica, Lemer, Izard, & Dehaene, 2004; Dehaene, Izard, Spelke, & Pica, 2008; Frank, Everett, Fedorenko, & Gibson, 2008), and species (Brannon & Terrace, 2000; Emmerton, 2001; Cantlon & Brannon, 2007).

The approximate number system has been characterized two ways in prior literature. One formalization assumes that number gets mapped to a linear psychological scale, but the fidelity of representation decreases with increasing numerosity (Gibbon, 1977; Meck

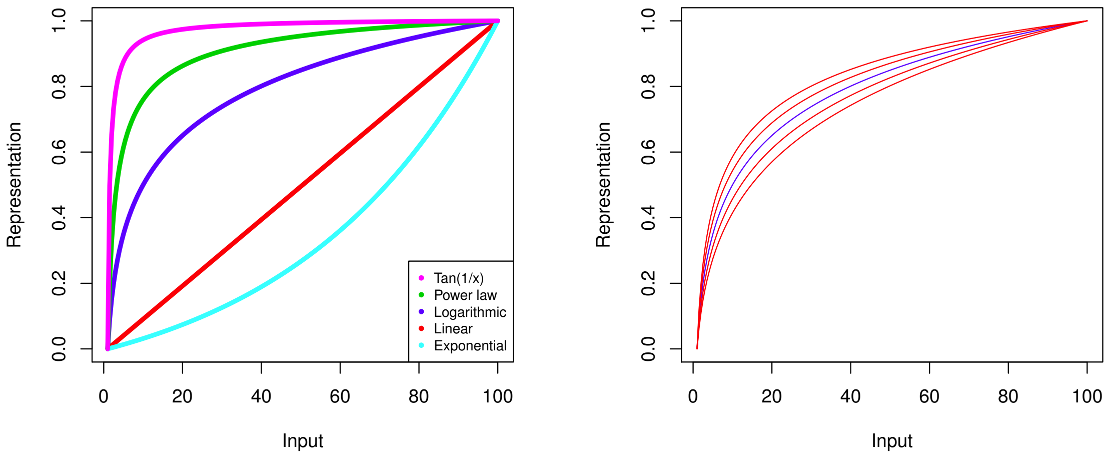
& Church, 1983; Whalen, Gallistel, & Gelman, 1999; Gallistel & Gelman, 1992). If, for instance, the standard deviation of a represented value n is proportional to n, this model can explain the ratio effect in which the confusability of x and y depends on x/y. An alternative to a linear mapping with variable noise is a *logarithmic* mapping with constant noise (Dehaene, 2003; Dehaene & Changeux, 1993). In this model, a number n is mapped to a representation $\psi(n)$ given by $\psi(n) \propto \log(n)$. Properties such as the psychological confusability of numbers are determined by distance in the logarithmically-transformed psychological space. Because $\log x - \log y = \log(x/y)$, this framework can also explain the ratio effect. Some work has argued for the neural reality of the logarithmic mapping by showing neural tuning curves that scale logarithmically (Nieder, Freedman, & Miller, 2002; Nieder & Miller, 2004; Nieder & Merten, 2007; Nieder & Dehaene, 2009), although other behavioral phenomena appear less well described by either model (Verguts, Fias, & Stevens, 2005).[1]

The present paper aims to investigate why cognitive systems may represent higher numbers with decreasing fidelity in the way that they do. To answer this question, we choose to analyze the logarithmic mapping in detail. Analysis is far more straightforward for the logarithmic model since it can be captured by considering a single function ψ. The linear model with scale variability cannot be as simply captured with a single one-dimensional function since one must explain both its linearity, the shape of its noise, and the change of noise with numerosity, all together presenting a much more complex analytical situation.[2] Therefore, we focus here on understanding how the logarithmic system might be derived, noting that the general method and approach developed here is potentially applicable in future work on the properties of the linear model, as well as other psychophysical domains.

In principle, the cognitive system supporting number could likely implement a large number of mappings (though see Luce, 1959). It is easy to imagine other possibilities, such as where the input stimuli are mapped to representational space according to other functions such as exponentials, power-laws, or polynomials. Several of these examples are shown in Figure 1a, where numbers $1, 2, \ldots, 100$ are mapped into a bounded psychological space, arbitrarily denoted $[0, 1]$. Here, the distance between numerosities in psychological space (difference along the y-axis) is meant to quantify measures such as confusability or generalization among particular representations (in the sense of Shepard, 1987). Thus, representations which are given high fidelity are further away from their neighbors; close numbers such as the higher numbers are more likely to be confused because they are nearby in psychological space. This figure illustrates the key challenge faced by a cognitive system: the psychological space for any organism is bounded—we do not have infinite representational capacities—so our cognitive system must trade-off fidelity among representations. One cannot increase fidelity for one stimulus without paying the cost of effectively decreasing

[1] Other work has criticized logarithmic scaling by arguing that it predicts nonlinear addition and subtraction (Stevens, 1960; Livingstone et al., 2014), since $\psi(x) + \psi(y) = \log x + \log y \neq x + y$. This critique erroneously assumes that addition in numerical space of $x + y$ *must* correspond to addition in the psychological space (e.g. $\psi(x) + \psi(y)$). In actuality, numerical addition can be correctly implemented in psychological space using other functions (e.g. f such that $f(\psi(x), \psi(y)) = \psi(x + y)$) and such functions have long been worked out in computer science (e.g. Swartzlander & Alexopoulos, 1975).

