A Theory of Reaction Time Distributions

Fermín MOSCOSO DEL PRADO MARTÍN
Laboratoire de Psychologie Cognitive (UMR-6146)
CNRS & Université de Provence (Aix–Marseille I)
Marseilles, France

Draft of December 28, 2008

Abstract

We develop a general theory of reaction time (RT) distributions in psychological experiments, deriving from the distribution of the quotient of two normal random variables, that of the task difficulty (top-down information), and that of the external evidence that becomes available to solve it (bottom-up information). The theory is capable of accounting for results from a variety of models of reaction time distributions and it makes novel predictions. It provides a unified account of known changes in the shape of the distributions depending on properties of the task and of the participants, and it predicts additional changes that should be observed. We show that a number of known properties of RT distributions are homogeneously accounted for by variations in the value of two easily interpretable parameters, the coefficients of variation of the two normal variables. The predictions of the theory are compared with those of many families of distributions that have been proposed to account for RTs, indicating our theory provides a significantly better account of the data across tasks and modalities. In addition, a new methodology for analyzing RT data is derived from the theory, and we illustrate how it can be used to disentagle top-down and bottom-up effects in an experiment. Finally, we show how the theory links to neurobiological models of response latencies.

Since their introduction by Donders (1869), reaction times (RTs) have been an important measure in the investigation of cognitive processes. As such, a lot of research has been devoted to the understanding of their properties. An issue that has raised some attention is the peculiar probability distributions that describe RTs, which have proved difficult to ac-

---

This work was partially supported by the European Commission through a Marie Curie European Reintegration Grant (MC–EIF–010318).
The author wishes to thank Anna Montagnini for having seeded the thoughts contained in this paper, and Xavier Alario, Xavier ..., Borís Burle, Yousri Marzouki, and Jonathan Grainger for discussion and suggestions on these ideas. Correspondence can be addressed to:
fermin.moscoso-del-prado@univ-provence.fr

count for by most general probability distribution families. This has in many cases led to the proposal of sophisticated *ad-hoc* distributions, specific to the domain of RTs (see Luce, 1986 for a comprehensive review of the field). A particular consequence of this is that the proposed distributions have gone further than being specific to RTs, but have become specific even to particular experimental tasks and modalities. In this study we attempt at putting these apparently different distributions under one general theoretical framework, show that they can all be grouped together in a single general purpose probability distribution. In addition, we discuss how this theory fits into both the high-level probabilistic models, and lower-level neurobiological models of processing. The theory that we propose makes new predictions, and has methodological implications for the analysis of RT experiments

Our theory can be stated in a relatively trivial form: RTs are directly proportional to the difficulty of the task, and inversely proportional to the rate at which information becomes available to solve it. To obtain a probability distribution from here one only needs to add that both the task difficulty and the incoming information are normally distributed and are possibly inter-correlated. As we will show, this simple statement has rich and novel implications for the shapes that distributions of RTs should take. The theory that we propose fully derives from the statement above without further additions.

The methodology that derives from the theory can be summarized in the following points. First, outlier removal based only on the magnitude of the RTs (whether absolute or relative, short or long) should be avoided. Second, in most cases, the adequate transformation for analyzing RT data is the reciprocal. Third, random effects, both of participant and stimulus (and other related to the particular experiment) should be explicitly included in the analyses as much as possible. Fourth, effects should be investigated both on the (reciprocal) mean and on its standard deviation.

We will discuss this problem in five stages. We will begin by providing an overview of one particular theory on the distribution of RTs in decisional tasks. This is the LATER model that was introduced by Carpenter (1981), and has since then received support from a range of studies. We will take Carpenter's LATER model as the starting point for our theory building. In the following section we will show how a simple extension of LATER leads to a surprisingly general model, capable of accounting for responses *across* participants, types of tasks, and modalities. Here we also discuss how our theory can account for some of the known properties of RT distributions. Having provided a basic description of our theory, we will continue by showing that our theory can also be taken as a generalization of some current neuro-biological models of decision making. We will pay special attention to the integration of our theory with the family of Drift Diffusion Models (DDM; Ratcliff, 1978), as these have proved very useful in explaining the RT distributions in many tasks, and offer a natural link to the properties of neural populations. The theoretical sections are followed by a methodological section introducing new techniques for the analysis of RT data that are derived from the theory. We continue by comparing our theoretical predictions with those of other commonly used RT distributions, paying special attention to the now very common Ex-Gaussian distribution (McGill, 1963). For this we make use of four lexical processing datasets across types of tasks (decision *vs.* recognition; comprehension *vs.* production) and modalities (visual *vs.* auditory). After this, we will focus on analyzing in detail the largests of these datasets, those of the lexical decision and word naming tasks provided by the English Lexicon Project (ELP; Balota, Yap, Cortese, Hutchinson, Kessler, Loftis,

Neely, Nelson, Simpson, & Treiman. 2007). We will show that our theory provides a better account of the distribution of responses, whether considered by individual subject or overall. Finally, we will illustrate how the implications of the theory can be used for disentangling top-down from bottom up-effects in a picture naming experiment. We conclude with a discussion of the theoretical and methodological implications of our theory.

The LATER Model

The LATER model ("Linear Approach to Threshold with Ergodic Rate"; Carpenter, 1981) is one of the simplest, and yet one of the most powerful models of reaction time distributions in decision tasks. Starting from the empirical observation that human response latencies in experimental tasks seem to follow a distribution whose reciprocal is normal, Carpenter proposed a remarkably simple model: He assumed that some decision signal is accumulated over time at a constant rate until a threshold is reached, at which point a response is triggered. Crucially, he added that the rate at which such decision signal accumulates is normally distributed across trials (see Figure 1, left panel). Despite its elegant simplicity, Carpenter and collaborators have – in a long sequence of studies – shown that such a model can account for a surprisingly wide variety of experimental manipulations, extending across different types of stimuli (auditory, visual, tactile) and response modalities going from button presses to ocular saccades (*e.g.*, Carpenter, 1981, 1988, 1999, 2000, 2001; Carpenter & Reddi, 2001; Carpenter & Williams, 1995; Reddi, Asrress, & Carpenter, 2003; Reddi & Carpenter, ,2000; Oswal, Ogden, & Carpenter, 2007; Sinha, Brown, & Carpenter, 2006).

In mathematical terms, the model is rather easily specified. If the response is triggered when the evidence – starting from a resting level (S_0) – reaches a threshold level (θ), and evidence accumulates at a constant rate (r) which, across trials, follows a distribution $N(\mu_r, \sigma_r^2)$, the response latency (T) is determined by:

$$T = \frac{\theta - S_0}{r} = \frac{\Delta}{r}. \tag{1}$$

If one further assumes that both S_0 and θ are relatively constant across trials, the the distribution of the times is the reciprocal of a normal distribution:

$$\frac{1}{T} \sim N\left(\frac{\mu_r}{\theta - S_0}, \left(\frac{\mu_r}{\theta - S_0}\right)^2\right). \tag{2}$$

This distribution is what Carpenter terms a *Recinormal distribution*.

The Recinormal Distribution

In order to compare the predictive value of LATER with other existing models of reaction time distributions, it will be necessary to have an explicit description of Carpenter's Recinormal distribution. In general, we can define the Recinormal distribution by its probability density function:

$$f_r(t; \mu, \sigma) = \begin{cases} \frac{1}{t^2\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(\mu t - 1)^2}{2\sigma^2 t^2}} & \text{if } t \neq 0, \\ 0 & \text{if } t = 0. \end{cases} \tag{3}$$

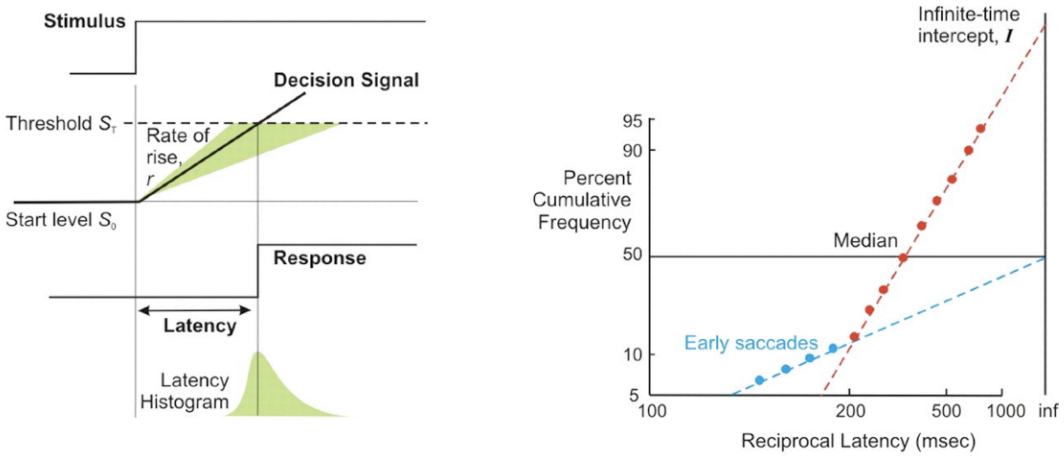

Figure 1. *Left panel*: Schema of the LATER model. Evidence accumulates from an initial state (S_0) to a decision criterion (θ). The rate (r) at which the evidence accumulates varies according to a normal distribution with mean μ_r and variance σ_r^2, giving rise to the typical skewed distribution of response latencies (left-bottom). *Right panel*: A "Reciprobit plot". When plotted against the theoretical quantiles of a normal distribution, the reciprocal response latencies (with changed sign) appear to form a straight line. This is indicative of them also following a normal distribution. In addition, a small population of early responses seems to arise from a different normal distribution. *Taken from Sinha et al., 2006 – permission pending.*

This distribution is continuous and defined along the whole real line. The parameters μ and σ correspond to the mean and standard deviation of the distribution of its reciprocal, which is a normal distribution (further details of the Recinormal distribution are provided in Appendix A).

Probabilistic interpretation

LATER can be directly interpreted at the computational level as an optimal model of hypothesis testing. The main parameters of the LATER model are the decision threshold (θ), the starting level for the evidence accumulation process ($S(0)$), and the mean and standard deviation of the evidence accumulation rate (μ_r and σ_r). If we take $S(0)$ to represent the logit prior probability of an hypothesis (H) being tested (*e.g.* a stimulus is present, the stimulus is a word, *etc*) on the basis of some evidence provided by a stimulus (E) arriving at a fixed rate r, then we have by Bayes theorem:

$$S(T) = \log \frac{P(H|E)}{1 - P(H|E)} = \log \frac{P(H)}{1 - P(H)} + \int_0^T \log \frac{P(E|H)}{1 - P(E|H)} \mathrm{d}t = S(0) + rT. \quad (4)$$

Therefore, interpreting the rate of information intake as the logit of the likelihood (*i.e.*, the log *Bayes factor*; Kass & Raftery, 1995) of the stimulus, and the prior information as the logit of the prior probabilities (the log *prior odds*), the accumulated evidence is an optimal estimate of the logit of the posterior probability of the hypothesis being tested (the log *posterior odds*) in an optimal inference process. Notice however, that this optimality is

dependent on two assumptions: first, on having a normal distribution of the incoming evidence, and second, on having a constant rate of evidence income. The first is in fact a rather safe assumption, Bayes factors are known to be asymptotically log-normally distributed (*cf.*, Kass & Raftery, 1995). The second one can in principle be more problematic though. If the evidence fluctuates during the decision process, even if this is a normal fluctuation with a mean equivalent to μ_r, the time that would be taken by an optimal evidence accumulation process corresponds to the first passage time of a linear Brownian motion with a drift, and that time is described by an Inverse Gaussian distribution rather than by the Recinormal that LATER advocates.

Reciprobit plots

LATER proposes using the "Reciprobit plot" as a diagnostic tool to assess the contribution of different factors to an experiment's results. This plot is the typical normal quantile-quantile plot (a scatter-plot of the theoretical quantiles of an $N(0,1)$ distribution, versus the quantiles from the observed data) with the axes swapped (the data are plotted on the horizontal axis and the theoretical normal on the vertical axis), and a (changed sign) reciprocal transformation on the data ($d = -1/RT$). In addition, the labeling of the axes is also changed to the corresponding RT values on the horizontal axis, and the equivalent cumulative probability on the vertical axis (see the right panel of Figure 1). Observing a straight line in this plot is in general a good diagnostic of a normal distribution of the reciprocal.

Variations in slope and intercept of the Reciprobit line are informative as to the nature of the experimental manipulations that have been performed. The Reciprobit plot is a representation of the distribution of the reciprocal of the RT:

$$\frac{1}{T} = \frac{r}{\theta - S(0)} = \frac{r}{\Delta}. \tag{5}$$

If the rate r is normally distributed with mean μ_r and variance σ_r^2, and Δ is a constant, then $1/T$ will also be normally distributed with mean and variance:

$$\mu = \frac{\mu_r}{\Delta},\ \sigma^2 = \frac{\sigma_r^2}{\Delta^2}, \tag{6}$$

and the slope and intercept of the Reciprobit are given by:

$$\text{slope} = \frac{1}{\sigma} = \frac{\Delta}{\sigma_r}. \tag{7}$$

$$\text{intercept} = \frac{\mu}{\sigma} = \frac{\mu_r}{\sigma_r}. \tag{8}$$

Therefore, variation in the Δ (prior probability or threshold level) will be reflected in variation in the slope only, while variation in the μ_r, the rate of information income, will affect only the slope of the Reciprobit plot.

These consequences have been experimentally demonstrated. On the one hand, variations in top-down factors such as the prior probability of stimuli, result in a change in the slope of the Reciprobit plot (Carpenter & Williams, 1995). In the same direction, Oswal *et al.* (2007) manipulated the variability of the foreperiod (*i.e.,* the SOA) by controling the

hazard rate of stimulus appearance (*i.e.*, the probability that a stimulus is presented at any moment in time given that it has not appeared before). The found that the instantaneous hazard rate correlated with the slope of the corresponding Reciprobit plots, giving further evidence that the expectation of observing a stimulus affects the starting level (S_0) of the decision process. Similarly, Reddi and Carpenter (2000) observed that if one manipulates the response threshold by introducing a variation in the time pressure with which participants perform the experiment, one also obtains a variation in the general slope of the line. On the other hand, Reddi *et al.* (2003) showed that changes in the information contained by the stimulus itself – the rate at which the evidence is acquired – are reflected in changes in the intercept of the Reciprobit plot. This was shown by proving that the proportion of coherently moving points in a random dot kinematogram are reflected in the intercept value on the Reciprobit plot.[1]

Neurophysiological evidence

In addition to providing a good fit to experimental data, some neurophysiological evidence has been presented that can support this type of model. Hanes and Schall (1996) found that, before saccadic onset, visuomotor neurons in the frontal eye fields show an approximately linear increase in activity. The rate of this increase varies randomly from trial to trial, and the time at which the saccade actually occurs has a more or less constant relation to the time when the activity reaches a fixed criterion. Furthermore, neurons in the superior colliculus also show rise-to-threshold behavior, with their starting level depending on the prior probability of the stimulus (Basso & Wurtz, 1997, 1998), and this decision based activity seems to be separate from that elicited by perceptual processes (Thompson, Hanes, Bichot, & Schall, 1996; see Nakahara, Nakamura, & Hikosaka, 2006, for an extensive review of the neurophysiological literature that provides support for LATER).

As it can be appreciated in the Reciprobit plot of Figure 1, there appears to be an additional population of very fast responses which do not follow the overall Recinormal distribution of the remaining latencies. These short responses are attributed to a different population of sub-cortical neurons that – very rarely – would overtake their cortical counterparts in providing a response (Carpenter, 2001; Carpenter & Williams, 1995; Reddi & Carpenter, 2000).

As we have seen, LATER provides a very good account of individual-level RT. The theory is simple and yet remarkably powerful. It seems like a good starting point on which to build a more general theory of reaction time distributions. In the following section we will describe a generalization of LATER that permits a wider coverage of experimental situations across different participants and items.

General Theory of RT Distributions

We have seen that RTs appear to follow a Recinormal distribution. However, this result holds only as long as the difference between the resting level and the threshold ($\Delta = \theta - S_0$) remains fairly constant. For several reasons, it is difficult to assume that

[1] Carpenter and colleagues in fact assume a constant vertical intercept at infinite time, and variation in the horizontal intercept only. In our opinion this is not so clear or informative, therefore we concentrate on variations on the intercept in general.

this quantity will remain constant in a psychological experiment. First, most interesting RT experiments will involve different types of stimuli, and in most cases these stimuli will be presented to multiple participants. Clearly, in many situations different stimuli will have different prior probabilities. As discussed above, variation in prior probability leads to variation in S_0 (Carpenter & Williams, 1995; Reddi & Carpenter, 2000). Furthermore, experimental participants themselves are also likely to show variations in both resting levels and threshold, depending on factors like their previous experience, age, etc. Finally, even in experiments of the type shown by Carpenter and colleagues, where the analyses are performed on individual subjects responding to relatively constant types of stimuli, it is not difficult to imagine that there is a certain degree of variation in the resting level due to – among other possibilities – random fluctuations in cortical activity, fatigue, and normal fluctuations in the participants' level of attention during an experimental session.

Therefore, in order to account for most of the experimental situations of interest in psychology, it will become necessary to explicitly include the possibility of fluctuations in both the information gain rate (r) and in the resting level to threshold distance (Δ). To keep consistency with LATER, we assume that Δ is also normally distributed with mean μ_Δ and standard deviation σ_Δ. If we keep the linear path assumption of LATER – we will show below that the distributional properties are not dependent on this particular path – the RT will be given by:

$$T = \frac{\Delta}{r}, \quad r \sim \mathrm{N}(\mu_r, \sigma_r^2),\ \Delta \sim \mathrm{N}(\mu_\Delta, \sigma_\Delta^2)\ . \tag{9}$$

Therefore, once we also allow for normal variation in the in Δ factor, the RT will follow a distribution corresponding to the ratio between two normally distributed variables. Notice that, under this assumption, both the RTs and the inverse RTs will in fact follow the same type of distribution: that of the ratio between normally distributed variables.

A further complication needs to be addressed. Up to the moment, and in line with other models that also propose to take this variation into account (Brown & Heathcote, 2008; Nakahara *et al.*, 2006), we have implicitly assumed that the values of r and Δ are statistically independent of each other. In reality, this seems over-optimistic. It is not rare that the perceptual properties of stimuli are in fact correlated with their prior probabilities. Consider for instance experiments that involve the presentation of written words. It is long known that the length in characters of a word is negatively correlated with its frequency of occurrence in a corpus of text (Zipf, 1949). However, some studies have shown that both of this variables seem to have different contributions to the recognition of a word, and these contributions depend on the amount of bottom-up or top-down processing that the task requires (*e.g.*, Baayen, Feldman, & Schreuder, 2006). It is reasonable to assume then that word frequency influences the prior probability of words (cf., Norris, 2006, Moscoso del Prado, 2008) while word length will have a stronger influence of the rate of intake of visual information (*e.g.*, Bernard, Moscoso del Prado, Montagnini, & Castet, 2008). The correlation between these factors will result in a correlation between both normal distributions in the ratio. Therefore, an additional parameter ρ representing the correlation between r and Δ needs to be taken into account.

Fieller's normal ratio distribution

The distribution of the ratio of possibly correlated normal variables is well-studied and known in analytical form. Fieller (1932) derived the expression for its density function, and Hinkley (1969) further studied it, crucially providing a normal approximation with explicit error bounds and conditions of application (See Appendix B for more details on this distribution.). I will henceforth refer to this distribution as *Fieller's distribution*.

Fieller's distribution is fully characterized by four free parameters.[2] If the random variables X_1 and X_2 follow a bi-variate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and a Pearson correlation coefficient of ρ, then the ratio between them follows a distribution:

$$\frac{X_1}{X_2} \sim Fieller(\kappa, \lambda_1, \lambda_2, \rho)$$
$$\kappa = \frac{\mu_1}{\mu_2}, \quad \lambda_1 = \frac{\sigma_1}{|\mu_1|}, \quad \lambda_2 = \frac{\sigma_2}{|\mu_2|}. \tag{10}$$

The shape parameters λ_1 and λ_2 represent the coefficients of variation (CoV) of each of the normal variables. As we will see below, their values have important consequences for the predictions of our model.

Especial cases of Fieller's distribution

An interesting property of Fieller's distribution is that, for particular values of its CoV parameters λ_1 and λ_2, it reduces to more familiar probability distributions. Table 1 shows the most notable of these cases. The most salient – and least interesting – reduction happens when both CoV parameters take a value of zero. This indicates that neither the numerator nor the denominator exhibit any variation, that is, the RT is a constant (*i.e.*, it follows a degenerate distribution with all probability mass concentrated in one point, a Dirac impulse function).

More importantly, when the CoV of the denominator (λ_2) is zero, Fieller's distribution reduces to a plain normal distribution with mean κ and variance $((\kappa\lambda_1)^2)$. This corresponds to the intuitive notion that if λ_2 is zero, the denominator is just a plain constant that divides the normal distribution on the numerator. In the reverse case, when λ_1 is the one that is zero (*i.e.*, the numerator is constant), Fieller's distribution reduces to Carpenter's Recinormal distribution, with reciprocal mean $1/\kappa$ and reciprocal variance $(\lambda_2/\kappa)^2$. Finally, when both the CoV parameters λ_1 and λ_2 approach infinity, the situation is tat of a ratio between two zero-mean distributions. In this case Fieller's distribution converges rather fastly to a Cauchy distribution (also known as Lorentz distribution). The convergence of the ratio distribution to Cauchy for high values of the CoV parameters is well-known in the theory of physical measurements (see *e.g.*, Brody, Williams, Wold, and Quake (2002) for a recent application of this property to biological measurements). These four particular cases of Fieller's distribution are summarized in Table 1.

[2] Hinkley (1969) used five parameters to describe it, but only four of those parameters are actually free. We have chosen to use instead a four parameter characterization from which the 5 parameters can be reconstructed. In addition, these parameters enable more direct inferences on the properties of the distribution

Table 1: Particular cases of Fieller's distribution. The numbers in brackets indicate estimated thresholds below or above which the reduction still applies.