[2] A satisfying analysis might try to derive the *two*-dimensional function $g(x, v)$, giving the probability that a psychological representation of x would be at value v, where the conditional distribution of v given x is a Gaussian centered at x, with a width proportional to x.



(a) (b)

Figure 1. : (a) Several logically possible representation systems for approximate number with with $\psi(1) = 0$ and $\psi(100) = 1$. Each mapping takes an input cardinality (n) to a representation ($\psi(n)$). The real puzzle is not which of these five example curves is chosen, but which out of the infinite number of possible mappings is chosen. (b) Logarithmic and power-law mappings: the red lines represent power laws with $\alpha = 1.7, 1.85, 2.15, 2.3$. Power-law mappings (reds) closely approximate a logarithmic mapping (blue) for $\alpha \approx 2$.

it for another.

The logarithmic mapping in this setup makes higher numbers closer in psychological space (e.g. $|\log 98 - \log 99| < |\log 4 - \log 5|$), reserving higher fidelity for lower numbers. Other mappings would not have this property—for instance, the mapping $\psi(x) = \tan 1/x$ which reserves almost all fidelity for the very lowest cardinalities (e.g. 20 and 21 are about as confusable as 98 and 99), or the exponential curve in Figure 1a which actually reserves fidelity for *higher* cardinality (e.g. 98 and 99 are *less* confusable than 4 and 5).

The present work aims to explain why the mapping for numerosity appears to be at least close to logarithmic, building on prior derivations as well as work arguing for information processing explanations in perception more generally (e.g. Wainwright, 1999). The derivation presented here shows that a logarithmic mapping is an optimal psychological scaling for how people use numbers. The derivation is similar in spirit to Smith and Levy (2008), who derive the logarithmic relationship between reading time and probability in language comprehension. The present derivation shows that the logarithmic mapping is optimal relative to the probabilities with which different numerosities must be represented, the *need probabilities.* This approach differs from that undertaken previous by, for instance, Luce (1959), who created a set of psychophysical axioms—such as how the representation should behave under rescaling of the input—and studied the laws permitted by such a system. Building on recent work by Portugal and Svaiter (2011) and Sun et al. (2012), we formalize an optimization problem over a broad class of possible psychological functions, and show that solution of this optimization yields the logarithmic mapping in the case of number. In deriving the optimal representation for number, we show how power-law mappings may *also* be derived, for distributions similar in parametric form to the number need distribution. Thus, much as other accounts have attempted to unify or collapse logarithmic and power-law psychophysical functions (MacKay, 1963; Ekman, 1964; Wagenaar, 1975; Wasserman, Felsten, & Easland, 1979; Krueger, 1989; Sun et al., 2012), the present work demonstrates that both are optimal under different situations, starting from the same base assumptions. This work therefore provides an alternative to previous derivations of powers laws based on aggregation (Chater & Brown, 1999). We begin with an overview of previous derivations of the logarithmic mapping.

Previous derivations of the logarithmic mapping

In psychology, the most well-known derivation of the logarithmic mapping is due to Fechner (1860). Fechner started by assuming the validity of *Weber's law*, which holds that the just noticeable difference in a physical stimulus is proportional to the magnitude of the stimulus. Fechner then showed how this property of just noticeable differences gives rise to a logarithmic mapping, although his mathematics has been criticized (Luce & Edwards, 1958; Luce, 1962); more recent formalizations of Fechner's approach have used conceptually similar methods to the functional analysis we present here (Aczél, Falmagne, & Luce, 2000). As argued by Masin, Zudini, and Antonelli (2009), Fechner's approach of presupposing Weber's law has led to the persistent misperception that Weber's law is "the foundation rather than the implication" of the logarithmic mapping. Indeed, it makes much more sense to treat Weber's law as a description of behavior and to seek independent principles for explaining the psychological system that gives rise to this behavior.

In this spirit, Masin et al. (2009) review two alternative derivations which do not rely on Weber's law. One by Bernoulli (1738) predates Fechner by over a century and operates in the setting of subjective value; another, by Thurstone (1931) makes assumptions very similar to Bernoulli but is framed in terms of intuitive economic quantities like "motivation" and "satisfaction." While these examples importantly illustrate that a logarithmic mapping can be derived from principles other than Weber's law, they—without full justification—stipulate equations that lead directly to the desired outcome. Bernoulli, for instance, assumes that an incremental change in subjective space should depend inverse-proportionally on the total objective value; any other relationship would not have yielded a logarithmic mapping.

However, there is a more important disconnect between modern psychology and the derivations of Fechner, Thurstone, and Bernoulli. Their work predated an extremely valuable methodological innovation: *rational analysis* (e.g. Anderson & Milson, 1989; Anderson, 1990; Anderson & Schooler, 1991; Chater & Oaksford, 1999; Geisler, 2003). From the perspective of rational analysis, the question of why there is a logarithmic mapping can only be answered by considering the context and use of the approximate number system. Any derivation that does not take these factors into account is likely to be missing an important aspect of why our cognitive systems are the way they are. As it turns out, a logarithmic mapping is well-adapted *specifically* to the observed need probabilities of how often each cardinality must be represented or processed. A strong prediction of this type of rational approach is that the mapping to psychological space would be constructed differently (either throughout development or through evolutionary time) if we typically had to represent a different distribution of numerosities.