Value of λ_1	Value of λ_2	Distribution	Normal QQ-plot
0	0	Dirac(κ)	
any	0 ($< .22$)	N(κ, $(\kappa\lambda_1)^2$)	straight line
0 ($< .22$)	any	ReciN$\left(\frac{1}{\kappa}, \left(\frac{\lambda_2}{\kappa}\right)^2\right)$	straight line (on reciprocal plot)
∞ ($> .443$)	∞ ($> .443$)	Cauchy$\left(\rho\kappa\frac{\lambda_1}{\lambda_2}, \frac{\lambda_1}{\lambda_2}\kappa\sqrt{1-\rho^2}\right)$	horizontal line and two vertical lines at edges

Diagnostic tools

These reductions are of interest for the analysis of RT experiments. Most obviously, the normal and Recinormal cases are particularly useful. They permit the analysis of RTs using common data analysis techniques such as correlations and linear models of different types without violation of their normality assumptions, only requiring minor transformation of the RTs. Furthermore, using the techniques that we will describe in the following section, in combination with the tools and inferences presented by Carpenter and colleagues, when the RT distributions approach the normal and Recinormal areas we are able to distinguish factors that selectively affect either the rate r or the start to threshold distance Δ. As we have discussed, this is of vital importance if one wants to separate bottom-up, perceptual effects, from top-down, experience-based ones. For this reason, it is important to have objective diagnostic tools to decide when it is safe to analyze the data according to one of the reduced distributions.

Figure 2 illustrates the effect on the shape of the RT distribution of varying values of λ_1 and λ_2. The curves represent the values of these parameters in the horizontal axes. The vertical axes plot a Montecarlo estimate of the negentropy of the resulting distributions.[3] The two upper panels plot the estimated negentropy of the actual variable (*i.e.*, deviation from normality), while the lower panels plot the estimated negentropy of its reciprocal (*i.e.*, deviation from recinormality). This is not dependent on the values of the remaining parameters of Fieller's distribution (we replicated exactly the same results for many different combinations of parameter values). The upper panels show that, independently of the value

[3]The negentropy of a distribution (Brillouin, 1956) is the Kullback-Leibler divergence between a given distribution and an normal distribution of equivalent mean and variance. This measure is usually employed to estimate the amount of deviation from normality shown by a distribution, for instance in Independent Component Analysis (Comon, 1994). The negentropy of a distribution is zero if and only if that distribution is normal. Any non-normal distribution will have a strictly positive value.

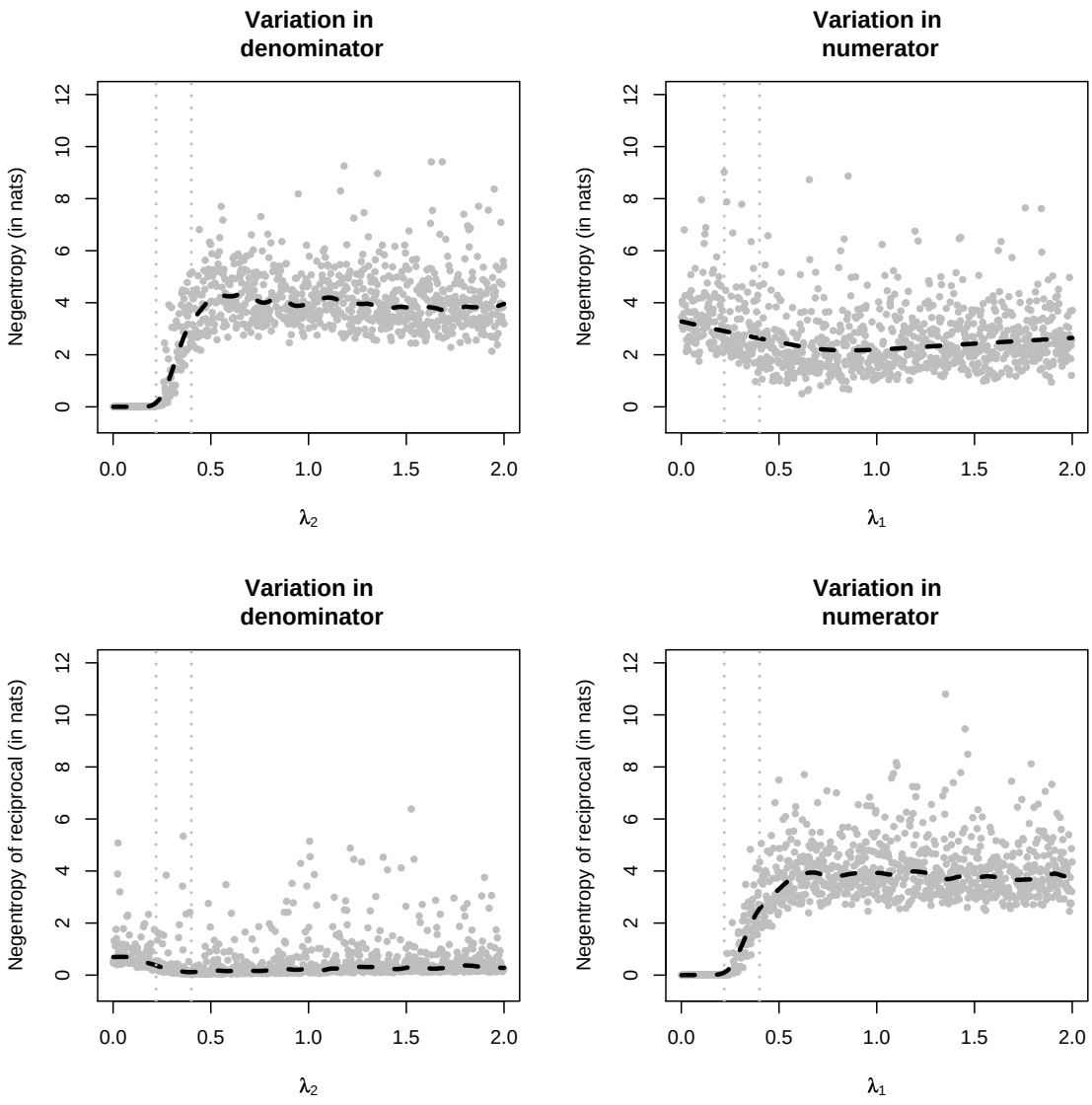

Figure 2. Deviations from normality and recinormality for different values of the coefficients λ_1 and λ_2. The upper panels plot a Montecarlo estimate of the negentropy for changing values of λ_2 fixing the value of λ_1 (top-left panel), and changing the values of λ_1 while keeping λ_2 constant (top-right panel). The bottom panels show the equivalent variations in the value of the estimated negentropy of the reciprocal, while keeping constant either λ_1 (bottom-left) or λ_2 (bottom-right). Each point represents a Montecarlo estimate of the negentropy based on a sample of 10,000 items from a Fieller's distribution. As fixed parameters, we used realistic values estimated from visual lexical decision data ($\kappa = 693.37, \lambda_1 = .26, \lambda_2 = .38, \rho = .59$). The vertical dotted lines indicate the approximate locations of phase changes, .22 and .4. The black dashed lines are non-parametric regression smoothers.

of λ_1, if $\lambda_2 < .22$, the distribution is in all respects normal.[4] In what follows we refer to this as the normal zone. On the other hand, as soon as λ_1 and λ_2 rise above around .4, it stabilizes itself at a value of approximately four nats, which is the approximate average negentropy of a Cauchy distribution.[5] We refer to the area in the graphs where both λ_1 and λ_2 are above .4 as the Cauchy zone. When the value of λ_2 lies between .22 and .4, there is a linear, rapidly growing deviation from normality towards the Cauchy distribution. We refer to this area of the plots as the linear zone. The same pattern is observed for the variation of λ_1 in the plots of the reciprocal, where again we find a Recinormal zone, a linear zone, and a Cauchy zone, defined by the same threshold values of .22 and .4. In sum, we can say that as long as λ_1 or λ_2 remains below .22, we will be able to safely analyze our data using the respectively the Recinormal or normal distribution (without the need for outlier removal).

The transition from the normal/Recinormal zones into the Cauchy zones can also be detected in normal quantile-quantile (NQQ) plots. While in the normal zone, the NQQ plot appears as a straight line. As we enter the linear zone, the plot gets split into a horizontal line covering most of the space at the center, and two vertical lines at the extremes. If one zooms in into this plot excluding the extremes, the vertical components at side progressively disappear, leaving a straight line as a result. The closer one gets to the Cauchy zone, the more zooming-in is necessary to obtain a straight line. Finally, once in the Cauchy zone, no matter how much zooming-in is is done, the plot will continue to have three components. The same diagnostic tool is valid for the Recinormal, but using the NQQ plot of the reciprocal. In sum, while remaining in the linear zone, the data can still be used for analysis using normal or Recinormal techniques, requiring only the removal of some outliers. Once in the Cauchy zone, normal analysis techniques cannot be used, and it is necessary to result to non-parametric methods or to methods that assume explicitly a Cauchy distribution.

In fact, as it will become clear in the data analysis sections, RT experiments normally exhibit distributions with both λ_1 and λ_2 in the linear zone, with λ_1 usually approaching the Recinormal zone around .25, and λ_2 closer to the .4 threshold. As shown by Hayya, Armstrong, and Gressis (1975), when the distribution is in the linear zone (which they define to lie under .39), the data can still be safely analyzed using normal techniques. This has the implication that, in most cases, a Recinormal-based analysis (*i.e.*, a reciprocal transform on the RTs) will provide a fairly good account of the data, possibly with some cleaning of outliers. Furthermore, when one reduces the variability in the experiment, for instance by analyzing single subject data, or by averaging across subjects or items, the CoV coefficients will be naturally reduced. If λ_1 is already close to the Recinormal zone, these decreases in the variability can fully place the distribution fully into the Recinormal zone, explaining

[4] Marsaglia (1965) gives a theoretical estimate of about $\lambda_2 < \frac{1}{3}$ (he expressed it in terms of the reciprocal CoV) for this threshold (in uncorrelated variables). In our experience with both simulated and real experimental data (which are usually correlated) this estimate proved to be too lax. As can be appreciated in Figure 2, between .22 and .33 there is already a significant departure from normality.

[5] Strictly speaking, the negentropy of a Cauchy distribution does not exist, as it would require a value for its standard deviation and all moments of the Cauchy distribution are undefined. However, bearing in mind that in RT we will always be concerned with truncated data (always at zero, and usually also at some upper bound RT), for which moments can be defined, by Montecarlo estimation we obtained that it has an empirical median value of 3.73, and an empirical mean value of $3.96 \pm .01$.

the observations of Carpenter and colleagues. When one needs to deal with individual responses of many subjects to many items, the distributions can become more complex. In some cases, the high values of the CoV parameter may put the distributions in, or very close to, the Cauchy zone. This can be problematic; the moments of Cauchy distribution are undefined, and so are those of its reciprocal (which is also Cauchy-distributed) and thus empirical means and variances calculated in one experiment, no matter how large, are likely not to be replicated by other experiments (cf., Brody et al., 2002). Fortunately, as we will see in the next section, some additional transformations, together with the natural truncation of RTs might alleviate this problem. Otherwise, when the RT distribution falls into the Cauchy zone, it will be safer to work on the median of the distribution (as does for instance Carpenter in most of his experiments), which is in all cases defined and replicable.

'Aggregativity'

One important property of the distribution that we are proposing is what we term its 'aggregativity', that is, if RTs are collected from a relatively homogeneous population of subjects and for a relatively homogeneous population of stimuli in the same task, the joint distribution of RTs across all stimuli and subjects should in the limit also follow an instance of Fieller's distribution.

To see this, consider that subjects are sampled from a normal population. Both their means μ_Δ and μ_r will then be normally distributed with means m_Δ and m_r, and variances s^2_Δ and s^2_r. Therefore, in the limit situation of infinite subjects the responses will follow a distribution of the ratio of two convolutions of normal variables, which are themselves normal. If the population is homogeneous, this limiting behavior will be reached rather quickly. The same argument holds for the distribution of items.

On the other hand, if either the population of items or that of subjects is not homogeneous, then we might expect the heterogeneity to propagate to the aggregate distribution of responses, leading to a mixture type distribution. Notice that this mixture consequence is also to be expected if the responses from each participant come themselves from a clear mixture population, as the very fast responses of sub-cortical origin would predict.

This has the implication that, when analyzing large datasets across many participants and items, the combined RTs will be Fieller distributed, although with λ_1 and λ_2 parameters with higher values than for the individual subjects or items. We will test this prediction in the data analysis section.

Hazard functions

When comparing the properties of different candidate probability distributions to describe RTs in auditory tasks, Burbeck and Luce (1982) suggested that crucial discriminating information is provided by the hazard functions, that is, the probability of a particular reaction time given that it was not shorter than that particular value:

$$h(t) = -\frac{\mathrm{d}\log\left(1 - F(t)\right)}{\mathrm{d}t} = \frac{f(t)}{1 - F(t)}, \tag{11}$$

where $f(t)$ and $F(t)$ are respectively the probability density function of the times and its cumulative probability function. Burbeck and Luce remarked that the shape of this function is notably different for different RT distributions. In particular, Luce and Burbeck contrast

distributions that show a monotone non-decreasing hazard function such as the normal, the Gumbel, and the Ex-Gaussian distributions, those that show a constant value as the exponential distribution, distributions that depending on their parameter values can show either increasing of decreasing hazard functions as is the case with the Weibull distribution, and those that show a peaked hazard function such as the Fréchet, the log-normal, the inverse Gaussian, and the RT distribution predicted by Grice's non-linear random criterion model (Grice, 1972).

Using evidence from the shape of hazard functions of RTs to auditory stimuli of varying intensities Burbeck and Luce (1982) argue that RTs to auditory stimuli arise from a competition between "level detectors" and "change detectors". In their view, the former give rise to RT distributions with monotonically increasing hazard functions, and RTs dominated by the later are characterized by peaked hazard functions. For instance, they show that RTs to high-intensity auditory signals show peaked hazard functions and thus are dominated by the change detectors, whereas RTs to low-intensity signals exhibit increasing hazard functions which would be the consequence of dominance of the level detectors. However, as Burbeck and Luce also noticed, it is in practice very difficult, if not outright impossible, to classify the RTs in one particular task as arising exclusively from either level detectors or change detectors. In any given task, even if the stimuli intensity is strongly manipulated, one will find that the reaction times arise from a combination of level and change detectors, although the contribution of each will vary.

Strictly speaking, the RT distribution that we are advocating belongs to those that have peaked hazard functions, although some considerations need to be made. As with the rest of the distribution's properties, the shape of the hazard function is determined by the CoV parameters λ_1 and λ_2 and the correlation coefficient ρ. Figure 3 illustrates the three basic shapes that the hazard function of a Fieller's distribution can take. The simplest, most common case are hazard functions of the type shown by the black line, that is a function growing up to early sharp peak, after which the function decreases monotonically eventually asymptoting to the constant decrease characteristic of power-law distributions such as the Pareto distribution. The location of this peak is controlled by the λ_2 parameter. As λ_2 approaches zero, the peak location goes to infinity, ultimately becoming a monotonically increasing function – a Gaussian – of which the dark grey line is an example. Finally, the light grey line illustrates a Recinormal case of Fieller's. In this case, the hazard function also shows an initial monotonically decreasing phase up to a very early minimum, after which the function evolves in a manner similar to the black line. In fact this minimum is also a general property of the distribution (arising from its bi-modality). Normally, for moderately high values of λ_1 this peak is very small and thus it is effectively not seen. However, when one approaches the Recinormal zone (the minimum fully disappears as λ_1 goes to infinity). The location of this minimum is controlled by the correlation coefficient ρ. Positive values of ρ (as the distributions showed) will place this minimum to the right of zero, making it more apparent, and possibly giving rise to a population of early responses in an apparent bi-modality.

As we have seen, Fieller's distribution can generalize both of the types of distribution discussed by Burbeck and Luce (1982). If the relative variability of the rate of evidence accumulation is very low, RT distributions will become more and more similar to a normal distribution, and thus the monotone increasing hazard. On the other hand, as the relative

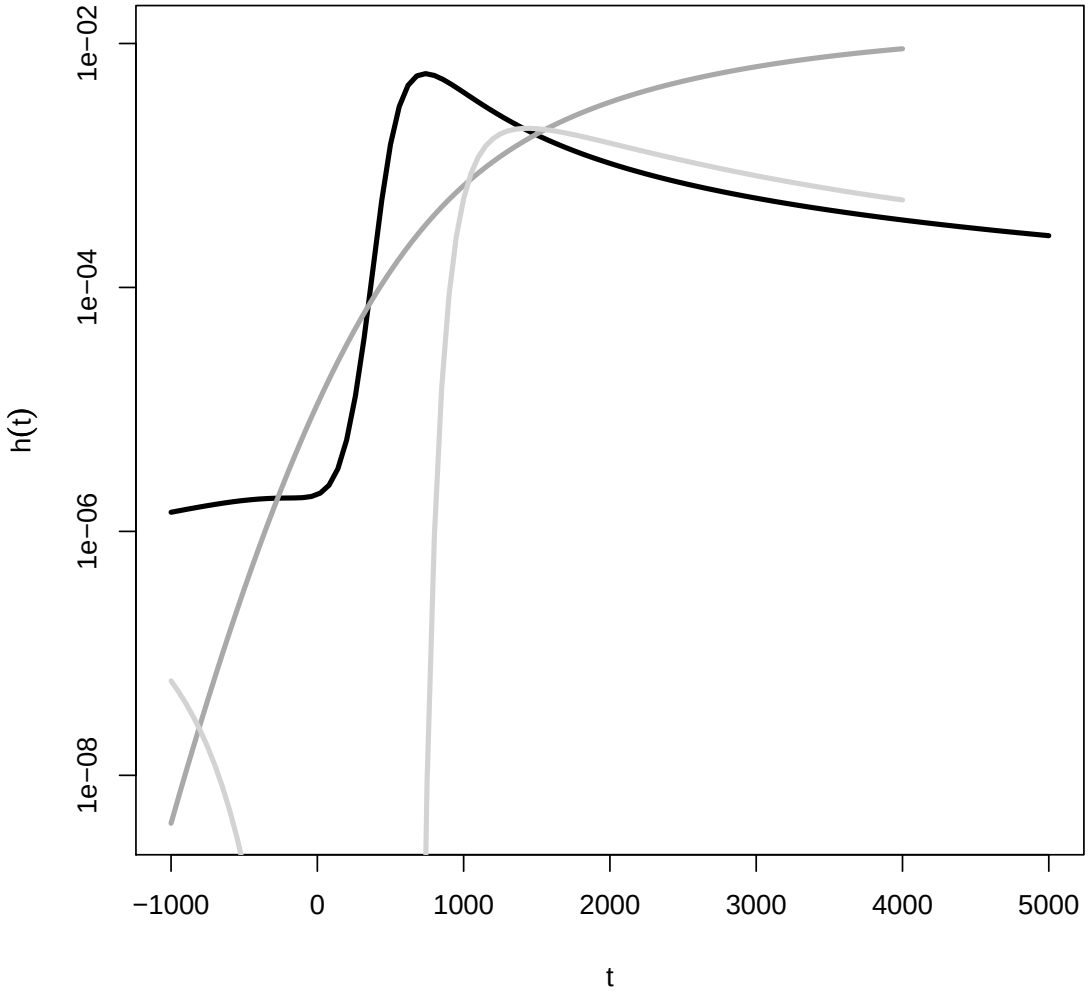

Figure 3. Typical shapes of the hazard function for three instances of Fieller's distribution. The black line plots a typical instance in a word naming experiment ($\lambda_1 = .49, \lambda_2 = .4$), not far from the Cauchy zone, the dark grey line plots an instance of a Fieller's distribution deep into the normal zone ($\lambda_2 = .05$), and the light grey line plots an instance of a distribution well into the Recinormal area ($\lambda_1 = .05$). Note that the abcises are plotted in a logarithmic scale to highlight small differences.

variability of the rate increases, the distribution becomes more similar to that predicted by a level detector type of model, described by distributions of the type of the inverse Gaussian or the log-normal. Importantly, Fieller's distribution also enables us to account for all types of balances between both types of processes, which are the case in most types of psychological experiments.

Right tails

Perhaps the most valuable information in order to discriminate between competing probability distributions is contained in the shape of their right tail, that is, the very slow responses. In fact, considering only the relatively fast reaction times located in the vicinity of the mode of a distribution can lead to serious problems of 'model mimicry', that is, completely different models can give rise to distributions that are in practice indistinguishable around their modes (*e.g.*, Ratcliff & Smith, 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). This problem is greatly attenuated when one examines the right tails of the distributions. In this area, different distributions give rise to qualitatively different shapes. It is therefore important to describe what our theory predicts in terms of the shape of the right tail of the distribution, and how does this contrast with other theories.

Clauset, Shalizi, and Newman (2007) provide a useful classification of possible shapes of the right tails of distributions. They distinguish between distributions whose right tails are equivalent to those of the *exponential* distribution, *power-law* type distributions (such as the Pareto distribution), distributions that show a power-law behavior up to a certain high value $t_{\max}$, from where they exhibit more of an exponential pattern (*power-law with cut-off*), distributions with *log-normal* tails, and distributions whose tail is thicker than an exponential but thinner than a power-law (*stretched exponential* type) such as the Weibull distribution.

Table 2 classifies several common RT distributions according to the taxonomy proposed by Clauset *et al.* (2007), we have added a class to accommodate the Gaussian (Clauset and colleagues consider only thick-tailed distributions). The classification has been performed by considering the dominant term in the probability density functions of each distribution. It is important to notice that the great majority of distributions that have been propose to describe RTs, have exponential type tails, including the Gamma distribution (*e.g.*, Christie, 1952; Luce, 1960; McGill, 1963), the Inverse Gaussian or Wald distribution (*e.g.,* Lamming, 1968; Stone, 1960), the Ex-Gaussian (*e.g.*, McGill, 1963; Hohle, 1965; Ratcliff & Murdock, 1976; Ratcliff 1978), the Ex-Wald (Schwarz, 2001), the 'large-time' series describing the first passage times in the diffusion model (*e.g.*, Luce, 1986; Ratcliff, 1978; Ratcliff & Smith, 2004; Ratcliff & Tuerlinckx, 2002; Tuerlinckx, 2004), and the closed form approximation to the DDM introduced by Lee, Fuss, and Navarro (2007). In general, any distribution that results from the convolution of an exponential with another one will belong to this group, except in cases where the other distribution in the convolution is of a power-law or stretched exponential type.

The stretched exponential type of distributions includes the Weibull, which has been argued as a model of RT distributions (Colonius, 1995; Logan, 1988; 1992; 1995). This type of distributions show thicker right tails than the exponential type distribution, but still thinner than one would observe in a power-law type that we are proposing. In the especial cases where Fieller's distribution is in the Recinormal or Cauchy zones, from a

Table 2: Classification of distributions according to the shape of their right tails. The DDM-large corresponds to the 'large-time' infinite series expansion of the first passage times of the (linear) Drift Diffusion Model given by Feller (1968). The DDM-small is the 'small time' expansion of Feller (1968). The DDM-Approximate corresponds to the closed-form approximation given by Lee *et al.*, (2007). Fieller's (general) refers to the general case of Fieller's distribution outside the normal, Recinormal, or Cauchy zones.

Distribution	Type	Dominant term	Shape (on log scale)	Shape (on log-log scale)
Exponential Gamma Inverse Gaussian Ex-Gaussian Ex-Wald DDM-large DDM-approximate	Exponential	$e^{-\lambda t}$, $\lambda > 0$	Linear decrease	Exponential decrease (slow)
Normal	Quadratic-exponential	e^{-kt^2}	Quadratic decrease	Exponential decrease (fast)
Log-normal	Log-normal	$\frac{1}{t}e^{-(\log t)^2}$	Quasi-linear decrease	Quadratic decrease
Pareto Cauchy Recinormal Fieller's	Power-law	$t^{-\alpha}$ $\alpha > 1$	Logarithmic decrease (from $t_{\min}$)	Linear decrease (from $t_{\min}$)
DDM-small	Power-law (with cut-off)	$t^{-\alpha}e^{-\lambda t}$ $\alpha > 1$, $\lambda > 0$	Power-law until $t_{\max}$ and linear from $t_{\max}$	Power-law until $t_{\max}$ and exp. from $t_{\max}$
Weibull	Stretched exponential	$t^{\beta-1}e^{-\lambda t^{\beta}}$ $\lambda, \beta > 0$	Above-linear decrease	Below-linear decrease

certain value $t_{\min}$, the tail will be a power law with exponent α of 2 (Jan, Moseley, & Stauffer, 1999).[6] More generally, Fieller's distribution will show a power-law tail behavior, with exponent value between 2 and 3. Finally, in cases of very small values of λ_2, when the distribution approaches the normal zone, the value of $t_{\min}$ increases, eventually going to infinity as λ_2 goes to zero. These extremely thick tails provide a sharp contrast with the right tails predicted by most other models. Of the other models proposing very thick, supra-exponential right tails, we find that the 'short time' variant of the first passage time described by Feller (1968) and applied to RT distributions by van Zandt (2000), and van Zandt, Colonius and Proctor (2000), can give rise to this cutoff power-law behavior. Notice however, that according to Feller, this approximation is only valid for the *short* RTs, and thus not for the right tails (see Navarro & Fuss, 2008 for details).

As we have seen, our theory predicts much thicker right tails than would be predicted by most current theories, except for the distribution predicted by the original LATER model of Carpenter (1981). LATER's Recinormal distribution – from which our theory evolved – is also of a power-law right tail type (with exponent two). In order to test this distinct prediction, we will need to examine large datasets. By definition, events in the right tail are

[6] Jan *et al.*, 1999 provide derivation of the exponent value and demonstrate its application to an Ising model of magnetization at critical temperatures.

very rare, but still we are predicting that they should happen much more often than one would expect in other theories. This also implies that we should avoid truncating RT data on their right tail, as this can often contain the only information that enables discrimination among theories. Unfortunately, RT are in most situations truncated to a maximum value during data collection, so in many cases our power to examine the right tail will be severely hampered. However, the common practice of discarding RTs longer than 3,000 ms. (*e.g.*, Ratcliff, Van Zandt, & McKoon 1999), 2,500 ms. (*e.g.,* Wagenmakers *et al.*, 2008) or even a short as 1,500 ms. (*e.g.*, Balota *et al.*, 2008). In this respect, it is important to contrast our proposal, with the outlier cleaning recommendations of Ratcliff (1993) who, based on simulations using the Ex-Gaussian and Inverse Gaussian distributions (both of the exponential tail type) recommended truncating the data at a fixed cut-off between 1,000 ms. and 2,500 ms. In the data analysis sections we will test these predictions.

'Express' responses

Carpenter's motivation for positing the presence of a separate population of very fast responses in the LATER model comes from the apparent deviations from recinormality that are observed in some experimental situations (Anderson & Carperter, 2008, Carpenter, 2001, Carpenter & Williams, 1995, Reddi & Carpenter, 2000). Figure 4 reproduces some results of Reddi and Carpenter (2001) in this respect. Notice that, especially in the time pressure condition, a separate population of fast responses seems to arise, represented by the lower slope regression lines.

Carpenter and colleagues attribute these 'express responses' to units in the superior colliculus responding to the stimuli before the cortical areas that would normally be in charge of the decision have responded. These units are assumed to have a higher variability in rise rate than the cortical neurons have, and are expected to have a mean rise rate close to zero. This would result in them anticipating the cortical response only in a very few cases when they happen to fire much faster than they usually would (but within their normal range of variation). The fast sub-population arises more frequently in some conditions than others. First, as it is evident from Figure 4, the differentiated fast responses arise more clearly in subjects or conditions that elicit faster responses. In Reddi and Carpenter's study, these were more apparent in the condition including time pressure than in the condition that did not include it. In addition, from the graph it appears that the less accurate participants showed a greater presence of these responses. Second, Carpenter (2001) showed that variability in the order of stimuli can also affect the proportion of very fast responses. Ratcliff (2001) showed that Carpenter and Reddi's data were also well modeled by the DDM, and also accepted the need for a separate population of slow responses. In his model, Ratcliff replicated the change in the left tail of the distribution by adding a small population of fast responses sampled from a uniform distribution.

Although the neuro-physiological mechanism that is argued to justify the very short latencies is very plausible, there are several indications that make it difficult to believe that this mechanism is responsible for the greater part of these short latencies. First, as can be appreciated in Figure 4, the transition between both populations is not really a sharp one, as one could expect from a real mixture distribution. Rather, there is an, admittedly fast, but still rather smooth transition between both linear components (this is most apparent in participants AC and AM of the figure). Second, following Carpenter's argument, one would

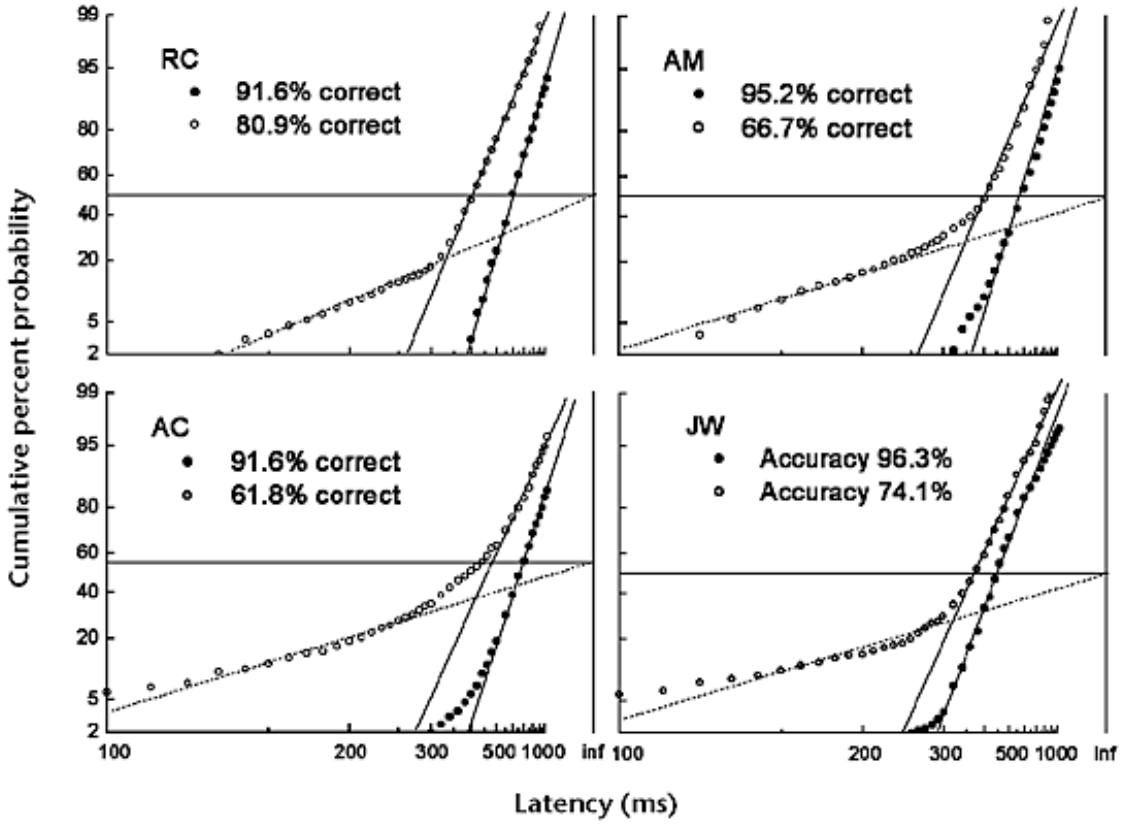

Figure 4. Evidence for the presence of a separate population of express responses. Notice that each of these Reciprobit plots can clearly be fitted by two straight lines, one for a minority of very fast responses, and one for the bulk of experimental responses. The open circles represent a condition in which participants responded under time pressure, while the filled dots plot the results of responding without such pressure. *Figure taken from Reddi and Carpenter (2000) – permission pending.*

expect that such sub-population only accounts for a very small percentage of responses. However, as can also be seen in their graph, in Reddi and Carpenter's results the fast sub-population accounts for over 40% of the responses in the time-pressure situation of participants AC and AM (in fact participant AC seems to show a *majority* of short responses in the time pressure condition), and similar very high percentages of fast responses are found in other studies (*e.g.*, Anderson & Carpenter, 2008).

What the high proportions and smooth transitions between both Reciprobit lines seem to suggest, is that those fast responses actually belong to the same distribution that generates the slower ones. In this direction, Nakahara, Nakamura, and Hikosaka (2006) suggested that this deviation would partially arise in an extension of the LATER model – ELATER – that allows for uncorrelated variations in the starting level to threshold distance (Δ). Indeed, the results of Nakahara and collaborators suggest that a single population may suffice to account for both the bulk of the responses, and the left tail ones.

As mentioned before, the studies of Carpenter and collaborators focused on analyzing single participant data separately, normally also separating the analyses for each condition. In most cases, this keeps the CoV parameter λ_2 relatively low (below the .22 threshold or very close to it), and assures that the results stay to a large degree within the Recinormal zone of Fieller's distribution. However, when the value of of this parameter is slightly increased, even in these very controlled single subject analyses, λ_2 can easily exit the Recinormal zone, entering the linear zone (.22 – .443). As we discussed before, once in this zone, Fieller's distribution very quickly departs from normality or recinormality, approaching the full Cauchy situation. It is thus interesting to see what shape this distribution takes when in the linear zone, possibly still close to a Recinormal zone.

Figure 5 illustrates the typical effect of taking a Fieller-distributed variable from the recinormal zone into the beginning of the linear zone. The points were randomly sampled from a Fieller's distribution with parameter $\lambda_1 = .3$ (the other parameters were kept to realistic values taken from the analysis of an English lexical decision experiment). The population of short responses arises very clearly, and the resulting reciprobit plot seems to be well-fitted by two straight lines, just as was observed in the experimental data. However, just as in Reddi and Carpenter's results, there is a smooth transition between both lines. We can see that in fact, contrary to the arguments of Carpenter and colleagues, a small modification of the LATER model predicts that the majority of fast responses belong to *the same* population as the slower ones. Just a very slight increase in the variability of Δ is sufficient for them to arise.

This result does not argue for the non-existence of a separate population of very short responses, that arises through a different neural mechanism. Rather it argues for a gradual change in the shape of the RT distribution, together with rather abrupt changes in shape when the parameters cross certain thresholds, leading to the bilinear aspect of the reciprobit plots. As we will see later in the analyses of actual experimental data, a population of very fast responses that fits the precise description provided by Carpenter and colleagues, is indeed present on the experimental data. Crucially however, this corresponds to a true minority of the responses (around one per thousand in the datasets we present). Thus, the neurophysiological mechanisms that have been positted to account for fast responses, also have behavioral consequences. Furthermore, we have provided the mechanism by which the precise shape of this bi-linearity can be predicted.

Non-decision times

Most models of reaction times in psychological tasks include a component of time that is unrelated to the actual task. This 'non-decision' time comprises the delays that arise from physiological factors such as neural conduction delays, synaptic delays, etc. Different models have accounted for this time in different ways. Some models have considered this to be a constant (*e.g.*, Luce & Green, 1972; Scharf, 1978; Tiefenau, Neubauer, von Specht, & Heil, 2006; Woodworth & Schlosberg, 1954) some others have assumed this to follow an exponential distribution (*e.g.*, McGill, 1963), and others have considered that this time will vary according to a normal distribution (*e.g.*, Hohle, 1965).

In the original LATER model, this pre-decision component of the time is assumed to have little variability, and not to have much influence on the final distribution of latencies. This assumption seems to match the empirical evidence relatively well – the reciprocal

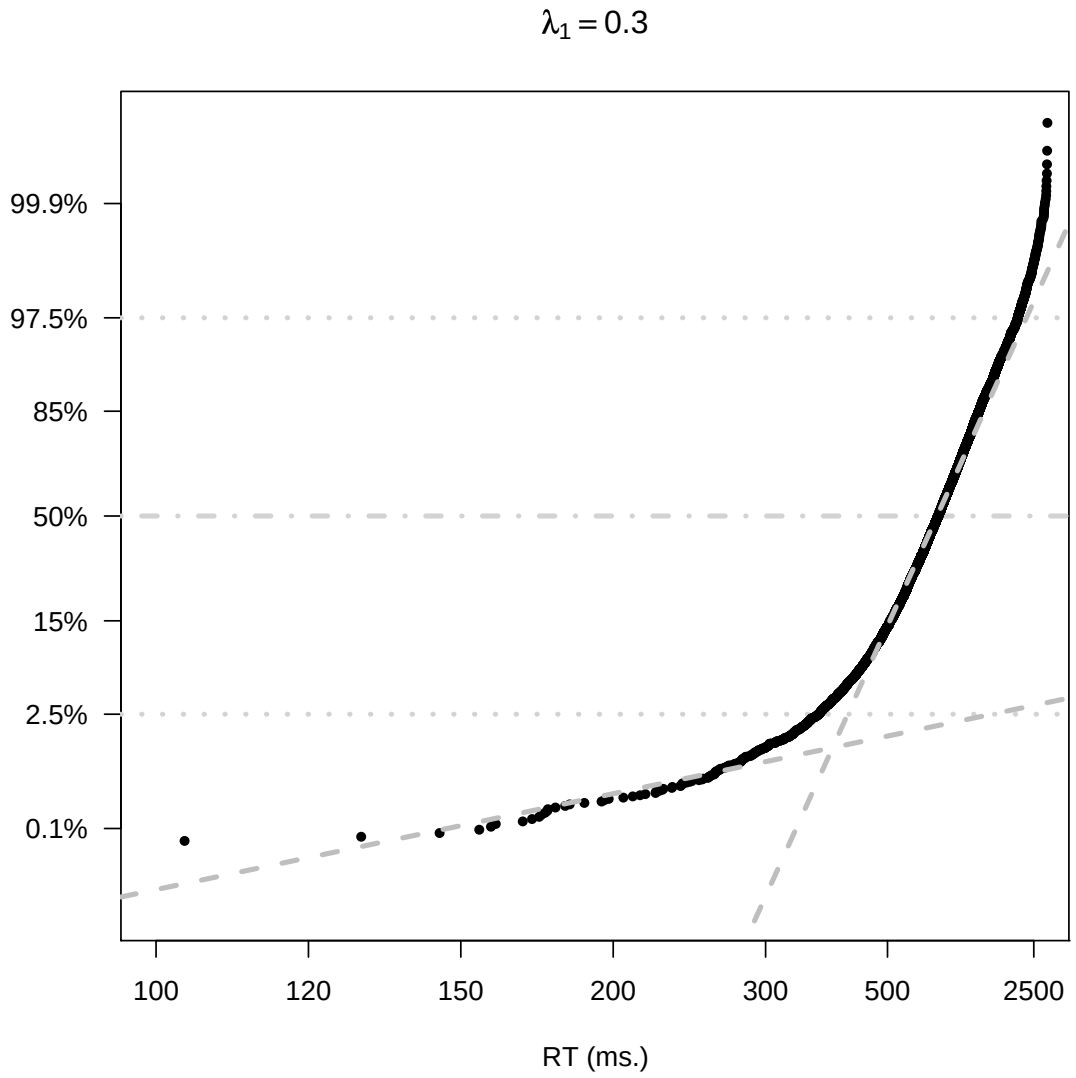

Figure 5. Typical Reciprobit plot of Fieller's distribution bordering the Recinormal zone. The data were sampled from a Fieller's distribution with parameter $\lambda_1 = .3$, that is, outside the Recinormal zone, but not yet reaching the Cauchy zone. The horizontal lines mark the median and 95% interval. The parameters used to generate the dataset were taken from the analysis of lexical decision latencies, with the only modification of λ_1. The remaining parameter values were $\kappa = 695$, $\lambda_2 = .38$, and $\rho = .6$. After sampling, the data were were truncated, keeping only the values in the interval from 1 ms. to 4000 ms., as typically happens in experimental situations.

RTs in Carpenter and colleagues' experiments do in most cases seem to approach a normal distribution. However, when one considers the magnitudes of non-decision times in typical psychological tasks, it is difficult to assume that they will not affect the distribution of latencies. Consider for instance the case of a lexical decision experiment. In such experiments, the non-decision component of the latencies are usually estimated to lie in the range of 350 ms. to 450 ms., depending on the particular modelling assumptions and experimental conditions (*e.g.*, Norris, in press, Ratcliff, Gomez & McKoon, 2004; Ratcliff & Smith, 2004; Wagenmakers *et al.*, 2008; but see for instance Wagenmakers, Steyvers, Raaijmakers, van Rijn, & Zeelenberg, 2004 for an estimate as short as 200 ms.). We can therefore say that the total time T is the sum of a non-decision component (T_n), which can either be constant or be itself a random variable, and a decision component (T_d) that arises from the evidence accumulation process. The decision component of the time is derived from the ratio betwen Δ and r. Taking these processes together, the expression for the response time becomes:

$$T = T_n + T_d = T_n + \frac{\Delta}{r} = \frac{\Delta + T_n \cdot r}{r}. \tag{12}$$

The literature suggests T_n does not have a very high CoV. If T_d follows a Fieller distribution, then the distribution of T will also be well-described by Fieller's distribution. Therefore, we could perform our analysis without dealing with the T_n factor (see Appendix B for details). The estimated parameters of the corresponding distribution are:

$$T \sim Fieller(\widehat{\kappa}, \widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{\rho}), \tag{13}$$

where $\widehat{\kappa}$ would be our estimate of the ratio between the means of $\Delta + rT_n$ and r, $\widehat{\lambda}_1$ estimates the CoV of $\Delta + rT_n$, $\widehat{\lambda}_2$ estimates the CoV of r, and $\widehat{\rho}$ is our approximation of the Pearson correlation coefficient between r and $\Delta + rT_n$. This estimation would have some problems. This would introduce a distortion by T_n in all parameters of the distribution. Of particular importance is the distortion that would be introduced into $\widehat{\rho}$. In the cases where the CoV of r (λ_2) is much higher than the CoV of Δ (λ_1), our estimate $\widehat{\rho}$ will become very high, as most of the variance of the numerator would also be driven by r. Fortunately though, in these cases the distribution will be well into the Recinormal zone, with or without the added term, and thus the analyses can still be safely performed. However, in cases where the distribution does not reduce neither to the Recinormal nor to the normal special case, this can be problematic, and it might be necessary to remove the contamination before proceeding to the analyses. In addition, $\widehat{\lambda}_1$ will also be significantly contaminated by the non-decision time. As this now estimates the CoV of a sum of random variables, its value will necessarily be higher that the values of λ_1 that one would have obtained without the contamination of T_n.

It is important to notice that, strictly speaking, it is *impossible* to remove the non-decisional component of the task without *a-priori* knowledge of the correlation between Δ and r. This is because for any normal ratio distribution, independently of its correlation coefficient ρ, there exists a constant such that one can find another normal ratio distribution with $\rho = 0$ (or in fact with whatever correlation coefficient we desire) that results from subtracting the constant from the target distribution (cf., Marsaglia, 1965; see Appendix B for more details).

Anticipations & non-responses

An interesting issue becomes apparent our formulation of the theory. Assuming that the variation in Δ is normally distributed implies that it has a non-zero probability of being negative. In turn, this would imply that in a sufficiently large sample of points we will find negative values of Δ. Such negative values can be interpreted as the resting level being above the response threshold *before* any stimulus was presented. If we state that a response is triggered as soon as the accumulated evidence surpasses the threshold, in these cases we would actually predict a small percentage of responses to be triggered before stimulus presentation.

In many experimental situations, participants respond before the stimulus itself was presented, or before it has been physically possible to process it and trigger a response (for instance, the response latency is significantly shorter than the expected conduction and synaptic delays in the motor nerves, indicated that the response must have been initiated before the stimulus was presented). In many cases these are just not measured, for instance in situations when the responses are only recorded at a certain period *after* stimulus presentation. Therefore some of these responses will fall in the "no-response" category, even though a response was indeed provided. In other cases, anticipations will appear as "double-response" cases, being attributed to the previous stimulus.

Figure 6 illustrates some of the situations in which different types of anticipations might happen. The figure plots a hypothetical bivariate normal distribution of Δ and r. The area is divided into three colored zones. First, the dark grey shaded area corresponds to the area of normally observable responses, that is, responses that are observable and happen after stimulus presentation. Note that this area extends into the second quadrant. This is to account for responses whose RT becomes negative once T_n is subtracted. The diagonal dashed line in the second quadrant plots $y = -T_n$, that is, points situated to the left of that line will correspond to cases when the response was provided *before* the stimulus (if we assumed a constant value of T_n). This is the area shaded in light grey in the figure. Points falling within this area are what we refer to as anticipations. In this cases, Δ has taken a negative value. This has the interpretation that, due to the variation is resting level or in threshold (or in both), the accumulator was above the threshold before the stimulus presentation, thus triggering a response. In addition, some of the points to the right of the $-T_n$ line, but still within the second quadrant (and even some of the points in the first quadrant close to the $\Delta = 0$ axis, if we also allow for random variations of T_n) will also correspond to anticipations. Although the responses were provided after stimulus presentation, this was still too early for having been triggered after the stimulus.

Finally, the white area corresponds to negative values of r with positive values of Δ. In this case, the accumulator will never reach the threshold in positive time. This will result in either no response being provided, or an error arising from the response of some other accumulator.