Recent work by Portugal and Svaiter (2011) and Sun et al. (2012) has moved studies towards an idealized rational analysis. Both studies assume that in representing a cardinality, the neural system maps an input number n to a quantized form $\hat{n}$, and that the "right" thing to do is minimize the expected value of the *relative error* of the quantization (Sun & Goyal, 2011), given by $\mathbb{E}_n\left[|n-\hat{n}|^2/n^2\right]$ (for history and related results, see Gray & Neuhoff, 1998; Cambanis & Gerr, 1983). This is not the same as assuming Weber's law, but rather assumes an objective function (over representations) that is based in relative error. Portugal and Svaiter (2011) show that a logarithmic mapping is the one that optimizes the worst-case relative quantization error. Sun et al. (2012) show that under a particular power law need distribution, the logarithmic mapping is the one which minimizes the expected relative quantization error.

Formulating the optimization problem in terms of *relative* error is very close—mathematically and conceptually—to assuming Weber's law from the start, since it takes for granted that what matters in psychological space are relative changes. Such an assumption is also how Stevens (1975) justified the power law psychophysical function. He wrote that a power law resulted from a cognitive system that—apparently—cares about relative changes in magnitude rather than absolute changes. Of course, such explanations are post-hoc. At best, they show that *if* an organism cares about relative changes, it will go with either a power law or logarithmic mapping. But why should an organism care about relative changes in the first places? One of our goals is to show from independent principles why relative changes might matter to a well-adapted psychological system.

We also aim to move beyond several other of the less desirable assumptions that Portugal and Svaiter (2011) and Sun et al. (2012) required. Their work is primarily formulated in terms of quantized representations (though see Sun et al., 2012, Appendix A), meaning that they assume that the appropriate neural representation is a discrete element or code. However, individual neurons are gradiently sensitive to numerosity (Nieder et al., 2002), meaning that a more biologically plausible analysis might consider the representational space to be continuous. Additionally, though Sun et al. (2012) show that logarithmic and power-law psychophysical functions can both be achieved by the same framework with slightly different parameters on the need distribution, they do not establish that the empirically-observed need distribution is the same one that leads to the logarithmic mapping. Indeed, the present results indicate that the most plausible input distribution does *not* lead to a logarithmic mapping under Sun et al.'s analysis.

Our goal is to address all of these limitations with a novel analysis. Most basically, we do not assume from the start that relative changes are what matter to the psychological system. Instead, we begin only from the assumption that cognitive systems must map numerosities from an external stimulus space to an internal representation space. We assume that the form of the mapping is optimized to avoid confusing the most frequently used representations. This setup allows us to derive a general law relating need probabilities to a mapping into psychological space: the rate of change of the mapping should be proportional to the square root of the need probability. The analysis shows that this derives exactly the logarithmic mapping for the need distribution of number, and more generally, a power law mapping (Stevens, 1957, 1961, 1975) for other stimuli which follow a power-law need distribution.

Optimal mappings into subjective space

We use ψ to denote the function mapping observable stimuli to internal representations. Thus, an input n will be mapped to a representation $\psi(n)$. For simplicity, we will assume that both the internal and external domains are continuous spaces. This can be justified by imagining that the representational system handles enough numbers that they well-approximate a continuous function—unless, of course, the system for approximate discrete numbers is identical with the system for continuous magnitude/extent.

The analysis requires some very basic properties of ψ (c.f. Luce, 1959): (i) the range of ψ must be bounded, (ii) ψ is monotonically increasing and (iii) ψ is twice-continuously differentiable. Boundedness comes from the assumption that psychological space has a limited representational capacity. For this, we assume that $\psi(n) \in [0,1]$ with $\psi(1) = 0$ and $\psi(M) = 1$, where M is the largest cardinality that people can represent. Monotonicity means that the mapping from external cardinality to internal number is "transparent," not requiring sophisticated computations, since a larger external magnitude always maps to a larger internal one. It also guarantees that the mapping will be invertible, so we can always tell what real-world numerosity a representation stands for. Finally, (iii) is a technical condition meaning that ψ is well behaved enough to have a well-defined rate of change (first derivative) and second derivative. This rules out, for instance, step functions with sharp corners. There are many functions that meet these criteria—including, for instance, all in Figure 1a—and our analysis aims to find the "best" ψ out of the infinitely many possible alternatives.

Figure 2 illustrates the setup. It cannot be the case that continuous representations are stored with perfect fidelity since that would require infinite information processing. Instead, we assume that any represented value $\psi(n)$ may be corrupted by representational noise from an arbitrary[3] distribution $\mathcal{N}$, independent of the value of $\psi(n)$. It is reasonable to suppose that $\mathcal{N}$ is, for instance, a Gaussian distribution as is observed in number and typical of noise, although our derivation does not require this. In our setup (Figure 2), an input cardinality n is mapped to a representation $\psi(n) \in [0,1]$. This value may be corrupted by noise to yield $\psi(n) + \epsilon$, where $\epsilon \sim \mathcal{N}$. We assume the noise $\mathcal{N}$ is constant over psychological space (ie. isotropic) in order see what properties arise without building in biasing factors into the structure of psychological space itself. A similarly uniform internal space is also an implicit assumption of psychological space models like that of (Shepard, 1987).

A rational goal for the system will then be to minimize the absolute difference between what a corrupted value represents (which is $\psi^{-1}(\psi(n)+\epsilon)$) and what we intended to represent (which is n). This expected difference is then,

$$\mathop{\mathbb{E}}_{n}\mathop{\mathbb{E}}_{\epsilon}\left|\psi^{-1}(\psi(n)+\epsilon) - n\right|. \tag{1}$$

[3] We require three technical requirements on $\mathcal{N}$: it must have a bounded absolute error, so that if $\epsilon \sim \mathcal{N}$, $\mathbb{E}_\epsilon |\epsilon| < \infty$, it must be independent of location in psychological space, and the typical error must be small relative to $1/\psi'(n)$.