The probabilities of the three types of points can be easily computed. Anticipations will happen whenever *either* of the accumulators has a negative value of Δ. These will correspond the sum of the whole left hemi-plane of Figure 6 for both accumulators. These points describe all observed responses shorter than T_n, and these are likely to be at chance level with respect to correctness (chance being their prior probabilities) since these responses

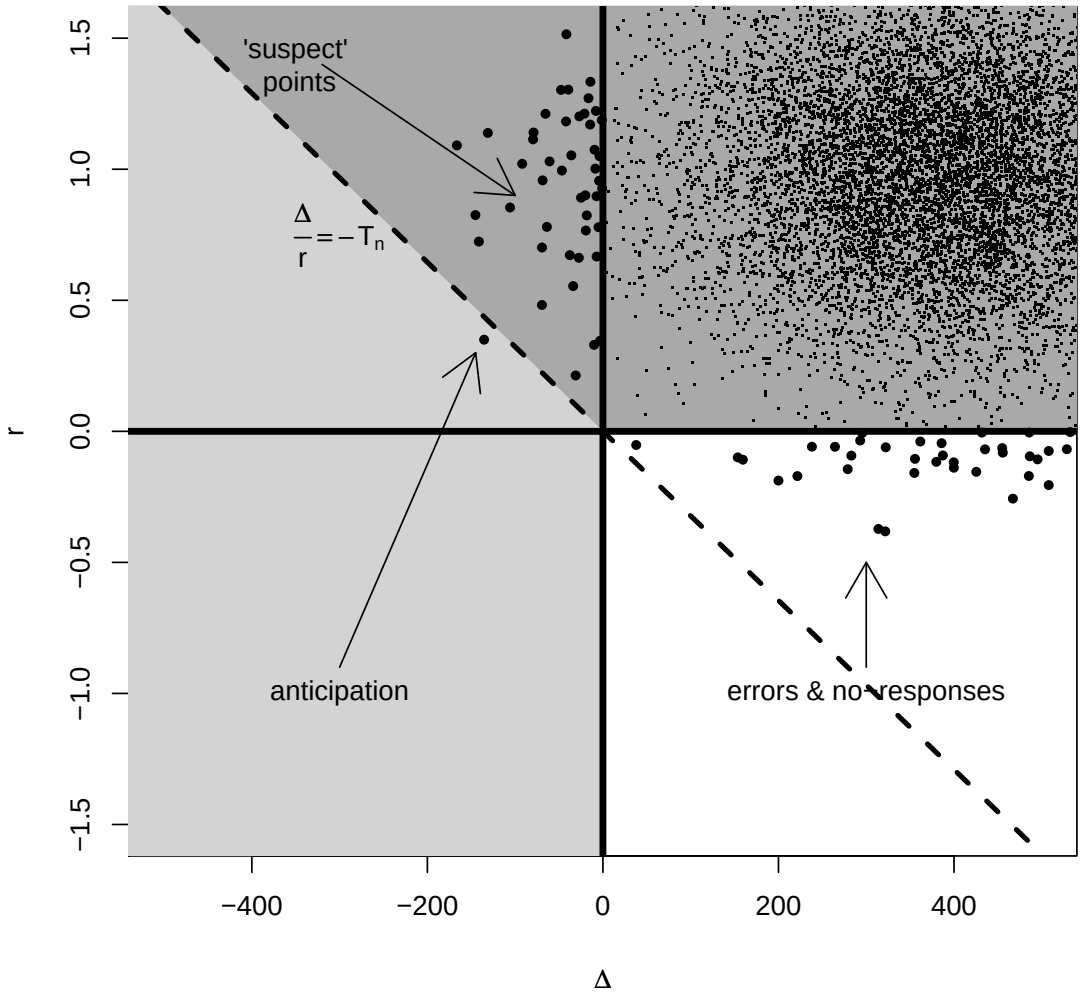

Figure 6. Anticipations and not-observable responses. The plot is divided into three areas: the observable zone (dark grey), the points that are observable before stimulus onset (light grey), and the non-observable zone (white). The diagonal dashed line marks $-T_n$. When observable, points to the left of this line will correspond to "negative" RTs. Points in the light grey area corresponds to cases where the accumulator was above threshold even before the stimulus was presented. Points falling into the white zone of the plot correspond to negative values of r, and thus they will never be observed, as they will either be no response cases (their trajectory never intersects with the threshold in positive time) or an error will be produced (some other accumulator eventually responds).

were triggered independently of the stimulus. In addition, there will be a proportion of cases in which both accumulators have negative rates, giving rise to no-response cases. Only the remaining responses will be the "valid" correct and incorrect ones, that happen as a result of a normal evidence accumulation process triggered by the stimulus. Using $\Phi(x)$ to denote the cumulative density function of the standard normal distribution, and $L(x, y; \rho)$ for the cumulative density of the standard bivariate normal distribution with correlation coefficient ρ, the probabilities of these three cases are:

$$P(\text{Anticipation}) = \Phi\left(-\lambda_{1,A}^{-1}\right) + \Phi\left(-\lambda_{1,B}^{-1}\right) - \Phi\left(-\lambda_{1,A}^{-1}\right)\Phi\left(-\lambda_{1,B}^{-1}\right) \quad (14)$$

$$P(\text{No}-\text{response}) = \left[L\left(\infty, -\lambda_{2,A}^{-1}; \rho_A\right) - L\left(-\lambda_{1,A}^{-1}, -\infty; \rho_A\right)\right] \cdot \cdot \left[L\left(\infty, -\lambda_{2,B}^{-1}; \rho_B\right) - L\left(-\lambda_{1,B}^{-1}, -\infty; \rho_B\right)\right] \quad (15)$$

$$P(\text{Valid}) = 1 - P(\text{Anticipation}) - P(\text{No}-\text{response}), \quad (16)$$

The theory that we propose predicts a certain number of anticipations. Anticipations can belong to two general groups. On the one hand, some anticipations might arise from the activity of non-cortical areas, responding in advance to a particular stimulus before having received cortical feedback. These correspond to the express responses mentioned by Carpenter & Williams (1995), which as we described above are not modeled by our theory. On the other hand, our theory still predicts an additional number of anticipatory responses arising at the cortical level, and falling within the general distribution of responses that we are proposing. In short, even after discounting express, sub-cortical responses, our theory predicts that a number of additional anticipations should occur. Crucially, the theory is sufficiently detailed to predict how many of these will be observed. Obviously, the correctness of this responses will be at chance level for each particular experiment. That is, although the anticipations will not be influenced by individual stimuli, they will be influenced by the overall properties of the experiment and of the participant. By manipulating these, one could in fact induce different probabilities of anticipations.

Errors and Alternative Responses

An issue that has become crucial when comparing theories of RTs in choice tasks is the success with which the are able to predict the proportion of errors in an experiment, and their RT distributions relative to the correct responses. This particular aspect has led to some serious criticism of many models. In particular, LATER has not fared particularly well in this part of the debate (*e.g.*, Ratcliff, 2001). Although Hanes and Carpenter (1999) provide some evidence that a race between multiple, laterally inhibited, accumulators could hypothetically explain error responses, they provided no detailed quantitative approach of it.

Recently, Brown and Heathcote (2005, 2008) proposed a family of 'ballistic' accumulator models that seem well-suited to account for error responses both in their proportion and in their RT distribution. Brown and Heathcote (2008)'s Linear Ballistic Accumulator (LBA) model is in fact very much the same as LATER, with only an additional component of uniformly distributed variation in the resting level of the system (S_0). As Hanes and Carpenter had proposed, LBA relies on a race between competing accumulators, and errors are produced when this race is won by the "wrong" accumulator. It is worth noticing here

that Brown and Heathcote also add that the different accumulators are independent of each other. Strictly speaking, a race of accumulators cannot account for the data without relying on the presence of some type of inhibition, be it lateral, feed-forward, or central. An easy way to see it is that, once a response has been given by one accumulator, it is necessary to 'call off the race' to avoid that several other competitors produce additional, and possibly conflicting, responses. Thus, Hanes and Carpenter's proposal of inhibition, restricted to some temporal limit (*i.e.*, a response cannot be inhibited from a certain threshold) is a necessary component of any theory (and has received substantial empirical support).

In the theory that we propose, errors arise from the competition between multiple accumulators, with some inhibition mechanism binding them together, whether this mechanism is central, lateral, or feed-forward is not relevant at our level of explanation, as they all can reduce to equivalent models (Bogacz, Brown, Moehlis, and Holmes, 2006). Different accumulators simultaneously integrate evidence, and the first one to reach a threshold triggers a response, inhibiting all the others.

Two alternatives.

In a simple two-alternative choice task, two accumulators A and B are integrating evidence. An error will be produced whenever the incorrect accumulator (B) reaches the threshold before the correct one (A) does. As discussed above the activities of the accumulators are proportional the logit probabilities of the particular stimuli. In these lines, the rates and prior probabilities can be interpreted as logit probabilities, that is, if $P(A)$ is the prior probability of accumulator A being the correct one, then its resting level will have an average increase of (or an average decrease in its threshold):

$$L_A = \log\left(\frac{P_A}{1-P_A}\right) \tag{17}$$

Notice that, in this interpretation, when the probability does not favor one option, it necessarily favors the opposite. Therefore $P_B = 1 - P_A$ and $L_B = -L_A$. The same interpretation is valid for the rates, which in this case would represent the average likelihood accumulated at each point in time in favor of other hypothesis.

Besides from hindering a probabilistic interpretation of the theory, not explicitly considering the presence of competition – as in the LBA model – has undesirable empirical consequences on the shape of the distribution. As noted by Brown and Heathcote (2008), the main effect of a race between independent accumulators is a general speed-up of the responses, which is more marked the larger the number of accumulators entering the race. Crucially, this speeding up is not uniform, but rather affects the right tail of the distributions more strongly: The more time it passes, the more likely it becomes that one of the accumulators reaches the threshold. However, when one considers the presence of competition, and of possible decay processes in the accumulators, the thinning of the right tail should become less pronounced. Although time does indeed increase the chances of the race having a winner, the more time it passes the more inhibition and decay (self-inhibition) the accumulators receive, making their responding less likely. However, this difference will only become apparent on very large datasets, with sufficient points for the very rare events in the right tail to become clearly visible.

Let us start by considering the pure race case, in the sense of Brown and Heathcote (2008). Denoting by F_A and F_B the Fieller cumulative probability distribution func-

tions with the parameters of accumulators A and B, and by f_A, f_B the corresponding probability density functions, in the case of an independent race then the probability density of an error or a correct response occurring at time $t > T_n$ would be given by:[7]

$$p(\text{Error}, t) = f_B(t)\,[1 - F_A(t)] \tag{18}$$

$$p(\text{Correct}, t) = f_A(t)\,[1 - F_B(t)] \tag{19}$$

And the probability of an error or a correct response occurring is the sum along the positive times (excluding times shorter than the non-decision time):

$$P(\text{Error}) = \int_{T_n}^{\infty} f_B(t)\,[1 - F_A(t)]\,\mathrm{d}t \tag{20}$$

$$P(\text{Correct}) = \int_{T_n}^{\infty} f_A(t)\,[1 - F_B(t)]\,\mathrm{d}t \tag{21}$$

These integrals can be evaluated numerically without much problem. Note however, that the (relative) closure under convolution and cross-correlation of Fieller's distribution (see the section on the stability of Fieller's distribution on Appendix **??** for details on this closure) simplifies this problem even further. The probability of an error is just the value of the cumulative distribution of the cross-correlation between B and A, evaluated at zero (with a correction for the anticipations and no-response cases). As we describe in the Appendix, this can be estimated directly by combining the properties of Fieller's distribution with those of the normal product distribution. Therefore, representing the cross-correlation between distribution by the $\star$ operator:

$$\begin{aligned} P(\text{Error}) &= P(\text{Valid})\,[F_B(t) \star F_A(t)]\,(0) \\ P(\text{Correct}) &= P(\text{Valid})\,[F_A(t) \star F_B(t)]\,(0) \end{aligned} \tag{22}$$

Notice that the probability of an error and the probability of a correct response do not add to one, as we saw above, the rest of the probability mass corresponds to possible anticipations and non-response cases.

Finally, the conditional density functions for correct and incorrect responses would be their joint distributions normalized by the probability of a valid error or the probability of a valid correct response:

$$\begin{aligned} p(t|\text{Error}) &= \frac{f_B(t)\,[1 - F_A(t)]}{P(\text{Error})} \\ p(t|\text{Correct}) &= \frac{f_A(t)\,[1 - F_B(t)]}{P(\text{Correct})} \end{aligned} \tag{23}$$

The application of (19), (21), (22), and (23) would describe the behavior of a 'pure race' model between LATER-style units without any lateral or central inhibition process,

[7] To be fully correct, an additional multiplicative term needs to be added to discount "positive" RTs resulting from points in the third quadrant of Figure 6. However, since points when *both* distributions will fall in this quadrant are extremely unfrequent, for notational simplicity we can overlook them.

in a manner very similar to the LBA model proposed by Brown and Heathcote (2008). In the limit, these race distributions should converge into instances of a Weibull distribution (*cf.*, Colonius, 1995; Logan, 1992; 1995). However, as it will become clear in the data analysis section, the presence of lateral inhibition and/or possible decay processes, results in fundamentally different RT distributions. In reality, the right tails of the distributions will become much thicker than one would predict using (23). This is caused by the lack of independence of the accumulators, which will severely delay responses.

Multiple alternatives.

The development of the theory for the two-choice case is valid without significant alterations for the general multiple choice, and recognition cases. In these cases, there is either one correct response among a finite set of possible candidates, or there is a *preferred* candidate among a finite set of possible responses, all of which could be considered correct as is the case in the picture naming example that we will discuss below.

Some recent models (*e.g.*, Brown & Heathcote, 2008) break down the competition between multiple accumulators to the individual accumulator level. Notice however, that this becomes of difficult application when the set of possible responses in not well-defined *a priori* (*e.g.*, the picture naming example). Rather, on an algorithmic level of description (in Marr (1982)'s sense), it is preferable to describe the process in terms of a correct or preferred response, and all the others. This corresponds – at the computational level – to the notion of probability. We have the probability of choosing a particular preferred response, and we have the probability of not choosing it (*i.e.*, choosing a different one).

With this in mind, we can define a generalized accumulator to account for all of the non-preferred responses. The generalized accumulator will have a negative average rate (for the probability of the preferred one to grow).

Some authors have argued that, once we allow for the competition between accumulators, the distribution of their joint first passage time through a threshold is after all a minimum distribution, and through the Theory of Extreme Values one should conclude that it will converge to on of the three generalized extreme value distributions (Gumbel, Weibull, or Fréchet; see Colonius, 1995 and Logan, 1995 for a detailed mathematical discussion of this point). However, this is strictly true if the race were to happen between *independent* accumulators (*e.g.*, assumption *c)* in Colonius, 1995), as in the case of the LBA model. In fact, if one considers the need for some inhibition mechanism binding the accumulators, their average hitting times do not constitute independent samples, but rather are strongly negatively inter-correlated ($r \simeq -1$) and the pre-conditions of Extreme Value Theory do not hold.

Neurophysiological Plausibility and Relationship to the DDM

The Drift-Diffusion Model (DDM; Ratcliff 1978) is perhaps the most successful family of rise to threshold models. As noted by Ratcliff (2001) in his response to the Reddi and Carpenter (2000) study, LATER and the DDM share many common characteristics, to the point that they might be considered convergent evidence models. In his letter, Ratcliff additionally points out that the DDM presents a number of advantages over LATER. The first of these if that the DDM also provides a direct mechanism to account and predict error probabilities and their latencies. The use of to opposed thresholds is crucial for this.

To this point, Carpenter and Reddi (2001) reply that the results of Hanes and Carpenter (1999) shown that a race between two accumulators would be able to explain error responses. In this direction, Brown and Heathcote (2005; 2008) showed that a 'ballistic' model (in the sense of constant increase in evidence, as is the case in LATER) including competing accumulators can exhibit similar properties to the DDM, both in terms of errors and RTs. However, as we will see in the empirical section below, when examined in detail on its tail predictions, Brown and Heathcote's models cannot possibly account for the RT distribution accurately. Crucially, in a recent study, Bogacz, Brown, Moehlis, Holmes, and Cohen (2006) have shown that linear rise to threshold models relying on competition between accumulators can all mathematically reduce to the DDM as long as there exists an inhibition mechanism across the accumulators in the model (independently of whether this inhibition is lateral, feed-forward, or central – 'pooled' in Bogacz *et al.*'s terminology). Hanes and Carpenter provided evidence that there might in fact be lateral inhibition between the accumulators (up to a certain temporal limit). The presence of lateral inhibition is also supported from human electro-physiological data (Burle, Vidal, Tandonnet, and Hasbroucq, 2004). Note however, that the results of Bogacz *et al.* are limited to linear accumulation models, and it is not clear that LATER is an instance of such. Although the rise to threshold is linear (leading to identical predictions of the median RTs), we will show below that the variance in the process is not so.

A second factor that seems to favor the DDM account over LATER is its suggestive approximation of the behavior of neural populations. Indeed, neural populations are very noisy and it is difficult to assume that they will show a constant rate increase in activities or firing rates. More likely, they will show a seemingly random fluctuation that, when sampled over a long time or across many measurements, will reveal the presence of a certain tendency or drift that pushes the level of oscillations up or down. These highly random fluctuations on a general accumulation can be observed both in animal single-cell recordings (Hanes & Schall, 1996) and in human electro-physiological data (cf., Burle *et al.*, 2004, Philiastides, Ratcliff, & Sajda, 2006). DDM-style or random walk models are naturally suited to deal with this random variations in the neural signal, and studies have demonstrated that the DDM can account well for the behavior of single neuron data (Ratcliff, Cherian, & Segraves, 2003; Ratcliff, Hasegawa, Hasegawa, Smith, & Segraves, 2007) although the introduction of non-linearities might be necessary (Roxin & Ledberg, 2008).

On the other hand, LATER presents a clear advantage, its elegant simplicity. The model is significantly simpler to fit than any model based on stochastic differential equations – which in many cases do not have known solutions. Its parameters have a clear interpretation on a computational level similar to those of the DDM (Carpenter & Williams, 1995). The distribution of RTs generated by our version of LATER is known in closed form, and it is relatively easy to fit. The tools for making predictions and analyzing data based on our theory are simple enough not to require advanced mathematical knowledge on the part of the experimentalist, and yet sufficiently powerful in the inferences that they enable. In a way, this theory resembles a general *law* of RT distributions. It seems that our extended LATER provides an excellent theory of RTs at Marr's *algorithmic level*. On the other hand, the DDM requires, even in its simplest linear versions, a very sophisticated mathematical framework in order to manipulate it, to the extent that a large number of studies have been published to describe techniques for fitting it (*e.g.*, Lee, Navarro, & Fuss, 2007; Navarro &

Fuss, 2008; Ratcliff & Tuerlinckx, 2002 Smith, 2000; Tuerlinckx, 2004; Voss & Voss, 2007; 2008; Wagenmakers, van der Maas, & Grasman, 2007; Wagenmakers, van der Maas, Dolan, & Grasman, 2008). Even in its simpler cases, the distribution of RTs that it predicts is not known in a closed form, but at best as an infinite sum of series or as a infinite recursive procedure (cf., Smith, 2000; but see also the approximations provided by Lee *et al.*, 2007 and Navarro & Fuss, 2008). We intend to show that – with some qualifications – the DDM family can be regarded as an *implementational level* description of the same process that our theory can explain at a higher level.

LATER's trajectory does not need to be linear

A first issue that could cast doubts on the plausibility of LATER as a model of activity accumulation in neurons (or more likely neural populations) is the constrained linear trajectory of the accumulation of evidence (this constrained linearity is in fact also problematic for a probabilistic interpretation of LATER). Even if we overlooked the noisy fluctuations that are observed in actual neural accumulations, the shape of the average accumulation itself does not seem to be linear, but rather seems to follow some type of exponential law (see for instance Burle *et al.*, 2004).

Fortunately, despite its explicit linear assumption, the predictions of LATER do not depend on the linear trajectories. This was already noticed by Kubitschek (1971).[8] In fact, any function f that is defined in the positive domain, and for which an inverse function f^{-1} exists, could serve as a model of the trajectory of LATER giving rise to an identical distribution of RTs, as long as the accumulated evidence is a function of the product of r and t. To see this, consider that the evidence at time t accumulates as a function f of the product of the rate and time rt:

$$S(t) - S_0 = f(rt). \tag{24}$$

Then, we can apply the inverse function on the left hand side of the equation, to obtain:

$$t = \frac{f^{-1}(S(t) - S_0)}{r} \tag{25}$$

Therefore, having any linear, non-linear, or transcendental invertible function (of the rt product) will produce identical results to those predicted by LATER as long as the "rate" parameter is normally distributed (which has a less clear interpretation in this generalized case).

To illustrate this point, consider that neural activity actually accumulates as an exponential function (as would for instance posterior probabilities). Then the equivalent expression for LATER would be:

$$S(t) = S_0 \mathrm{e}^{rt}. \tag{26}$$

Then we could use the logarithmic transformation to obtain:

$$t = \frac{\log\left(S(t)\right) - \log(S_0)}{r}. \tag{27}$$

[8] Kubitschek proposed a model identical to LATER to model cell-division times (see Kubitschek, 1966) which has since become the standard model of this type of processes. Interestingly, a random variation in the threshold level was also posited to refine the model (see Cooper, 1982).

In this case, it would be useful to define the starting level in a more appropriate way. If we define $s(t) = \log(S(t))$, then we can work with a new formulation of the resting level $s_0 = \log(S_0)$:

$$s(t) = s_0 + rt. \tag{28}$$

In this formulation, as long as r, $s(t)$ and s_0 are normally distributed (*i.e.*, $S(t)$ and S_0 are log-normally distributed), t will follow Fieller's distribution.

LATER reduces to a variant of the DDM

We propose that LATER provides a description at the algorithmic level, of what the DDM family implements at a more implementational level in the sense of Marr (1982). For this to be the case, we need to show how LATER can be implemented using a DDM process.

The accumulation of evidence by a linear DDM (*i.e.*, a Brownian motion with a drift and an infinitesimal variance) at any time point t is described by a normal distribution with mean $S_0 + \upsilon t$ and variance $s^2 t$, where S_0, υ, and s respectively denote the resting level (*i.e.*, the prior or starting value of the process), the mean drift and infinitesimal variance of the process. Similarly, the average accumulation of evidence by a LATER-style model with mean rate r is also described by a normal distribution centered at a mean $S_0 + rt$ (we will start our analysis using the constant Δ case and then extend it to the general case). Thus, equating the average drift υ with the average rise rate r will result on the same average accumulation of evidence. However, the variance at time t of the accumulated evidence in a LATER process with a variation in rate σ_r is $\sigma_r^2 t^2$. It is clear from this that there is no possible constant value of sigma that will reduce LATER to a classical DDM. Notice also that a compression of time will not produce the desired result, as it would also affect the mean accumulation. The most evident solution to achieve the same results is to define it as a diffusion model described by the Itô stochastic differential equation (SDE):

$$\begin{cases} \mathrm{d}S(t) &= r\,\mathrm{d}t + \sigma_r\sqrt{2t}\,\mathrm{d}W(t), \\ S(0) &= S_0 \end{cases} \tag{29}$$

where $S(t)$ denotes the accumulated evidence at time t and $W(t)$ is a standard Wiener process. At time t the accumulated evidence $S(t)$ is follows the desired normal distribution with mean $S_0 + rt$ and variance $\sigma_r^2 t^2$. We will refer to this reformulation of LATER in terms of the diffusion process as LATER-d. In turn the Itô SDE describing the classical DDM is:

$$\begin{cases} \mathrm{d}S(t) &= \upsilon\,\mathrm{d}t + s\,\mathrm{d}W(t), \\ S(0) &= S_0. \end{cases} \tag{30}$$

Comparing both equations, the only difference lies in the diffusion coefficient of both processes. While the DDM has a constant expression for it (s), that of LATER-d is a function of time ($\sigma_r\sqrt{2t}$). This expresses that the magnitude of the instantaneous fluctuation (*i.e.*, the 'average step size') at any point in time, in the LATER-d case is a function of time itself, whereas in the original DDM it remains constant. Therefore, although at the beginning of the process the variance of the accumulated evidence is likely to be smaller in LATER-d than in the classic DDM, with time LATER-d's variance overtakes that of the DDM.