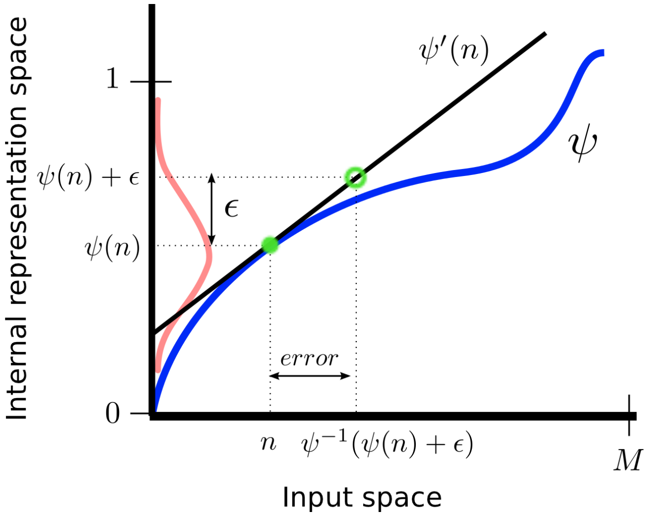


Figure 2. : The general setup of our analysis: an input n is mapped to a representation $\psi(n)$, which may then be corrupted by noise ϵ. We seek to minimize the amount by which this noise "matters" on the original input scale, given by $\psi^{-1}(\psi(n)+\epsilon)-n$. We compute $\psi^{-1}(n)$ using a linear approximation (see text).

Here, there is one expectation over n meaning that we should try to minimize the error for typical usage, thus more accurately representing the most frequently used numbers. There is also an expectation over ϵ, meaning that we try to minimize error, averaging over the uncertainty we have about how much the representations may be corrupted (as quantified by ϵ). Informally, by finding ψ to minimize (1), we are choosing a mapping into representational space such that when the represented values are altered by noise, the absolute amount in physical space that the change corresponds to is minimized.[4]

The difficulty with (1) is that it is stated in terms of ψ and its inverse, ψ^{-1}, making analytic analysis hard. We can, however, make a linear approximation to ψ near n (see Figure 2), and use the linear approximation to compute the inverse ψ' near $\psi(n)$. This approximation is valid so long as the noise ϵ is small, relative to $1/\psi'(n)$. In a linear approximation we use the differentiability (iii) of ψ and write

$$\psi(x)=\psi'(n)\cdot(x-n)+\psi(n), \qquad \text{for } x\approx n. \tag{2}$$

Then, the inverse function ψ^{-1} is

$$\psi^{-1}(x)=(x-\psi(n))\cdot\frac{1}{\psi'(n)}+n, \qquad \text{for } x\approx n. \tag{3}$$

Using this approximation, we can rewrite (1) as,

$$\begin{aligned}\underset{n}{\mathbb{E}}\,\underset{\epsilon}{\mathbb{E}}\left|\psi^{-1}(\psi(n)+\epsilon)-n\right| &= \underset{n}{\mathbb{E}}\,\underset{\epsilon}{\mathbb{E}}\left|(\psi(n)+\epsilon-\psi(n))\cdot\frac{1}{\psi'(n)}+n-n\right| \\ &= \underset{n}{\mathbb{E}}\,\frac{1}{\psi'(n)}\cdot\underset{\epsilon}{\mathbb{E}}\,|\epsilon|,\end{aligned} \tag{4}$$

[4] The following derivation therefore uses the norm $|\cdot|^p$ for $p=1$, but analogous derivations will work for others, giving rise to a different exponent at the end, including the squared error $p=2$.

where we have used the fact (ii) that ψ is monotonically increasing, so ψ' is positive. Writing out the expectation over n explicitly, this becomes,

$$\mathop{\mathbb{E}}_{\epsilon}\left[|\epsilon|\right] \cdot \int_1^M \frac{p(n)}{\psi'(n)} dn. \tag{5}$$

To summarize the derivation so far, we are seeking a function ψ mapping observed numbers into an internal representational space. Under a simple approximation that holds for relatively low internal noise, any potential ψ can be "scored" according to (5) to determine the amount by which noise corrupts the representation relative to the need distribution on numbers $p(n)$ and the internal noise ϵ.

To actually optimize (5), we first express the bound (i) in terms of ψ' rather than ψ. If $\psi(M) = 1$, then

$$\int_1^M \psi'(n) dn = 1. \tag{6}$$

Now we have stated an objective function (5) and a constraint (6) in terms of the rate of change of ψ, which is ψ'. It turns out that optimization of (5) subject to (6) over *functions* ψ is possible through the *calculus of variations* (see Fox, 2010; Gelfand & Fomin, 2000). This area of functional analysis can find minima or maxima over a space of functions exactly as standard calculus (or analysis) finds minima and maxima over variables (for similar applications of functional analysis to psychophysics, see Aczél et al., 2000). In our case, we write a *functional* $\mathcal{F}$—roughly, a function of functions[5]—that encodes our objective and constraints,

$$\mathcal{F}\left[\psi\right] = \mathop{\mathbb{E}}_{\epsilon}\left[|\epsilon|\right] \cdot \int_1^M \frac{p(n)}{\psi'(n)} dn + \lambda \left(\int_1^M \psi'(n) dn - 1 \right). \tag{7}$$