This last point is schematized in Figure 7. The left panel compares 500 trajectories randomly sampled from a DDM with 500 "ballistic" trajectories sampled from a LATER

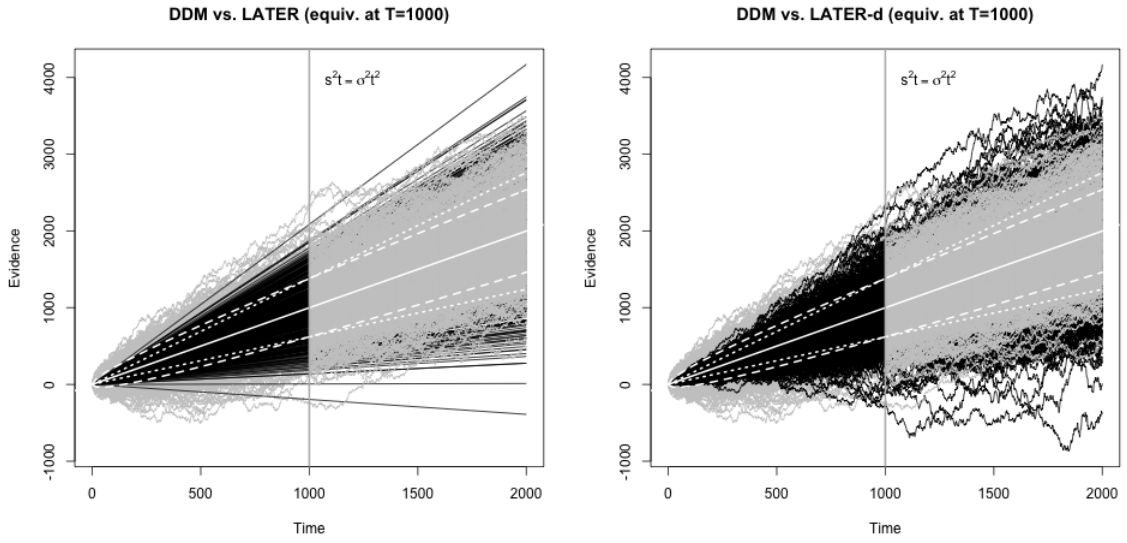

Figure 7. Comparison of LATER and DDM. The left panel overlaps 500 trajectories of the DDM (grey paths; $\upsilon = 1, s = 12.02$), with 500 trajectories of a LATER model (black paths; $r = 1, \sigma_r = .38$). The right panel plots the same DDM trajectories (grey paths), with trajectories sampled from LATER diffusion (black paths) equivalent to the process in the left panel. The solid white line marks the mean evidence. The dashed lines mark the 1 SD intervals of the DDM, and the dotted white lines show the 1 SD interval of the LATER models. The SDE simulations were performed using the R package "sde" (Iacus, 2008).

model. The parameters in the models were chosen on a realistic LATER scale, and were fixed to result in equal variances for both processes at time 1000. For visibility purposes, we have overlaid the LATER trajectories on top of the DDM's in the early times, and the DDM's on LATER's at times greater than 1000. It is apparent that, while LATER shows a triangular pattern of spread, the DDM results in a parabolic pattern, were the speed of growth of the spread decreases with time. The right panel shows how LATER-d has an identical behavior to the original fixed-trajectory version.

It remains only to extend LATER-d to consider the possibility of variability in Δ giving rise to Fieller's distribution of RTs. This is now trivial, the only thing that one needs to add is either variation in the resting level (S_0) or in the response threshold level (θ), or possibly in both. Figure 8 illustrates the effects that adding these additional noise components into the model. On the one hand, we can add a (normal) variation into the threshold level that is constant in time. We have represented this case as making the threshold fluctuate according to a distribution $N(\theta, \sigma_\theta^2)$, whose standard deviation is plotted by the grey dashed line in the picture. On the other hand, variation can be put directly in the starting point (*i.e.*, resting level) of the system. Then, at time zero, the accumulated evidence will follow a distribution $N(S_0, \sigma_s^2)$. As time progresses, at any point in time this variation combines with the variation of the drift (black dotted lines in the figure). Taking into account that the drift and the resting level can be correlated (the parameter ρ of Fieller's

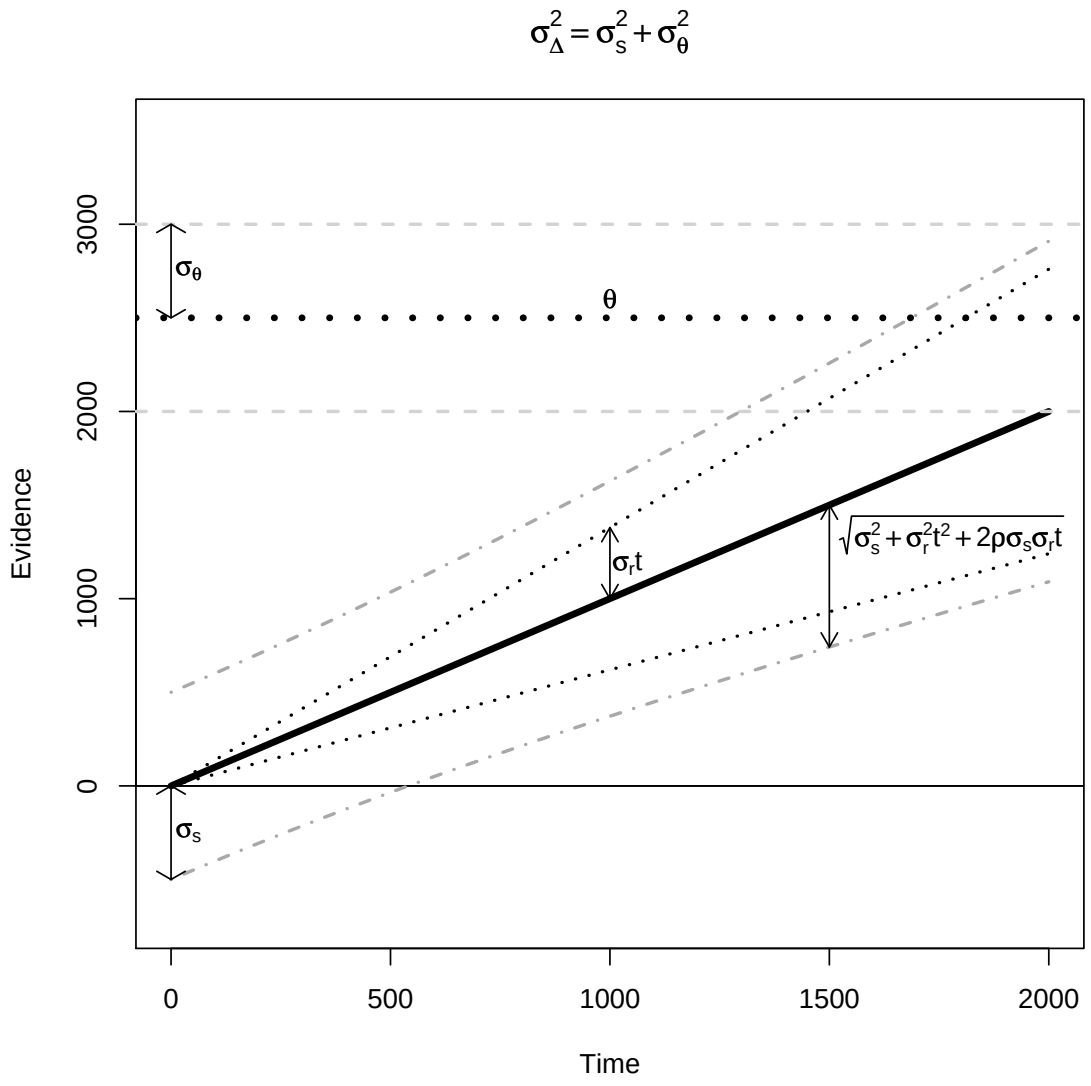

Figure 8. Adding variation in the threshold (σ_θ) and/or starting level (σ_s) to LATER-d.

distribution) the accumulated evidence follows a distribution $N(S_0+rt, \sigma_s^2+\sigma_r^2t^2+2\rho\sigma_s\sigma_r t)$, which is depicted by the grey dash-dotted lines in the figure. The only constraint is that both of these variances must sum up to the overall variance of the resting level to threshold distance ($\sigma_\Delta^2 = \sigma_s^2 + \sigma_\theta^2$). As described in the previous section, the first crossing times of this system will follow Fieller's distribution.

An notable issue that becomes apparent in Figure 8 is that the longer the reaction time, the lesser the influence of the variation in Δ. For graphical convenience, consider the case where we place all of σ_Δ in σ_s, leaving $\sigma_\theta = 0$. It is clear from the Figure that the additional variance added by σ_s on the increasing variance caused by σ_r becomes very small. This can be observed in the asymptotic convergence between the black dotted lines and the grey dash-dotted line. This has the implication, that, for tasks with very long reaction times, or for long responses in a particular, there will effectively be little deviation from the Recinormal case presented by Carpenter. This also explains why the "express responses" arise more often in the left side of the Reciprobit plot than on the right side, and why the faster conditions clearly show more of it than the slower conditions (see the experimental results of Reddi and Carpenter (2000) reproduced in Figure 4).

Techniques for Analyzing RT Data

In the previous sections, we have developed a theory of RT distributions. The theory has implications on how the analysis of RT data should be conducted. In this section, we elaborate on the actual methods that derive from the theory, explaining how they can be practically applied.[9]

Assessing recinormality

A first, relatively subjective, assessment of the recinormality of the by-item datasets is provided by examining the Reciprobit plots, as proposed by Carpenter (1981) and subsequent studies a Recinormal variable should give rise to a relatively straight Reciprobit plot. In addition, given that in the positive domain the Recinormal distribution is a (shifted) special case of the Box-Cox power-normal family of distributions (Box & Cox, 1964; Freeman & Modarres, 2006; see Appendix A for details) one can obtain a more objective assessment of recinormality by investigating the optimal value of the power parameter of the Box-Cox transformation in order to transform the data into a normal distribution. Recinormal distributions should result in optimal values of the power parameter close to -1, whereas log-normal type distributions should reveal values closer to zero. Values of the power parameter close to 1 are indicative of normality of the untransformed data.

Separating top-down from bottom-up effects

We now return to Carpenter's interpretation of the Reciprobit plots. As we discussed above, the distinct variations in the intercept and slope of such plots that are caused by different experimental manipulations are meaningful with respect to the nature of the manipulation. On the one hand, top-down factors such as manipulations in prior probability

[9] All methods described in this section have been implemented using the freely available R software for statistical computing, and the code for all routines is available under the GPL licence from the Open Knowledge Foundation Knowledge Forge website (`http://www.knowledgeforge.org`).

or urgency to respond alter the slope of the Reciprobit plot (Carpenter & Williams, 1995; Oswal *et al.*, 2007; Reddi & Carpenter, 2000). On the other hand, manipulating the general difficulty of perceiving or distinguishing a stimulus is reflected in a change in the plot's intercept (Reddi *et al.*, 2003). Therefore, in order to distinguish between these two types of effects we need an analysis technique that is able to reliably separate what is affecting the slope and what is affecting the intercept.

If we have a recinormally distributed variable, with reciprocal mean μ and reciprocal standard deviation σ, we can see that the slope (a) and intercept (b) of the Reciprobit plot are given by:

$$a = \frac{1}{\sigma}, \quad b = \frac{\mu}{\sigma}. \tag{31}$$

That is to say, the slope of the Reciprobit plot is determined by the precision of the reciprocal variable and the intercept is its inverse CoV. Notice that both the slope and the intercept are proportional to the precision of the reciprocal. This implies that there will necessarily be a strong correlation between Reciprobit slopes and intercepts. As we saw in our overview of the LATER model, these slopes and intercepts are respectively proportional to the average distance from starting level to threshold (μ_Δ) an to the average rate of information income (μ_r).

We need to be able to separately study effects on these two factors (slope and intercept). Fortunately, modern regression techniques enable this type of analyses. Several techniques exist that allow for separate effects on the location and scale parameters of a distribution. In particular, we will concentrate on the GAMLSS technique (Generalized Additive Models with Location, Scale, and Shape parameters; Rigby & Stasinopoulos, 2005) that presents a number of advantages, including a great flexibility of distributions to fit the response variable, and the possibility of considering random effects and different types of smoothers.

Using this technique, we can perform a regression analysis with two separate model formulae. On the one hand, one of the formulae refers to the factors that can influence the mean of a distribution, as in most traditional regression techniques. On the other hand, the additional formula analyzes the factors that affect the standard deviation of that distribution. If we take the Recinormal assumption that the reciprocal of the RT is distributed according to $N(\mu, \sigma^2)$, we can perform a regression analysis on the μ and σ parameters separately. More formally:

$$\begin{aligned} \mu &= \sum_i (\beta_i x_i) + \beta_0 + \epsilon_\mu \\ \log(\sigma) &= \sum_j (\gamma_j x_j) + \gamma_0 + \epsilon_\sigma, \end{aligned} \tag{32}$$

where the x_i and x_j are predictors related to our experiment, the β_i and γ_j are regression coefficients for each predictor (β_0 and γ_0 being intercept terms), ϵ_μ and ϵ_σ are error terms, and the log has been added as a link function to avoid negative values of σ. GAMLSS also allows to further break the error terms into different strata, allowing for the inclusion of subject- or item-specific random effects.

In this kind of regression, if a factor shows a significant β coefficient, and no significant γ coefficient, we will be able to infer that such a factor affects only the intercept b of the

Reciprobit plot. On the other hand, if a factor has a significant γ coefficient, this indicates that it may affect either the slope or the intercept of the Reciprobit plot, or *both*. However, by comparing its γ and β coefficients, we will be able to assess to what extent that effect is restricted to the intercept.

This analysis relies on the assumption that the data are recinormally distributed, that is, that they fall in the Recinormal zone of Fieller's distribution ($\lambda_1 < .22$). As we will see below, this is not always the case, in some experiments we will find that the individual subjects show also some significant degree of variation in Δ, leading to their responses lying in the linear zone of Fieller's ($.22 < \lambda_1 < .4$). Fortunately, in our own experience, even when in this linear zone (such as the word naming experiment discussed below), subject tend to be rather proximal to the Recinormal zone. Thus, reducing the level of intra-subject variation by the introduction of random effect of stimulus identity, and of inter-subject variation using random effect of subject identity, will lead to sufficient reduction in the value of λ_1 for this technique to be valid. In addition, as we will see in the data analysis section, significant departures from recinormality require very large datasets to be observable. With this in mind, it is advisable to verify that this is a sensible assumption before proceeding to the analyses. A simple way of doing this is to verify that the individual by-subject and by-item Reciprobit plots look approximately straight.

Empirical Evidence

We have extended the LATER model to be able to account for the typical psychological experiments, where different stimuli with different properties, are presented to different subjects in different experimental situations. In this section, we proceed to analyze experimental data to see if the predictions of the theory hold in real-life datasets, and how well does it compare to other proposals for RT distributions.

We will proceed in three steps. First, we will see that our theory can account for datasets where responses have been summarized across subject or items by taking by-item or by-participant mean response latencies. Next, we will continue by investigating the fully complex situation in which large datasets of single trials across participants and stimuli are individually considered. Here we will pay special attention to the behavior of the left and right tails of the distribution, where the predictions of our theory considerably diverge from previous proposals. In addition, we will investigate the differences that arise in experiments dominated by decisional components, in relation to experiments that are dominated by recognition components. Finally, we will illustrate the use of the methodological techniques derived from the theory to separate top-down and bottom up effects.

We will investigate four experiments which involve stimuli in two different modalities (visual and auditory), two types of responses (button presses and vocal) and – importantly – two different kinds of experimental tasks (decision-dominated and recognition-dominated). The datasets employed are summarized in Table 3.

Most of the properties of the theory that we are proposing are shared with several other proposed RT distributions, or at least small modifications on them could give rise to very similar features. The details in which our theory diverges more seriously from previous proposals concern mostly very rare events, that is, the tails of the distribution. In most analyses of RT distributions these tails receive little consideration, in many cases they are outright ignored. In order to have a detailed picture of what happens in these very short and

Table 3: Datasets used in the analyses.

Experiment	Language	Stimuli	Response	Dominant component	Number of items	Number of participants
visual lexical decision (Balota *et al.*, 2007)	English	visual	button-press	decision	37,424 (400)	816 (200)
word naming (Balota *et al.*, 2007)	English	visual	vocal	recognition	40,481 (400)	450 (200)
auditory lexical decision (Balling & Baayen, 2008)	Danish	auditory	button-press	decision	156	22
picture naming (Moscoso del Prado *et al.*, in prep.)	French	visual	vocal	recognition	512	20

very long reaction times, it is necessary to have very large bodies of data available. Fortunately, in recent years, some datasets of this type have begun to be collected. In particular, we will concentrate on the analysis of two recent massive datasets, the visual lexical decision and word naming datasets of the English Lexicon Project (Balota *et al.*, 2007). These datasets use very large numbers of different lexical stimuli, collected from many different individuals, in six different research centers. Therefore, we expect these datasets to present a sufficient number of very short an very long responses to enable the direct comparison of models.

Averaged datasets

We will start our assessment of the theory by analyzing datasets that have been summarized either by item or by participant, as these are the more common ways in which psychological data have traditionally been analyzed up to recent times. As we discussed in the previous sections, by removing part of the heterogeneity due to different participants and different stimuli, this averaging process should decrease the estimated value of both CoV parameters λ_1 and λ_2, putting the average RT distributions close to the Recinormal and/or normal areas. In fact, as already noticed by Carpenter (1981), there is a residual degree of variation in the rate of information intake, even in measurements of a single subject responding to a single type of stimulus. Therefore, we can expect the average distributions to lie in or close to the Recinormal zone, more often than to the normal zone.

In what follows, we study the details of the by-item and by-participant summaries of the datasets presented in Table 3. In order to put both English Lexicon Datasets in a scale that is comparable to a typical psychological experiment, we have performed the summarized analyses on subsamples of 400 randomly selected items for the by-item analyses, and 200 randomly sample participants in the by-participant analyses.

By-item analyses.

Figure 9 provides the Reciprobit plots and Box-Cox estimations for each of the four by-

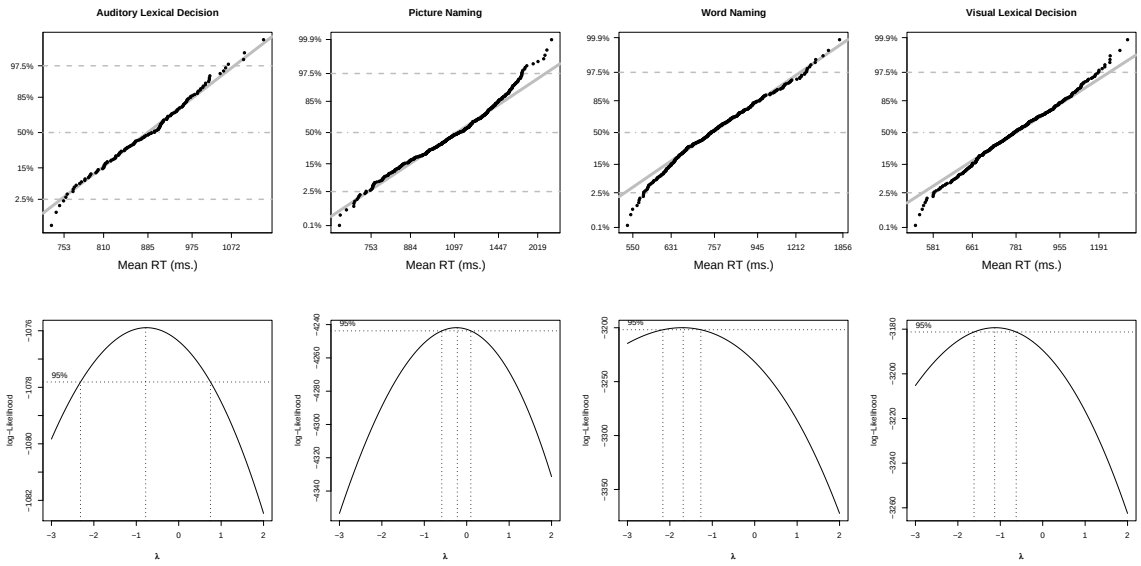

Figure 9. Reciprobit plots (*upper panels*) and log-likelihood (*lower panels*) of the power parameter in the Box-Cox transformation for each of the four by-item datasets. The horizontal lines in the Reciprobit plots represent the median and 95% intervals of each dataset. Recinormal distributions are characterized by straight lines in these plots. The vertical lines in the Box-Cox plots indicate the maximum likelihood estimates and estimated 95% confidence interval for the optimal value of the parameter. Values close to -1 are characteristic of Recinormal distributions, values close to 0 would be the signature of log-normal distributions. The estimations of the Box-Cox parameter have been obtained using function `boxcox` in *R* package `MASS` (Venables & Ripley, 2002).

item datasets. Notice that, although by inspection of the Reciprobit plots alone one would be guided to conclude that the four datasets are recinormally distributed, more detailed objective examination of the Box-Cox estimates gives a slightly different picture. On the one hand, the RTs in both "decision" tasks (*i.e.*, visual and auditory lexical decision) are very clearly Recinormal, with optimal values of the power parameter remarkably close to -1. However, when it comes to the "recognition" datasets, the picture appears less clear. The word naming dataset suggests an optimal power parameter of around -2, suggesting that the Recinormal might not produce such a good fit to the data. In particular, it appears that there are slightly less short latencies than one would expect under a Recinormal assumption (see the Reciprobit plot) Notice, however, that the difference in log-likelihood to -1 is still very small, and the optimal would have probably have been around -1 had we included prior information on the possible values of the parameter (*i.e.*, a value of -2 is theoretically meaningless). Still, this divergence could be indicative of not such a good fit to the data by a LATER-style model.

The picture naming dataset presents yet a different case. Here, the optimal value of the Box-Cox power parameter is close to zero, which would be the stamp of a log-normal distribution. Notice, that the log-normal is itself one of the candidate distributions for RT data (*e.g.*, Woodworth & Schlosberg, 1954), and is in fact the underlying assumption

Table 4: Comparison of estimated maximum likelihood fits to the four by-item averaged datasets. The Weibull, Log-normal, and Inverse Gaussian distributions have been fitted with an additional shift parameter whose value ranges between zero and the minimum of the dataset. Both parameters of the Recinormal, and the standard parameters of the Log-normal, and Inverse Gaussian distributions have been fitted using the analytical forms of their maximum likelihood estimations. Fieller's distribution, the Ex-Gaussian, and the shape and scale parameters of the Weibull were fit using a Nelder-Mead optimization of their log-likelihood functions. The additional shift parameters of the Weibull, Log-normal, and Inverse Gaussian were fit by separately optimizing the log-likelihood (with the other parameters determined above) using a quasi-Newton method with box constraints of the range of parameter values (method ''L-BFGS-B'' in *R* function `optim` – general `stats` package). The AIC rows show the Akaike's Information Criterion for each of the fits, and the BIC row lists the corresponding values of Schwartz's "Bayesian" Information Criterion. The numbers in bold indicate the best fits to each dataset according to each criterion.

Distribution	Stat.	Lex. Dec. (Visual)	Word Nam.	Pic. Nam.	Lex. Dec. (Auditory)	TOTAL
	Range (ms.)	502 – 1,370	540 – 1,788	673 – 2,848	736 – 1,167	
Ex-Gaussian	AIC	5,104.25	5,135.07	6,669.18	1,892.46	18,800.96
	BIC	5,116.23	5,147.05	6,681.70	1,901.61	18,846.59
Fieller	AIC	5,106.05	5,142.66	6,660.87	1,814.74	18,724.32
	BIC	5,122.01	5,158.63	6,677.57	1,826.94	18,785.15
Recinormal	AIC	5,101.46	5,152.64	6,660.55	1,810.81	18,725.46
	BIC	5,109.44	5,160.62	**6,668.90**	**1,816.91**	18,755.87
Weibull (3-param.)	AIC	5,097.57	5,147.50	6,664.09	**1,809.42**	18,718.58
	BIC	5,109.48	5,159.47	6,676.61	1,818.17	18,763.73
Log-normal (3-param.)	AIC	5,097.96	5,132.30	**6,656.78**	1,812.55	18,699.59
	BIC	5,109.94	5,144.27	6,669.30	1,821.70	18,745.21
Inverse Gaussian (3-param.)	AIC	**5,096.61**	**5,131.68**	6,657.06	1,812.45	**18,697.80**
	BIC	**5,108.59**	**5,143.66**	6,669.58	1,821.60	**18,743.43**

for the log transformation that is often applied in the analysis of RT data. Therefore, unlike the case of word naming, it would be difficult to discard the zero value on the basis of prior information on possible distributions. Further inspection of the Reciprobit plot reveals that most of the divergence from recinormality arises from an overabundance of the long responses. In fact, as it will be described in the detailed analysis of this dataset, this might have been a by-effect of the mechanism for data collection, which from a certain point implicitly introduced a pressure to respond.

We can conclude from this that, although the four datasets conform relatively well to the predictions of a Recinormal distribution, there may be some traces of a difference between decision dominated and recognition dominated processes, where the later ones result in slightly worse Recinormal fits. This could be interpreted as mild support for the differences argued for by Carpenter and Reddi (2001), according to which LATER only applies to decision processes.

We can also compare more explicitly different candidate distributions o fit the data. Table 4 compares the information criteria. The first thing to notice is that, overall, the three-parameter versions of the Log-normal and Inverse Gaussian distributions seem to provide

Table 5: Comparison of estimated maximum likelihood fits to both of the by-participant averaged datasets. The fits were obtained in the same manner as for the by-item datasets. The AIC rows show the Akaike's Information Criterion for each of the fits, and the BIC row lists the corresponding values of Schwartz's "Bayesian" Information Criterion. The numbers in bold indicate the best fits to each dataset according to each criterion.

Distribution	Stat.	Lex. Dec (Visual)	Word Nam.	TOTAL
	Range (ms.)	502 – 1,370	508 – 1,044	
Ex-Gaussian	AIC	2,617.53	2,464.90	5,082.43
	BIC	2,627.42	2,474.79	5,102.21
Fieller	AIC	2,620.56	2,461.78	5,082.34
	BIC	2,633.75	2,474.97	5,108.72
Recinormal	AIC	**2,616.56**	2,464.52	5,081.08
	BIC	**2,623.16**	2,471.12	**5,094.28**
Weibull (3-param.)	AIC	2,623.17	**2,457.07**	5,080.24
	BIC	2,633.06	**2,466.97**	5,100.03
Log-normal (3-param.)	AIC	2,617.11	2,459.49	5,076.60
	BIC	2,627.01	2,469.38	5,096.39
Inverse Gaussian (3-param.)	AIC	2,616.86	2,459.32	**5,076.18**
	BIC	2,626.75	2,469.22	5,095.97

a better fit to the datasets. However, there is in general a very high level of 'mimicry' between the distributions, with all distributions providing quite adequate fits to the data. Notice also that the Ex-Gaussian and Fieller's distribution seem to perform slightly less well. We believe that this is in part a side effect of the fitting procedures, while for the other distributions there exist exact maximum likelihood estimators of the parameters, for these two distributions we had to rely on an iterative optimization procedure on a 3 or 4 dimensional space. Notice also that these fits do not sugegst any kind of contrast between recognition-dominated and decision-dominated tasks.

By-participant analyses.

We proceed now to compare datasets that have been averaged by-participant. Unfortunately, only the large ELP datasets provide a sufficient number of individual participants to allow any reliable comparison of distribution fits. Therefore we restrict the analyses to these two datasets.

Figure 10 provides the Reciprobit plots and Box-Cox estimations for each of the two datasets summarized using by-participant means. As it was the case in the item-analyses, the Reciprobit plots seem to indicate that both datasets are recinormally distributed. Once more, in the Box-Cox estimates it appears that the decisional task (*i.e.*, visual lexical decision) is more clearly Recinormal than the recognition task (*i.e.*, word naming). The former has a typically Recinormal estimated power parameter of around -1, while the later seems more like a log-normal case, with power parameter around zero. As before, this could be interpreted as a suggestion that recognitional and decisional task might be different.

The picture provided by distribution fitting, as shown in Table 5 is rather similar to what we observed in the by-item means. All distributions seem to perform more or

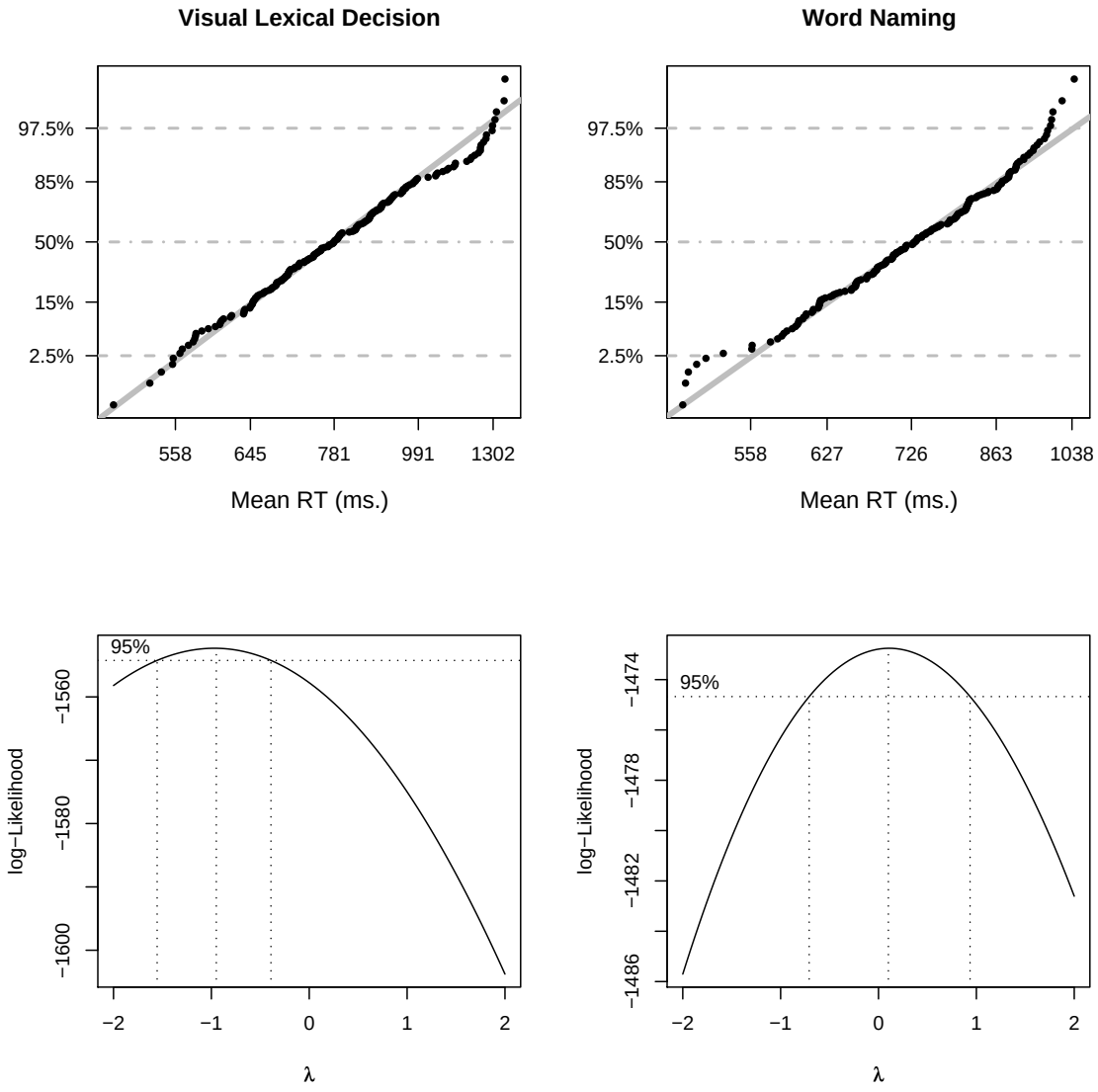

Figure 10. Reciprobit plots (*upper panels*) and log-likelihood (*lower panels*) of the power parameter in the Box-Cox transformation for each of the two datasets summarized by-participant. The horizontal lines in the Reciprobit plots represent the median and 95% intervals of each dataset. Recinormal distributions are characterized by straight lines in these plots. The vertical lines in the Box-Cox plots indicate the maximum likelihood estimates and estimated 95% confidence interval for the optimal value of the parameter. Values close to -1 are characteristic of Recinormal distributions, values close to 0 would be the signature of log-normal distributions. The estimations of the Box-Cox parameter have been obtained using function `boxcox` in *R* package `MASS` (Venables & Ripley, 2002).

less equally well on both datasets. Overall, the Inverse Gaussian and Log-normal provide slightly better fits than the other distribution overall. By particular tasks, it seems like the Recinormal or Fieller's provide the best fit to the decision data (note that the difference in AIC and BIC between the Recinormal and Fieller's is fully accounted for by the difference between 2 and 4 degrees of freedom), while the Weibull provides the best fits to the word naming data.

Large aggregated datasets

We have seen that neither the by-subject nor the by-item averaged datasets provides us with conclusive information that enables to discriminate between distributions (and the theories that they embody). Rather, we find that in this type of data, there appears to be a severe 'model mimicry' problem in the sense described by Ratcliff and Smith (2004) and Wagenmakers *et al.*, (2004). Averaging reduces the variation in the sample, placing the observations closer the the grand-mean. This has the consequence that it will become very difficult to see elements falling into the right tails. As we discussed in the theoretical section, it is precisely the right tail that provides the most discriminating information between distributions of different families. It becomes therefore clear that, in order to obtain contrastive information, we need to focus on the areas where the distributions make significantly different predictions, on their right tails. Making inferences on shapes of right tails requires as large datasets as possible. We now proceed to perform this comparison on large, individual response datasets. In this section we provide a detailed analysis of the aggregated datasets, that is, all RT measurements have been lumped together, irrespective of the participant or the stimulus. As we discussed above, if the data follow Fieller's distribution, by application of the Central Limit Theorem to the normal random variables in the numerators and denominators of the mixture, the aggregated data should also be described by an instance of Fieller's.

As we did with the averaged datasets, we begin our aggregated analyses by inspecting the Reciprobit plots and the estimated values of the Box-Cox power parameter. These are presented in Figure 11 for each of the four datasets. The first thing that one notices is that the shapes of the Reciprobit plots in the upper panels are dramatically different between the medium/small scale datasets (Auditory Lexical Decision and Picture Naming), than for the two massive datasets from the ELP (Visual Lexical Decision and Word Naming). On the one hand, both smaller datasets present a clearly Recinormal trace with straight lines on their Reciprobit plots. Only the slowest responses (*i.e.*, above 3 s.) from the picture naming dataset deviate from the main line. In fact, this corresponds to the responses when the participants implicitly received additional pressure to respond (as we discuss later, at this point the picture disappeared from the screen, although RT recording continued for 1 additional second). On the other hand, the two ELP datasets present the characteristic bi-linear pattern that Carpenter attributes to a separate minority populations of "express" responses. This corresponds to the lower slope lines depicted in each of the Reciprobit plots, which includes less than 5% percent of the data points in each dataset. This contrast between small and large datasets is also reflected in the Box-Cox estimates shown in the bottom panels of the figure. For both small datasets we estimate optimal values of the power parameter close to -1, as is characteristic of Recinormal distributions. However, the optimal estimates for the two large datasets are in fact close to a log-normal value of zero.

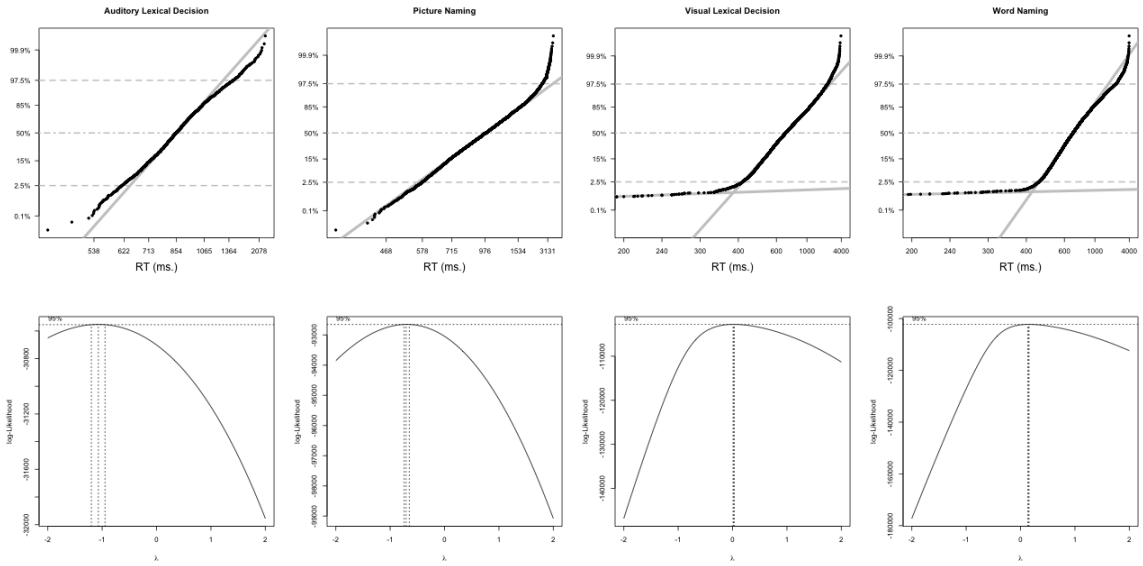

Figure 11. Reciprobit plots (*upper panels*) and log-likelihood (*lower panels*) of the power parameter in the Box-Cox transformation for each of the four individual trial aggregated datasets. The horizontal lines in the Reciprobit plots represent the median and 95% intervals of each dataset. Recinormal distributions are characterized by straight lines in these plots. The vertical lines in the Box-Cox plots indicate the maximum likelihood estimates and estimated 95% confidence interval for the optimal value of the parameter.

In addition, the shape of the log-likelihood is now changed, now taking high values also into the positive domain. Notice that, in this case, it becomes clear that the contrast between datasets has nothing to do with the recognition or decision component of the datasets, and it is solely determined by the mere size of the datasets.

We can also compare how well do different candidate distributions fit these aggregated data. Table 6 compares the quality of the best fits (fitted using KmD method described in Appendix D). In contrast with the averaged datasets, the three parameter versions of the Weibull, Log-normal, and Inverse Gaussian distributions cannot be used to fit these large datasets. As we have very early responses, the shift parameter would be forced to take a value of zero (otherwise the log-likelihood goes to minus infinity) and the fit provided by these distributions would be unacceptably low. Instead of the Inverse-Gaussian, we have included the Ex-Wald distribution proposed by Schwartz (2001), which is an Inverse Gaussian convoluted with an Exponential, and thus allows for a variable shift. We could not find any variable shift version of the Weibull, and we have thus not included it (fitting a 2 parameter version led to extremely poor fits). Finally, for reference purposes, we have also included a two parameter log-normal. The picture presented by the table is very similar to what we concluded from the Reciprobit plots and Box-Cox methods. Fieller's distribution is necessary to explain the large number of extremely long or short responses that happen in the large datasets. In the smaller datasets, where extreme responses (either long or short) are very unlikely to occur in a significative number, it appears that the Ex-

Table 6: Comparison of estimated maximum likelihood fits to individual trials in the four datasets. The fits were obtained in the same manner as for the by-item datasets.

Distribution	Stat.	Lex. Dec. (Auditory)	Lex. Dec. (Visual)	Word Nam.	Pic. Nam.
	Range (ms.)	446 – 2,327	1 – 3,997	1 – 3,997	370 – 3893
Ex-Gaussian	AIC	**43,555**	16,177,435	13,965,705	**105,427**
	BIC	**43,573**	16,177,471	13,965,740	**105,448**
Fieller	AIC	43,583	**16,151,249**	**13,830,087**	105,580
	BIC	43,607	**16,151,297**	**13,830,135**	105,608
Ex-Wald	AIC	48,709	16,599,825	14,658,385	108,633
	BIC	48,728	16,599,860	14,658,421	108,654
Log-normal (2-param.)	AIC	43,873	16,374,800	14,287,700	106,242
	BIC	43,886	16,374,824	14,287,724	106,256

Gaussian distribution does slightly better than Fieller's. However, the evidence from this datasets is much weaker than the evidence presented by the large ELP data. We can thus conclude that, in order to distinguish between different distributions, we need a large set of data points, so that events in the tails become sufficiently frequent. For large aggregated datasets, Fieller's distribution provides a significantly better fit than any of the alternatives.

We have seen that in terms of quality of fits, Fieller's distribution seems like a good candidate to account for the aggregated distributions of RTs in large datasets, both in recognition and decision tasks. An additional piece of evidence comes from the shape of the hazard functions. As discussed in the theoretical section, different distributions give rise to characteristic shapes of the hazard function (see Luce, 1986 and Burbeck and Luce, 1982 for details). On the one hand, both the Ex-Gaussian and the Weibull type distributions can only give rise to monotonic hazard functions. The Weibull hazard rates are either monotonically increasing or decreasing, with a flat function in the special case when the Weibull reduces to the exponential. The Ex-Gaussian has an initial monotonically increasing phase, which is followed by a flat (neither increasing nor decreasing) phase that arises in the left tail, where the exponential component becomes dominant. On the other hand, distributions of the type of the Inverse Gaussian, the Log-Normal, or the time distributions generated by different types of drift diffusion models have peaked hazard functions, with a monotonically increasing and a monotonically decreasing – in some cases flat – phase.

Figure 12 presents the estimated hazard rates (using the method described in Appendix E) for the four datasets under consideration. The two small datasets show slightly peaked hazard functions. Notice however, that the peaks seem very weak. In our own experience, if one generates Ex-Gaussian distributed random numbers, and then re-estimates the hazard function from the generated points, one often finds that the estimators have produced small peaks of the kind found in both small datasets. Therefore, these hazard estimates could be consistent both with monotonically increasing and with peaked hazard rates. The large datasets however, provide a much clearer peak, followed by decreasing phases. These cannot be the consequence of a monotonic hazard function. Therefore, they provide strong qualitative evidence against a Weibull or Ex-Gaussian distribution, much fa-

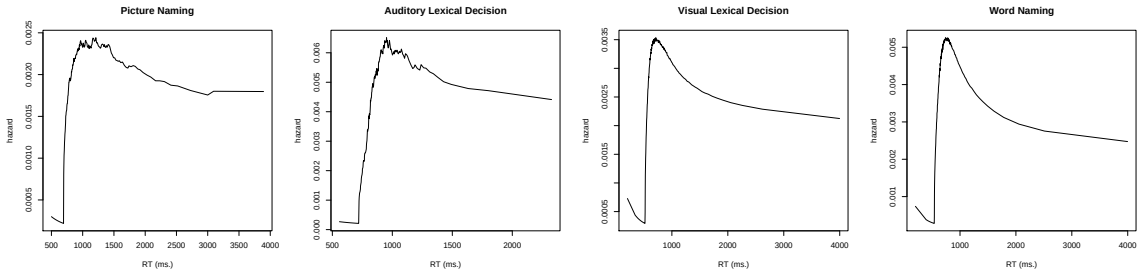

Figure 12. Estimated hazard functions for each of the aggregated datasets.

voring a peaked type distribution (*e.g.*, Log-normal, Inverse Gaussian, Recinormal, Fieller's, *etc.*).

We have discussed in the theoretical sections that, with respect to the tails, our theory predicts two clear things: there will be a higher number of anticipations with respect to other theories, and the log right tail of the distribution should follow a heavier than exponential pattern (*i.e.*, linear in log-log scale), rather than the linear decrease that would be predicted by distributions with exponential tails.