Equivalently,

$$\mathcal{F}\left[\psi\right] = \int_1^M \mathcal{L}(n, \psi(n), \psi'(n)) dn \tag{8}$$

where

$$\mathcal{L}(n, u, v) = \mathop{\mathbb{E}}_{\epsilon}\left[|\epsilon|\right] \cdot \frac{p(n)}{v} + \lambda \left(v - \frac{1}{M-1} \right). \tag{9}$$

This equation has added the constraint multiplied by the variable λ (providing the functional analysis analog the λ in the method of Lagrange Multipliers). Roughly, the λ allows us to combine objective function and constraints into a single equation whose partial derivatives can be used to compute the function ψ that maximizes $\mathcal{F}\left[\cdot\right]$ (for more on the theory behind this techinique, see Gelfand & Fomin, 2000).

The Euler-Lagrange equation solves the optimization in (8) over functions ψ, providing the optimal ψ by solving

$$\mathcal{L}_u(n, \psi(n), \psi'(n)) - \frac{d}{dn} \mathcal{L}_v(n, \psi(n), \psi'(n)) = 0, \tag{10}$$

[5] Other examples of functionals include, for instance, the functional for differential entropy, which takes a distribution and returns a number. Shannon (1948) for instance provides a simple functional analysis proof using similar techniques that normal distributions maximize entropy relative to a fixed mean and variance.

where $\mathcal{L}_u$ is the partial derivative of $\mathcal{L}$ with respect to its second argument, u, and $\mathcal{L}_v$ is the partial derivative of $\mathcal{L}$ with respect to its third argument, v. That is, (10) states that we compute two partial derivatives of $\mathcal{L}$ and evaluate them at the appropriate values (n, $\psi(n)$, and $\psi'(n)$), yielding a differential equation that must be solved to find the optimal ψ. In (10), $\mathcal{L}_u = 0$ since u does not appear in $\mathcal{L}$, and

$$\mathcal{L}_v(n, \psi(n), \psi'(n)) = -E_\epsilon\left[|\epsilon|\right] \cdot \frac{p(n)}{(\psi'(n))^2} + \lambda \tag{11}$$

so by (10) we seek a solution of

$$\frac{d}{dn}\left(E_\epsilon\left[|\epsilon|\right] \cdot \frac{p(n)}{(\psi'(n))^2} - \lambda\right) = 0. \tag{12}$$

Integrating both sides yields

$$E_\epsilon\left[|\epsilon|\right] \cdot \frac{p(n)}{(\psi'(n))^2} - \lambda = C \tag{13}$$

for some constant C, meaning that

$$\psi'(n) = \sqrt{\frac{p(n) \cdot E_\epsilon\left[|\epsilon|\right]}{C + \lambda}}. \tag{14}$$

Here, λ is chosen to satisfy the bound in (6), so the constants are essentially irrelevant. More simply, then, we can write the optimal ψ as satisfying,

$$\psi'(n) \propto \sqrt{p(n)}. \tag{15}$$

This result indicates that the optimal mapping in terms of minimizing error relative to the need probabilities makes the internal scale *change* proportional to the square root of the need probability $p(n)$.

Logarithmic and power-law mappings are optimal for power-law needs

The previous section showed that the optimal mapping into psychological space is proportional to the square root of the need distribution $p(n)$. In most cases, such as those reviewed by Stevens (1975), the need distribution $p(n)$ is not so clear: how often people need to encode the particular heaviness or velocity of a stimulus? Sun et al. (2012) examine the case of loudness and that the need distribution appears roughly log-normal or power-law distributed in intensity, or normally-distributed in decibels[6]. In the case of number, however, there is a clear way to measure the need probability: we can look at how often typical speakers of a language encode specific cardinalities as measured by number word frequencies. This provides a plausible measure for how often cognitive processing mechanisms need to *exactly* encode each number.

[6] Log-normal distributions are notoriously hard to distinguish from power laws (Malevergne, Pisarenko, & Sornette, 2011) and result from similar generative processes (see Mitzenmacher, 2004)

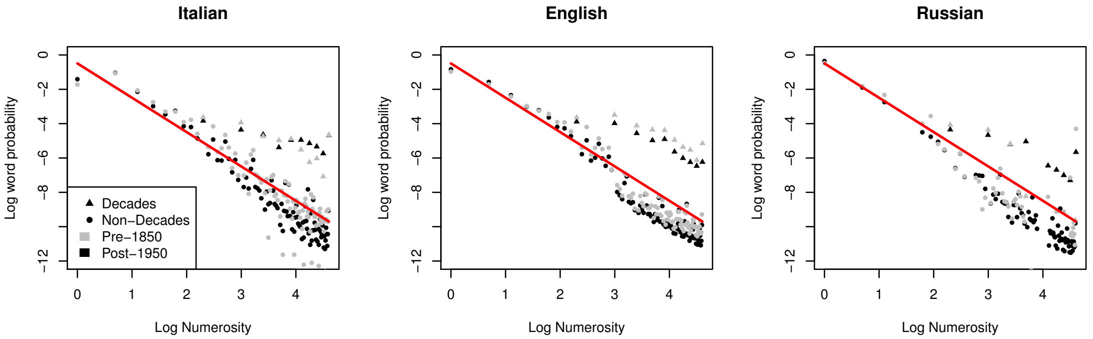


Figure 3. : The distribution of number word frequencies across Italian, English, and Russian according to the Google Books N-gram dataset (Li et al. 2012). This reveals a strong power-law distribution across time, language, and for both decades ("ten", "twenty", etc.) and non-decades. On these plots, the *linear* trend of the data corresponds to the *exponent* in the power law distribution. The red line shows a power-law distribution with $\alpha = 2$.