Figure 13 compares the quality of the fits provided by the Ex-Gaussian (dark grey solid lines), Ex-Wald (light grey solid lines) and Fieller's distribution (black solid lines) to the visual lexical decision (upper panels) and word naming (lower panels) datasets from the English lexicon project.[10] The right panels show that, when comparing these estimates of the density with a Gaussian KDE of the same data (dash-dotted grey lines), both distributions seem to provide very good fits, with hardly any difference between them, although the Fieller's fit already seem a bit better. However, when one examines in detail the log-densities of the distributions, one finds that the Ex-Gaussian fits radically diverge from the KDE estimates at both tails. Here, the Ex-Gaussian distribution underestimates the densities by many orders of magnitude (*i.e.*, logarithmic units). In contrast, Fieller's distribution provides an excellent fit of both datasets up to the far right tail, and a significantly more accurate fits of the left tail. Similar to the Ex-Gaussian, the Ex-Wald distribution also shows too light tails relative to the data.

The problems of using exponential tail distribution as a model of aggregated RTs is further highlighted by Figure 14. The figure compares on a log-log scale the fit of a power-law tailed distribution (Fieller – solid black lines), and an exponential-tailed distribution (Ex-Gaussian, solid grey lines, we have not plotted the Ex-Wald fits as they were clearly worse in all aspects) to the Gaussian KDE estimates for the lexical decision (top panel) and word naming datasets (bottom panel) – . The log-log scale emphasizes the problem of truncating the distributions. The vertical dotted lines show typical truncating points at 300 ms. and 2,000 ms., as recommended by Ratcliff (1994). Notice, that within that interval, there is basically no difference between exponential-tailed and power-law distributions. It is however precisely beyond these cutoff points where one finds information that can reliably discriminate between both types of distributions (and the underlying models that each

[10]The distributions were fit using the KmD technique described in Appendix D.

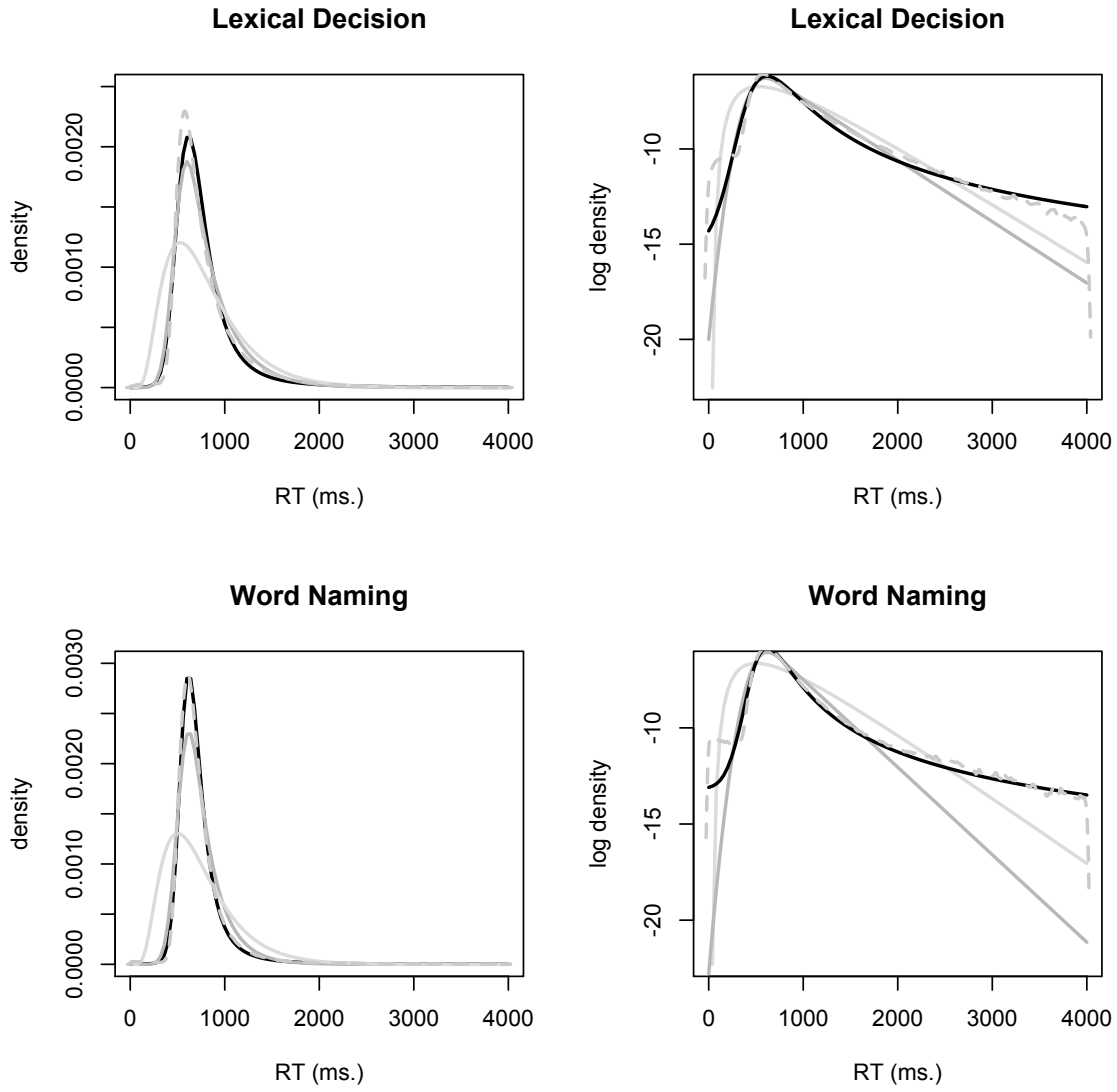

Figure 13. Comparison of the fits provided by Fieller's distribution (black solid lines), the Ex-Gaussian distribution (grey solid lines), and the Ex-Wald distribution (light grey solid lines) to the KDE estimates (grey dashed lines) of the aggregated visual lexical decision (*top panels*) and picture naming (*bottom panels*) latencies from the English lexicon project. The left panels show the estimated densities, and the right panels show the corresponding log-densities. Notice that the differences on the tails are only visible on the log-scale plots.

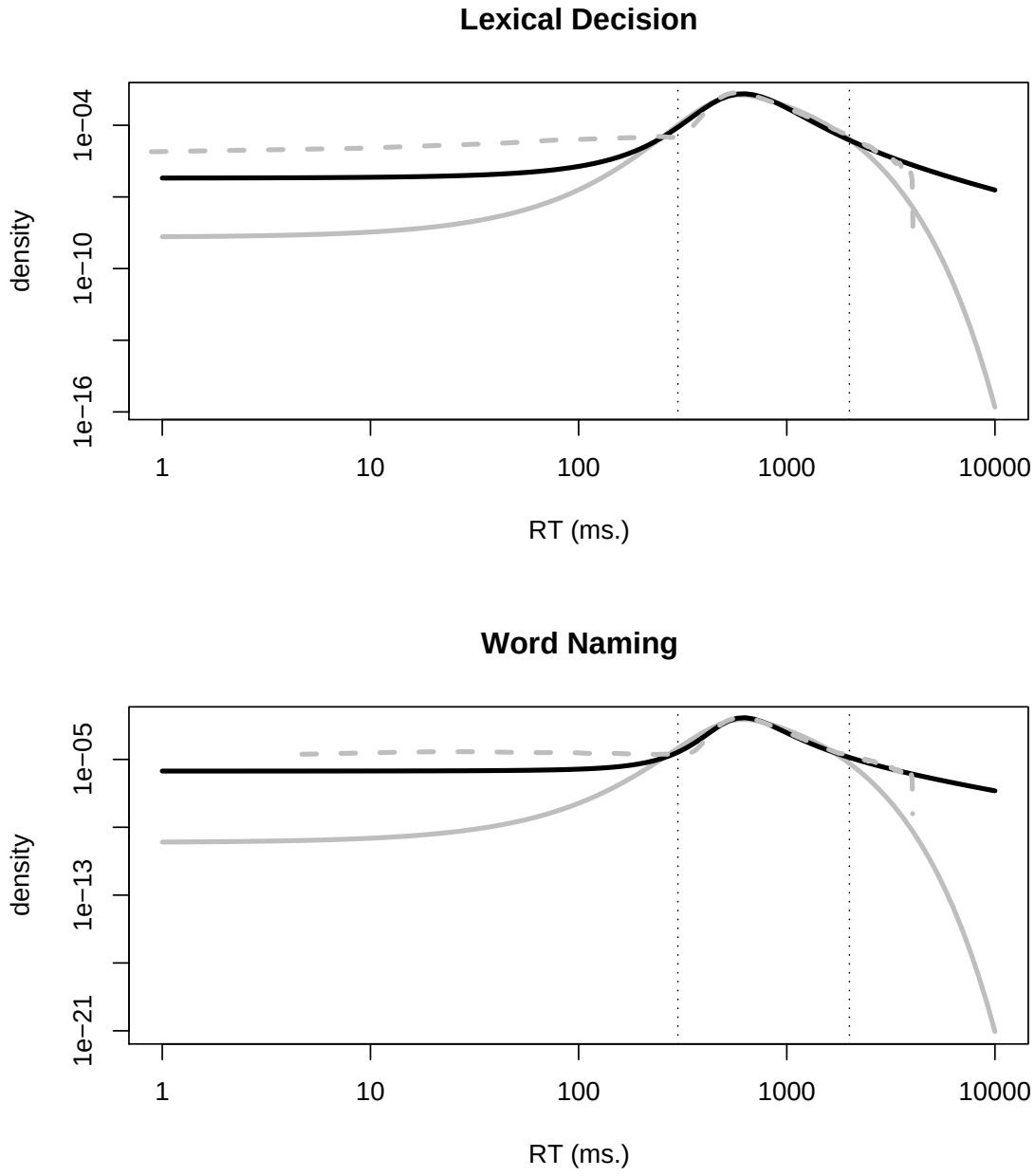

Figure 14. Log-log scale comparison of the fits provided by Fieller's distribution (black solid lines) and the Ex-Gaussian distribution (grey solid lines) to the KDE estimates of the aggregated visual lexical decision (*top panel*) and picture naming (*bottom panel*) latencies from the English lexicon project. The vertical dotted lines indicate typical cut-off points of 300 ms. and 2000 ms. The fits have been extrapolated up to 10,000 ms. to stress the different predictions that each makes.

implies).

We have seen that, when analyzing aggregated data, Fieller's distribution seems to outperform all other candidate distributions. This is confirmed by quantitative comparisons of the distributions (the information criteria of the different fits), and qualitative analyses of the shapes of both hazard functions and of the right tails of the distribution in log and log-log scale. The very slow responses on the right tails are certainly a very small minority of responses, on the order of 3% in the current datasets, and are in most studies discarded as outliers. However, the sheer size of both the ELP datasets makes that even such a small percentage of responses amounts to several tens of thousands, enabling a rather accurate estimation of their distribution. Looking at the distribution fit by the KDE estimates, it becomes clear that these points are in no way outliers. Rather, they follow a clearly regular law. Had they really been outliers, the KDE estimates would not look as the clean lines that appear in the graphs, but more like a very bumpy curve or one with a great amount of high frequency oscillations. It is clear that removing these points results in a serious loss of useful information, and this information is precisely the one that enables discrimination between exponential-tailed distributions such as the Ex-Gaussian or the Ex-Wald and power-law distributions such as Fieller's. In addition, the aggregate data show clearly peaked hazard functions, which could not be accounted for by neither Weibull-type nor Ex-Gaussian type distributions. We conclude from this aggregate data analysis that Fieller's distribution provides a good description of the joint distribution of responses across subjects and items. This is consistent with our prediction that the aggregation of Fieller's distribution should result in another instance of Fieller's.

Individual participants analyses of large datasets

In the previous section, we have validated that the aggregate distribution of data is in accord with Fieller's distribution. However, this is in a way indirect evidence in support of the theory. It could well be the case that, although the aggregate RTs are Fieller distributed, the responses by each individual participant are not. For instance, a few atypical participants producing more long responses than the rest could have bent the tail of the aggregate distribution.

We now analyze in more detail the distribution of responses of each individual subject in the ELP datasets. In these datasets, each subject responded to a relatively large number of words (an average of 1,414 correct responses to words per subject in the lexical decision dataset, and of 2,323 correct responses per subject in the picture naming dataset), thus enabling separate fits to each subject. Table 7 summarizes the results of fitting distributions individually to each subject. For simplicity we have only included the two distributions that produced the best fits for both datasets, Fieller's and the Ex-Gaussian, as they provide examples of distributions with power-law (Fieller's) and exponential tails (Ex-Gaussian). From the tables, it appears that both in the lexical decision and in the word naming datasets, Fieller's distribution overall outperforms the Ex-Gaussian in terms of average quality of fit. However, the lexical decision averages are misleading. Notice that, although in the mean, Fieller's distribution appears to provide a better fit to the data, further examination of the paired *median* difference reveals that both distributions are even, in fact with the possibility of a slight advantage for the Ex-Gaussian. The origin of this discrepancy lies in the distribution of the subject-specific differences between the

Table 7: Comparison of the estimated AIC and BIC for participant-specific maximum likelihood fits of the Ex-Gaussian distribution and Fieller's distribution, across all participants in the ELP visual lexical decision and word naming datasets for which both fits converged. Positive values in the difference row favor Fieller's fits, and negative values favor the Ex-Gaussian.

Distribution	Statistic	Lexical Decision		Word Naming	
		Mean ± Std. error	Median	Mean ± Std. error	Median
Ex-Gaussian	AIC	$19,151 \pm 56$	$19,202$	$30,351 \pm 98$	$30,433$
	BIC	$19,167 \pm 56$	$19,217$	$30,368 \pm 98$	$30,450$
Fieller	AIC	$\mathbf{19},000 \pm 77$	$\mathbf{19},158$	$\mathbf{30},086 \pm 98$	$\mathbf{30},182$
	BIC	$\mathbf{19},021 \pm 77$	$\mathbf{19},179$	$\mathbf{30},109 \pm 98$	$\mathbf{30},205$
Ex-Gaussian - Fieller (paired)	AIC	$+152 \pm 55$	-12	$+265 \pm 32$	$+155$
	BIC	$+146 \pm 55$	-17	$+259 \pm 32$	$+150$
Number of participants		759		432	
Correct resps. / participant		1,414		2,323	

information criteria for both fits. While in the picture naming dataset there was a clear preference for the Fieller's fit in most subjects, in the lexical decision datasets there was a huge inter-subject variability on the differences between estimated fits.[11] We confirmed this interpretation using linear mixed effect model regressions with the estimated AIC values as dependent variables, including fixed effects of distribution (Fieller's *vs.* Ex-Gaussian), a random effect of the subject identity, and a possible mixed-effect interaction between the distribution and the subject. In the picture naming data there was a significant advantage for Fieller's fits ($\widehat{\beta} \simeq 265$, $t = 8.4, p < .0001$, $\widehat{p}_{mcmc} = .0234$) and no significant mixed effect interaction between the participants and the fixed effect ($\chi^2_{6,2} = .61$, $p = .74$). In contrast, in the lexical decision data there might have been a slight trend in favor of the Fieller's fits ($\widehat{\beta} \simeq 151$, $t = 2.38$, $p = .0174$, $\widehat{p}_{mcmc} = .2150$) but it did not reach significance according to a Markov Chain Montecarlo estimate of the p-value, and there was a clear mixed effect interaction between subject identity and preferred distribution ($\chi^2_{6,2} = 73.55$, $p < .0001$).[12]

We interpret the above results as clear evidence in favor of Fieller's fits in the picture naming datasets, but roughly equal performance in the lexical decision dataset – if anything, a marginal advantage for Fieller's fits – and substantial differences across subjects. This is not difficult to understand. The lexical decision datasets included much less responses than the picture naming ones, and it is thus less likely for a subject to elicit relatively long responses than it is in the larger samples of the word naming dataset. As the main

[11] The actual distribution of AIC differences in lexical decision dataset is extremely thick-tailed, in fact very much resembling a scale-free Cauchy – centered at the median, and with a half-width at half-maximum a bit over one third of the sample standard error – for which it is pointless to estimate sample means and standard deviations. In contrast, for the word naming dataset this distribution very much resembled a one-tailed exponential on the Fieller side with but a few outliers supporting the Ex-Gaussian.

[12] We report both t-based (p) and Markov Chain Montecarlo estimates ($\widehat{p}_{mcmc}$) of the p-values because we found the former to be too lax in this dataset, as it can be observed in the estimates for visual lexical decision regression – see Baayen *et al.*, (2008) for a detailed discussion of this issue. The response variable AIC was squared prior to the analysis, as a Box-Cox transformation estimate suggested this would be most adequate. In addition, to avoid numerical error from large numbers, the AIC values were divided by 10,000 prior to squaring. The effect estimates ($\widehat{\beta}$) provided have been back-transformed to the original AIC scale.

difference between Fieller's distribution and the Ex-Gaussian is found in the heavier right tails, only subjects that showed some of the very rare long RTs would be better accounted for by Fieller's. This is indeed the case, Figure 15 shows that the relative dispersion of the maximum RT produced by each subject (*i.e.,* a z-score of the maximum produced by each subject relative to the subject-specific mean and standard deviation RT) was positively correlated with the magnitude and sign of the difference in AIC (Kendall's $\tau = .264$, $z = 10.883$, $p < .0001$), that is to say, subjects with more extreme responses were better fitted by Fieller's. However, as shown by the averages, the relative support in favor of Fieller's/power-law tail from the subjects that showed some long responses is proportionally much stronger than the support in favor of the Ex-Gaussian/exponential tail by the subjects that did not elicit any long responses.

Figure 16 illustrates the estimated RT distributions of a 'prototypical' subject in each of the tasks. To obtain these curves, we estimated the cumulative density functions of the RTs individually for each subject in each task (without any smoothing). From these we interpolated 50 points from each participant (the grey points in the figures) uniformly sampled in the interval between 0 ms. and 4000 ms. In order to do this, we fixed the values of the cumulative density at zero at 0 ms, and at one at 4000 ms. to enable extrapolation outside an individual subject's range of responses.[13] The interpolated probabilities were probit-transformed, and we performed a non-parametric locally weighted regression on the probit values. Finally, the resulting smoother in probit-scale was back-transformed to standard normal probability density scale, and then renormalized to integrate to one in the interval from 0 to 4000 ms. This results in the solid black lines in the figures approximating the density of RTs of an ideal 'average' subject. The dashed lines on the logarithmic plots are linear regressions on the log-tails, used to underline how both ideal distributions deviate from an exponential tail (which would fall onto the straight lines) that would be characteristic of most usually advocated RT distributions. Notice also that, in consonance with the individual subject analyses, the deviation from exponentiality is more marked (starts earlier) in the picture naming than in the lexical decision dataset.

Using these prototypical densities we can also inspect their corresponding hazard functions (see Figure 17). Note that both estimated hazard functions are of the peaked type (although the peak is admittedly lighter in the lexical decision curve). Only distributions that can have peaked hazards could account for these data. Therefore, the evidence from hazards also seems to rule out Ex-Gaussian and Weibull type distributions to account for the data.

Interpretation of the parameter values of Fieller's distribution

Above we have seen that Fieller's distribution presents an overall advantage over the other candidates to account for the distribution of RTs for individual subjects in terms of quality of fits, shape of the right tails, and hazard functions. A crucial point about this distribution is that its estimated parameter values are informative as to the properties of the task. We now proceed to interpret the estimated parameter values.

[13] Estimation without extrapolation would have overestimated the densities at the right tail, as these would be estimated only from the subjects that produced them, ignoring that most subjects in fact did not. This would exaggerate the power-law appearance, biasing in favor of Fieller's distribution.

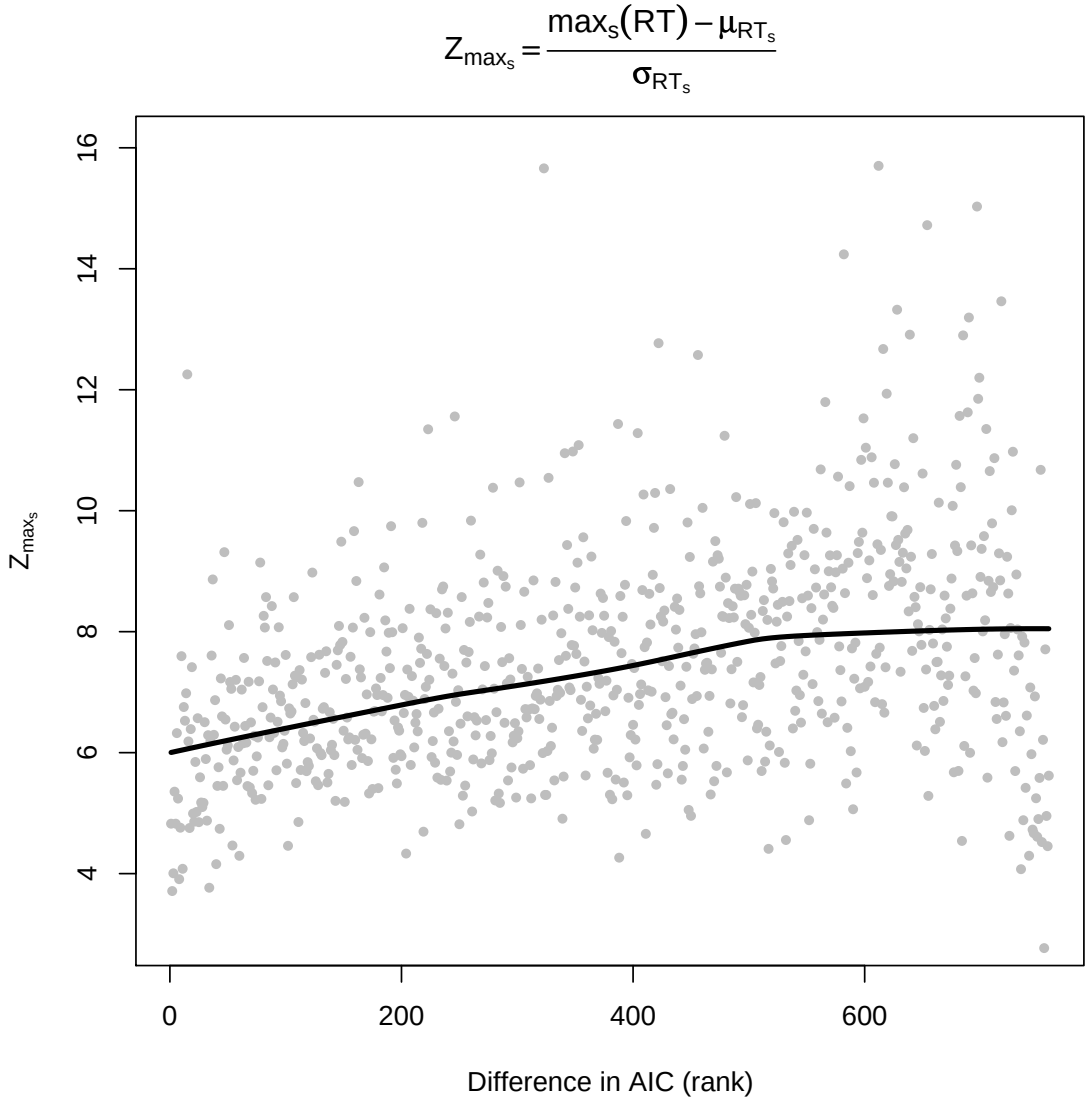

Figure 15. Correlation between the advantage of Fieller's over the Ex-Gaussian fits with the relative size of the maximum RT for individual participants in the ELP lexical decision dataset. The horizontal axis plots the rank of the difference in AIC between both fits. High values indicate a high preference for Fieller's, and low values indicate a higher preference for the Ex-Gaussian. The vertical axis plots the dispersion (distance from the subject mean measured in standard deviations) of the maximum RT value for each subject. The solid line is non-parametric regression smoother.