Figure 3 shows the distribution of number words in the Google Books N-gram dataset (Lin et al., 2012) on a log-log plot for three relatively unrelated languages, Italian, English, and Russian. First, the overall data trend is linear on this log-log-plot, indicating that number words follow something close to a *power law* distribution[7] (Newman, 2005):

$$p(n) \propto n^{-\alpha} \tag{16}$$

for some α. This type of power law distribution is famously observed more generally in word frequencies (Zipf, 1936, 1949), although the cause of these frequencies is still unknown (Piantadosi, 2014).

There is one important point about the particular power law observed: the exponent—corresponding to the linear slope in the log-log plot—is very close to $\alpha = 2$. The actual exponent found by a fit depends strongly on the details of fitting (see Newman, 2005)—in particular, how apparent outliers like "one" in Italian, and the decades are treated. Rather than obsess over the details of fitting, we have simply shown a power-law distribution with $\alpha = 2$ in red, showing that the trend of the data across languages and historical time, as well as for decades and non-decades, closely approximates this particular exponent $\alpha = 2$. Both this general pattern in number word distribution and the exponent $\alpha \approx 2$ accord with a more thorough cross-linguistic analysis undertaken by Dehaene and Mehler (1992). They show that across languages, this type of distribution generally holds, although there are interesting complications for numbers like unlucky 13 in some languages, or decades. The power law exponent that they report is $\alpha = 1.9$. This empirical fact about number usage could be called the *inverse square law for number frequency.*

The detailed patterns exhibited by these plausible number word need probabilities are also interesting. For instance the "decades" ("ten", "twenty", "thirty", etc.) have substantially higher probability than non-decades of similar magnitude, likely due to approximate

[7] A power law is linear on a log-log plot: the log-log linearity implies that $\log p(n) = C - \alpha \cdot \log n$ for some C, α, so $p(n) \propto n^{-\alpha}$.

usage, as originally noted by Dehaene and Mehler (1992). Additionally, in English words over 20 ($\log 20 \approx 3$) are somewhat less probable than might be expected by the frequency of the teens, although it is not clear whether this is somehow a corpus/text artifact of these words typically being written with a hyphen. Interestingly, even within these types of deviations, the decades, non-decades, and English words over 20 all follow a power law with exponent roughly 2, as evidenced by their slope similar to the red line's slope. Importantly, if the need probabilities $p(n)$ reflect "real" needs—not, for instance, some artifact of modern culture—they should be observed throughout historical time. The gray points show data for books published before 1850, demonstrating an effectively identical distribution. Indeed, the Spearman correlation of individual number word frequencies across the two time points is 0.96 in Italian ($p \ll 0.001$), 0.98 in English ($p \ll 0.001$), and 0.88 in Russian ($p \ll 0.001$), indicate a strong tendency for consistent usage or need.

The significance of the exponent $\alpha = 2$ is that it predicts precisely the logarithmic mapping when the optimal ψ is found by solving equation (15). In general, when $p(n)$ is a power law, the optimal mapping from (15) becomes

$$\psi'(n) \propto n^{-\alpha/2}. \tag{17}$$

So, when $\alpha = 2$,

$$\psi'(n) \propto \frac{1}{n}. \tag{18}$$

Integrating ψ' yields

$$\psi(n) \propto \log(n) + C, \tag{19}$$

a logarithmic mapping[8]. This explains why the mapping for number is at least approximately logarithmic[9]. Alternatively, when $\alpha \neq 2$, (17) becomes

$$\psi(n) \propto n^{-\alpha/2+1}, \tag{20}$$

yielding the power law psychophysical functions argued for by (Stevens, 1957, 1961, 1975).

What we have shown, then, is that the power-law *and* logarithmic mapping both fall out of the same analysis, resulting from different exponents on the need distribution. This reveals a deep connection between these two psychophysical laws: both are optimal under different exponents of the same form of distribution (as also argued by Sun et al., 2012). Indeed, because of the similarity between log-normal distributions and power-law distributions, we should expect functions very much like power-law mappings for a log-normal $p(n)$. Previously, this need distribution has been used to explain properties of number in other cognitive paradigms (Verguts et al., 2005).

Critically, the present analysis rests on the assumption that natural language word frequencies provide an accurate "need" distribution for how often each number must be represented.[10] To fully explain number, $p(n)$ must be a power law across evolutionary time

[8] Note that since we scale and shift the function so that it lies in psychological space $[0, 1]$, the constant C can be ignored.

[9] It is useful to compare this finding with Sun et al. (2012), who argue that the best representation will scale with $n^{-(\alpha-1)/3-1}$. So in their analysis they require an exponent of $\alpha = 1$ to recover the logarithmic function. This exponent does not match with the above data suggesting that for number, $\alpha \approx 2$.

[10] Note that Sun et al. (2012) find similar results in their derivation, although the exponent they require for logarithmic mapping is $\alpha = 1$ rather than the above empirically observed $\alpha \approx 2$.

and likely developmental time. Notably, there independent reasons beyond these corpus results why need probabilities are likely power law distribution. Anderson and Schooler (1991) find power laws need distributions across several domains in human memory; power laws are also very generally found in complex systems, resulting from a wide variety of statistical processes (see Mitzenmacher, 2004). In general, though, further work will be required to test the need distribution in common environments and determine whether this need distribution under (15) adequately models representational mappings.

This analysis did not require any strong assumptions about the error distribution $\mathcal{N}$ in psychological space. This is important because one might expect that the noise is generated according to other optimizing principles which may vary by domain. Indeed, our approach is consistent both with the approximately Gaussian noise observed for number (Nieder et al., 2002; Nieder & Miller, 2004; Nieder & Merten, 2007; Nieder & Dehaene, 2009), and even other kinds of confusability/generalization gradients such as exponentials also found throughout cognition (Shepard, 1987; Chater & Vitányi, 2003). In the case of Gaussian psychological noise, this setup recovers Weber's law and the standard psychophysics of number.

Curiously, one interpretation of these findings is that if the mapping is optimized as we suggest, it is very unlikely that the mapping is truly logarithmic: α is almost surely not *exactly* equal to 2, so the optimal mapping is almost certain to be a power law. However, these possibilities are not different in any interesting sense: Figure 1b shows the logarithmic mapping and power law mappings, appropriately bounded as in (i), for α near 2. What these illustrate is that the optimization we describe is "continuous" in that small changes to α do not lead to large changes in ψ, even though the written form of the function changes. In this sense, it is not a productive question to study whether the law is truly logarithmic or truly a power law, because the two are just part of the same continuum of functions. This may also be true in other perceptual domains.

Conclusion

It is worth summarizing the results in general terms. We imagine that an input n is mapped to a psychological representation $\psi(n)$, which may then be corrupted by noise, to give a corrupted representation $\psi(n) + \epsilon$. In physical space, the amount by which this noise ϵ "matters" can be quantified by $1/\psi'(n)$, one over the rate of change (ψ') of ψ at n. Our analysis sought to minimize the average effect of this noise, subject to bounded representational resources. When this optimization is performed over a wide range of functions ψ, we find that ψ should *change* according to the square root of the need probability $p(n)$ in order to minimize the effect of errors. This is a general fact about the optimal psychological mapping.

A plausible need distribution for number robustly follows a power-law need distribution, with a particular exponent $\alpha = 2$ such that making ψ change according to the square root of $p(n)$ yields a logarithmic mapping. Other domains that follow a power law need distributions with $\alpha \neq 2$ will give rise to power-law mappings. In the case of number, these results explain *why* relative changes in magnitude (e.g. Weber's law) are what matter psychologically: a psychological system has bounded representational resources and, subject to this constraint, the system with minimal absolute error uses a logarithmic mapping. Thus, unlike derivations where the logarithm comes from assuming relative changes are the relevant ones (Fechner, 1860; Sun et al., 2012; Portugal & Svaiter, 2011) or those explaining the law from lower-level architectural considerations (Stoianov & Zorzi, 2012), the present results derive this fact from a rational analysis of effective information processing. The logarithm arises in our approach because of the particular need distribution actually observed for number; a different distribution would have resulted in a different optimized mapping and the general form of this optimization is provided in (15).

As such, this approach is in principle applicable to other psychophysical domains such as brightness, loudness, and weight (Stevens, 1957). The challenge is that in these domains the need distribution is not as easily quantified. For acoustic loudness, Sun et al. (2012) show that their derivation recovers a plausible, near-logarithmic psychophysical function from a log-normal need distribution, and numerically solving (15) for log-normal distributions yields similar relationships for the analysis[11]. This indicates that this approach of optimizing functional mappings in the way we describe may plausibly explain psychophysics of other modalities, once future work determines plausible need distributions across these domains. In general, then, the results illustrate how core systems of representation (Feigenson et al., 2004; Carey, 2009) may be highly-tuned to environmental pressures and functional optimization over the course of evolutionary or developmental time.

Acknowledgments

I'd like to thank Dan Roy for providing crucial comments and advice on this work. Ted Gibson, Josh Tenenbaum, and Liz Spelke also provided important feedback on this and earlier versions of this paper. Three reviewers provided extremely helpful comments on this paper. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number F32HD070544. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[11] As standard symbolic math programs cannot perform the required integration for log-normal distributions, it is likely that the optimal mapping for log-normal distributions has no closed form. The non-existence of a closed-form solution may be possible to prove with, for instance, Liouville's theorem, but no proof is attempted here.

References

Aczél, J., Falmagne, J., & Luce, R. (2000). Functional equations in the behavioral sciences. *Math. Japonica*, *52*(3), 469–512.

Anderson, J. (1990). *The adaptive character of thought.* Lawrence Erlbaum.

Anderson, J., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

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

Bernoulli, D. (1738). *Specimen theoriae novae de mensura sortis* (Vol. 5).

Brannon, E., & Terrace, H. (2000). Representation of the numerosities 1–9 by rhesus macaques. *Journal of Experimental Psychology: Animal Behavior Processes*, *26*(1), 31.

Cambanis, S., & Gerr, N. L. (1983). A simple class of asymptotically optimal quantizers. *Information Theory, IEEE Transactions on*, *29*(5), 664–676.

Cantlon, J., & Brannon, E. (2007). Basic math in monkeys and college students. *PLOS Biology*, *5*(12), e328.

Cantlon, J., Safford, K., & Brannon, E. (2010). Spontaneous analog number representations in 3-year-old children. *Developmental science*, *13*(2), 289–297.

Carey, S. (2009). *The Origin of Concepts.* Oxford: Oxford University Press.

Chater, N., & Brown, G. D. (1999). Scale-invariance as a unifying psychological principle. *Cognition*, *69*(3), B17–B24.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Chater, N., & Vitányi, P. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369.

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

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145–147.

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

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, *320*(5880), 1217–1220.

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

Ekman, G. (1964). Is the power law a special case of Fechner's law? *Perceptual and Motor Skills*, *19*(3), 730–730.

Emmerton, J. (2001). Birds' judgments of number and quantity. *Avian visual cognition.*

Fechner, G. (1860). *Elemente der psychophysik.* Breitkopf & Härtel: Leipzig.

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

Fox, C. (2010). *An introduction to the calculus of variations.* Dover Publications.

Frank, M., Everett, D., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, *108*(3), 819–824.

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

Geisler, W. (2003). Ideal observer analysis. *The visual neurosciences*, 825–837.

Gelfand, I., & Fomin, S. (2000). *Calculus of variations.* Dover publications.

Gibbon, J. (1977). Scalar expectancy theory and weber's law in animal timing. *Psychological Review*, *84*(3), 279.

Gray, R. M., & Neuhoff, D. L. (1998). Quantization. *Information Theory, IEEE Transactions on*, *44*(6), 2325–2383.

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., Naiman, D., & 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.

Krueger, L. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*(02), 251–267.

Lin, Y., Michel, J., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus.

Lipton, J., & Spelke, E. (2003). Origins of number sense large-number discrimination in human infants. *Psychological Science*, *14*(5), 396–401.

Livingstone, M. S., Pettine, W. W., Srihasam, K., Moore, B., Morocz, I. A., & Lee, D. (2014). Symbol addition by monkeys provides evidence for normalized quantity coding. *Proceedings of the National Academy of Sciences*, *111*(18), 6822–6827.

Luce, D. (1959). On the possible psychophysical laws. *Psychological Review; Psychological Review*, *66*(2), 81.

Luce, D. (1962). Comments on Rozeboom's criticism of "On the Possible Psychophysical Laws.".

Luce, D., & Edwards, W. (1958). The derivation of subjective scales from just noticeable differences. *Psychological review*, *65*(4), 222–237.

MacKay, D. (1963). Psychophysics of perceived intensity: A theoretical basis for fechner's and stevens' laws. *Science*.

Malevergne, Y., Pisarenko, V., & Sornette, D. (2011). Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Physical Review E*, *83*(3), 036111.

Masin, S., Zudini, V., & Antonelli, M. (2009). Early alternative derivations of Fechner's law. *Journal of the History of the Behavioral Sciences*, *45*(1), 56–65.

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.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, *1*(2), 226–251.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, *46*(5), 323–351.

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

Nieder, A., Freedman, D., & Miller, E. (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. *The Journal of neuroscience*, *27*(22), 5986–5993.

Nieder, A., & Miller, E. (2004). Analog numerical representations in rhesus monkeys: Evidence for parallel processing. *Journal of Cognitive Neuroscience*, *16*(5), 889–901.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*, 1112–1130. Available from `http://colala.bcs.rochester.edu/papers/piantadosi2014zipfs.pdf`

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

Portugal, R., & Svaiter, B. (2011). Weber-Fechner Law and the Optimality of the Logarithmic Scale. *Minds and Machines*, *21*(1), 73–81.

Shannon, C. (1948). *The Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Smith, N., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 595–600).

Stevens, S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Stevens, S. (1960). The psychophysics of sensory function. *American Scientist*, 226–253.

Stevens, S. (1961). To honor Fechner and repeal his law. *Science; Science*.

Stevens, S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects.* Transaction Publishers.

Stoianov, I., & Zorzi, M. (2012). Emergence of a'visual number sense in hierarchical generative models. *Nature neuroscience*, *15*(2), 194–196.

Sun, J., & Goyal, V. (2011). Scalar quantization for relative error. In *Data compression conference (dcc), 2011* (pp. 293–302).

Sun, J., Wang, G., Goyal, V., & Varshney, L. (2012). A framework for Bayesian optimality of psychophysical laws. *Journal of Mathematical Psychology*.

Swartzlander, E. E., & Alexopoulos, A. G. (1975). The sign/logarithm number system. *IEEE Transactions on Computers*, *24*(12), 1238–1242.

Thurstone, L. (1931). The indifference function. *The Journal of Social Psychology*, *2*(2), 139–167.

Verguts, T., Fias, W., & Stevens, M. (2005). A model of exact small-number representation. *Psychonomic Bulletin & Review*, *12*(1), 66–80.

Wagenaar, W. (1975). Stevens vs Fechner: A plea for dismissal of the case. *Acta Psychologica*, *39*(3), 225–235.

Wainwright, M. J. (1999). Visual adaptation as optimal information transmission. *Vision research*, *39*(23), 3960–3974.

Wasserman, G., Felsten, G., & Easland, G. (1979). The psychophysical function: Harmonizing Fechner and Stevens. *Science*.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*(2), 130–137.

Xu, F., & Arriaga, R. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology*, *25*(1), 103–108.

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

Xu, F., Spelke, E., & Goddard, S. (2004). Number sense in human infants. *Developmental science*, *8*(1), 88–101.

Zipf, G. (1936). *The Psychobiology of Language.* London: Routledge.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort.* New York: Addison-Wesley.