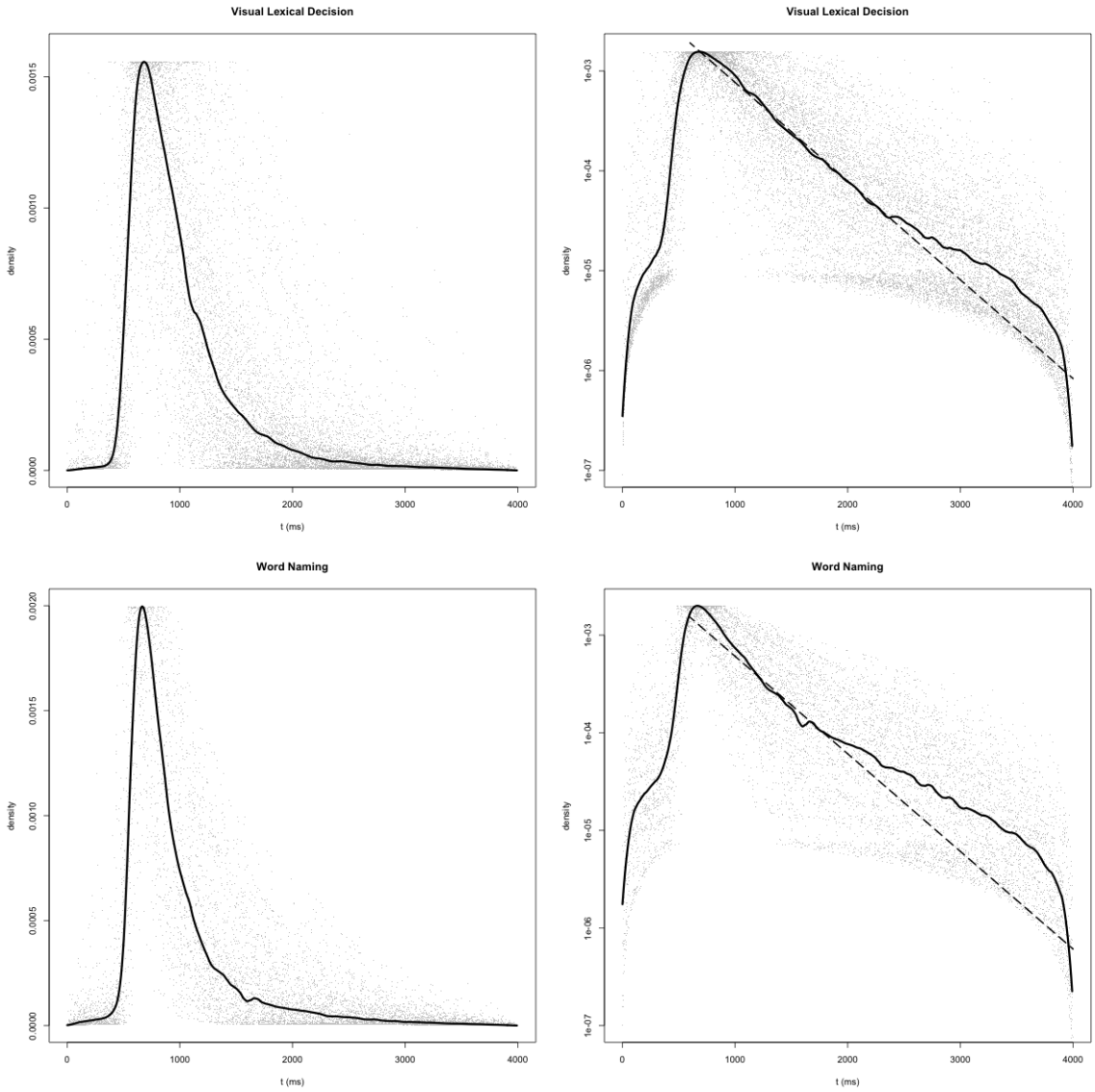

Figure 16. Ideal 'prototypical' subject in the lexical decision (*top panels*) and word naming (*bottom panels*) datasets. The left panels depict the densities, and the right panels are their equivalents in log-scale. The grey points are samples of 50 density points for each subject. The solid black lines plot the estimate prototypical density. The dashed black lines in the logarithmic plots correspond to linear regressions on the log right tail, showing how an exponential tailed fit to these data should look like.

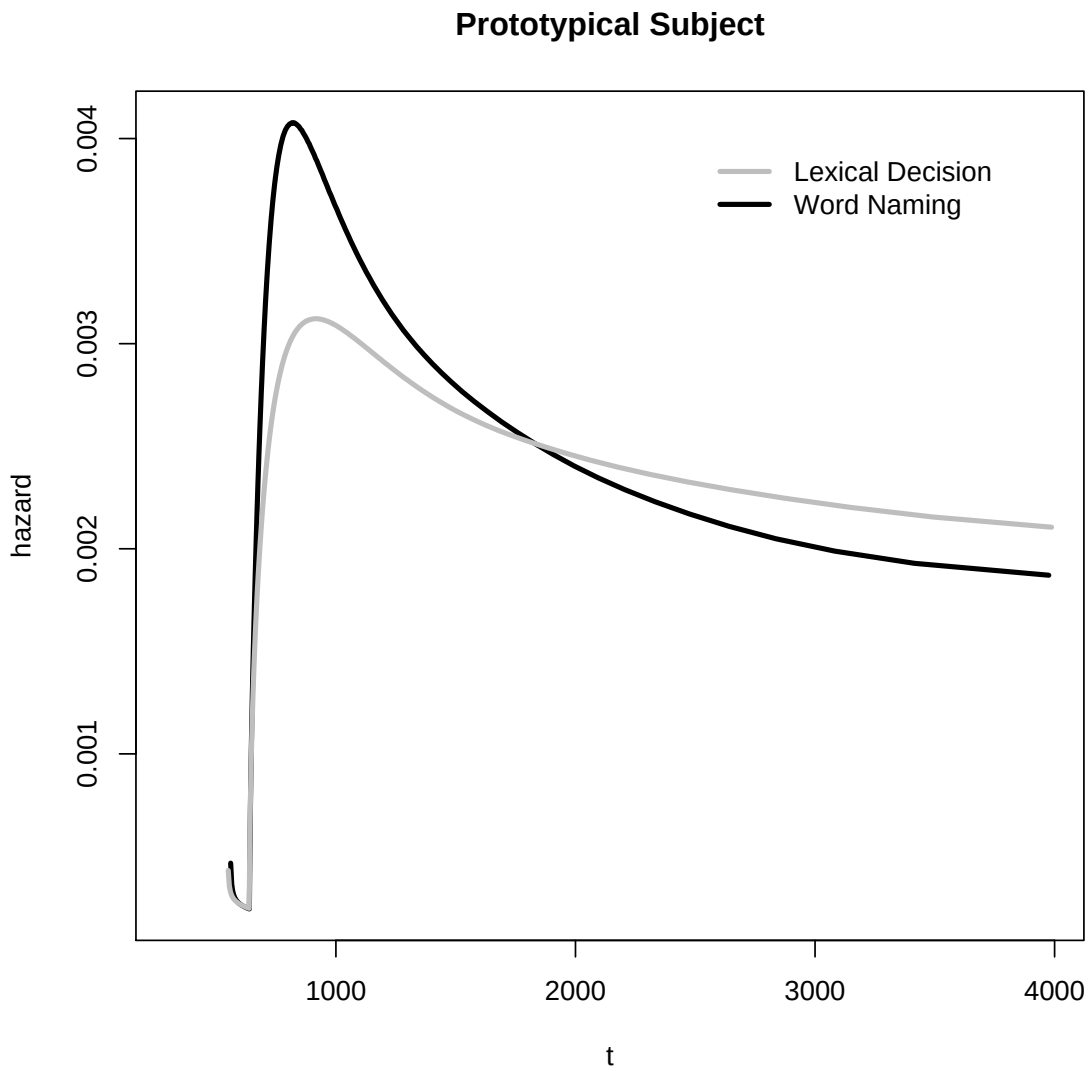

Figure 17. Estimated hazard function for the 'prototypical' subjects (see Figure **??**). The curves where estimated using the non-parametric method described by Burbeck and Luce (1982) on the estimated quantiles of the prototypical distributions.

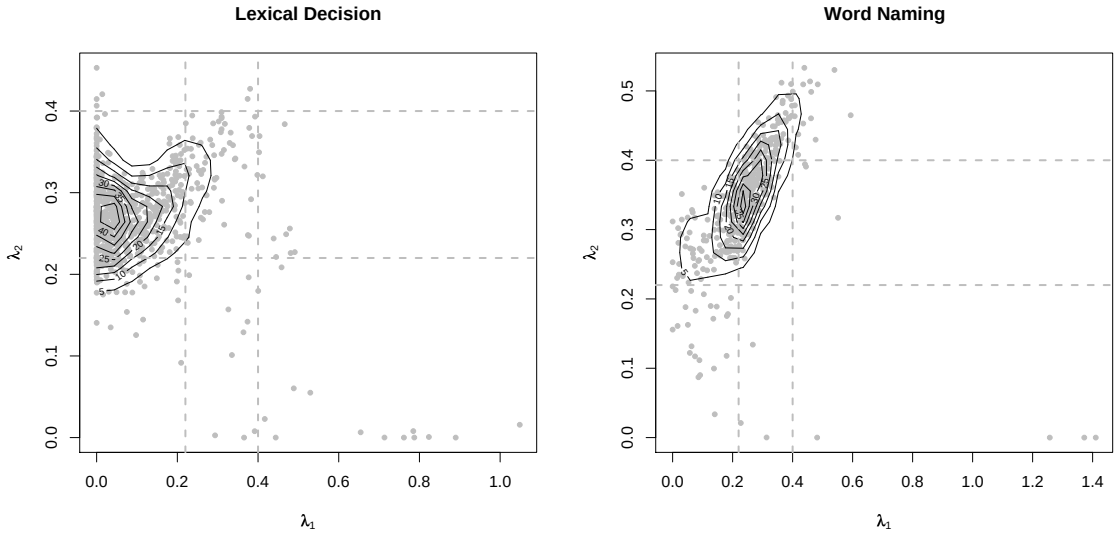

Figure 18. Values of the CoV parameters λ_1 and λ_2 obtained by fitting Fieller's distribution (using the KmD method (see Appendix D) to each individual participant in the visual lexical decision (*left panel*) and word naming datasets (*right panel*)of the ELP. Each point represents the fit obtained for an individual subject. The contours represent a 2-dimensional Gaussian kernel estimation of the density (obtained using function `kde2d` in *R* package `MASS`; Venables & Ripley, 2002). The horizontal and vertical grey dashed lines indicate the phase-change boundaries of Fieller's distribution. Points lying outside the centered 95% with respect to either λ_1 or λ_2 have been excluded from both graphs in order to avoid the large outliers resulting from non-converging fits.

The estimated values of the parameters of the Fieller fits to the aggregated data were ($\widehat{\kappa} = 695\,\text{ms.}, \widehat{\lambda}_1 = .27, \widehat{\lambda}_2 = .38, \widehat{\rho} = .6$) for the lexical decision dataset, and ($\widehat{\kappa} = 681\,\text{ms.}, \widehat{\lambda}_1 = .40, \widehat{\lambda}_2 = .44, \widehat{\rho} = .84$) for the word naming dataset. This puts both datasets in the linear zone in Fieller's distribution. However, the relatively high value of $\widehat{\lambda}_1$ for the word naming dataset in fact makes this distribution approach the Cauchy zone. This is indicative of a very high variability in the numerator of the ratio that gives rise to the distribution. If, following Carpenter and Williams (1995) we attribute this variability to variability in prior expectations, this fact becomes meaningful. While in the lexical decision experiment there were only two possible responses, which were matched in prior probability, in the word naming dataset different words will have a different prior expectation, causing a much greater variability across items. As we can see, this difference in the tasks is readily reflected in the fits of Fieller's distribution. This last issue is explored in more detail in Figure 18. The figure displays the estimated values of the λ_1 and λ_2 parameters in the separate individual subject fits. In the lexical decision data, the typical participant will show an estimated λ_1 value of around .05, with the vast majority of participants having an estimated value below the critical .22. This indicates that, in visual lexical decision the responses of each individual participant are well-described by a recinormal distribution, and thus the larger value of λ_1 in the overall fit is due only to inter-subject variation in threshold

or resting levels.

The situation is different in in the word naming participants. In this dataset the typical participant shows an estimated λ_1 just above the critical .22, already into the linear zone of Fieller's distribution, with a great proportion of the participants being significantly above this value. This indicates that in this case, there is a much greater heterogeneity in the threshold or resting levels from item to item. Interestingly, there is also a clear correlation in between the estimated λ_1 and λ_2 values ($\rho = .76$, $t(421) = 24.09$, $p < .0001$). This correlation reflects the interrelationship between the top-down and bottom-up properties of the stimuli (*e.g.*, frequency and word length). In these experiments, each participant saw a different subset of the estimuli, and thus there will be variation in both top-down and bottom up properties of the estimuli and these seem to be related to each other. In sum, variation in the prior probability of stimuli makes the intra-participant values of λ_1 greater in word naming than in visual lexical decision. In contrast, the estimated values of λ_2 are very similar in both experiments, being either just below or just above .3 in each experiment, indicating that both experiments exhibit a similar degree of variation in the bottom-up/perceptual properties of the stimuli, which are indeed identical in both experiments.

Distributions of correct and incorrect responses

As we have seen, the right tails of the distributions in both datasets are significantly thicker than one would predict by any theory that relies on an exponential-tailed distribution, and seem more prone to be described by theories that propose a power-law type of right tail (perhaps with a cutoff). However, as we noted in the theoretical section, distributions of the stretched exponential type, as is the Weibull proposed by Logan (1988), can also give rise to heavier than exponential tails. Furthermore, when discussing the distributions of correct and incorrect responses under a 'pure race' model, as the ones proposed by Brown and Heathcote (2005; 2008). In our theoretical analysis we advanced that these distributions would still predict too thin tails, below linear in log-log scale. We now proceed to investigate the distributions of correct and incorrect reponses that would arise using a race model as described by (19) and (22). For this, we investigate in more detail the conditional distributions of correct and incorrect responses to words in the ELP visual lexical decision dataset.

Figure 19 illustrates the RT distribution that would be predicted by a race of independent accumulators. The right panel shows that a relatively good fit of the density is obtained in comparison with Gaussian kernel density estimators (KDE) of the same distributions. However, when one examines in detail the quality of the fit in logarithmic scale (left-panel), one finds that the lack of inhibition has led to three important problems. The first of these problems is that the 'pointiness' of the mode is lost, giving rise to the more bell-shaped profile characteristic of a Weibull or Gamma distribution. The second is that, as predicted, the lack of any inhibition process has considerably thinned the right tail of the distribution, grossly underestimating the log-probability of responses above 1500 ms. Finally, the third problem lies in the diverging ratio of errors to correct responses. Whereas the empirical data seem to have a constant ratio (save for the very fast 'express' responses) of errors to correct responses, apparent in the parallel pattern of the KDE-estimated densities, the model desities have instead an initial diverging phase.

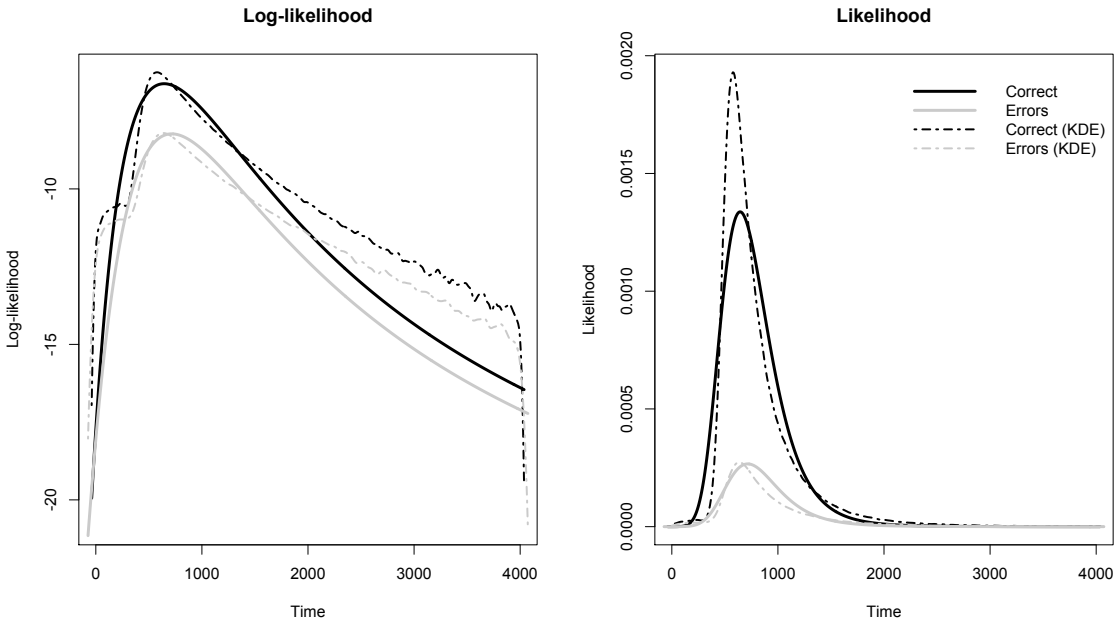

Figure 19. Lexical decision data from the ELP ($\sim$ 1.3 million individual responses) fitted as an inhibition-free competition between accumulators. The solid lines represent the predicted densities (right panel) and log densities (left panel) of the fitted model (the model was fitted with fixed variances for both accumulators). The discontinuous lines plot Gaussian kernel density estimators for log-densities and densities. Black lines plot correct responses, and grey lines plot error responses.

Figure 20 shows the effect of considering that, due to the compensation of the competition that would be provided by inhibition and decay mechanism, both the distributions of correct responses and errors can be modelled as plain instances of Fieller's distributions. To enable direct comparison with the fits of Figure 19, the parameters were fitted under identical constraints of equal variances and thresholds for both accumulators. Note that the three problems that were apparent in the free competition are greatly attenuated. The pointiness around the mode is now clear, and the ratio of errors to correct responses is now more or less constant. Finally, the fit of the right tail of the distribution is now very precise even in the logarithmic scale. There is still an apparently excessive 'bending' of the left tail relative to the KDE fits, but most of this is actually due to the population of very fast responses which is visible in the shoulder of the left tail of the logarithmic plots.

Very early responses

As can be appreciated in Figures 13 and Figure 14, neither the our distribution nor the exponential tail variants accurately models the very fast responses on the left tails of the curves. Even though Fieller's distribution provides a much better fit of these points also, it is still around two orders of magnitude below the KDE estimate from the data. Once again, despite being very rare (around 1% of the data counting all responses faster than 250

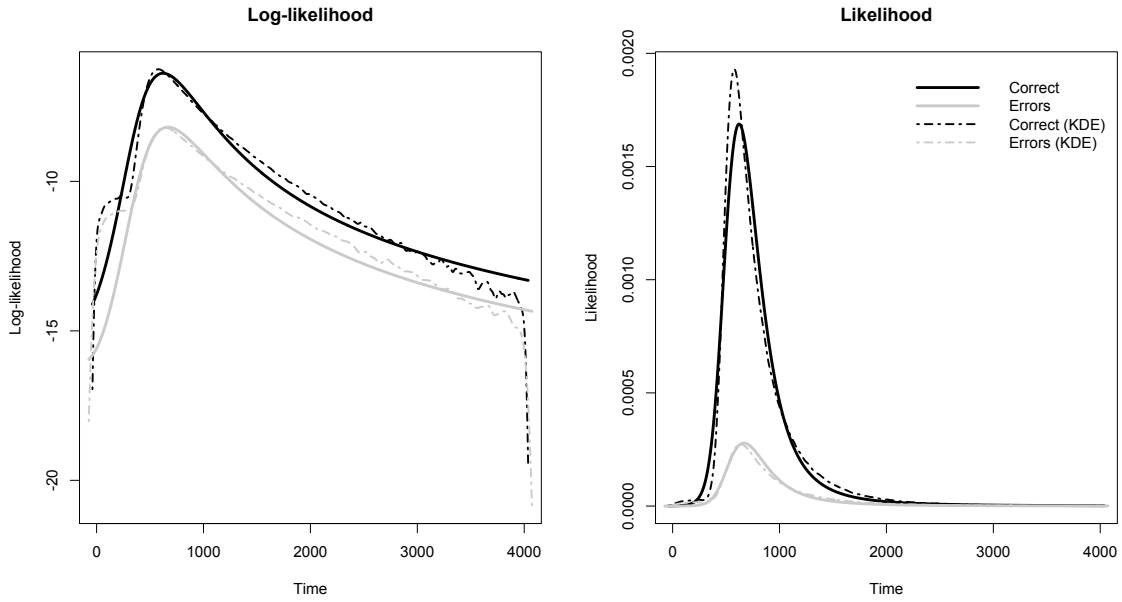

Figure 20. Lexical decision data from the ELP ($\sim$ 1.3 million individual responses) fitted as a competition including inhibition between accumulators. The solid lines represent the predicted densities (right panel) and log densities (left panel) of the fitted model (the model was fitted with fixed variances for both accumulators). The discontinuous lines plot Gaussian kernel density estimators for log-densities and densities. Black lines plot correct responses, and grey lines plot error responses.

ms.), there are still a large enough number (around 13,000 in each dataset) of these short responses to provide sufficiently good estimates of their distributions by KDE. However, it is evident in both logarithmic plots that these points form clearly separate 'bumps' in the log-density fit, giving rise to obvious shoulders in the distributions. In turn, this suggests that these points, or at least a great proportion of them, are indeed outliers in the sense that they originate from a different distribution than the one generating the rest of the points – they are generated by another process. Therefore, as we advanced above, these can indeed correspond to the 'express' responses hypothesized by Carpenter and Williams (1995) and Reddi and Carpenter (2000). Two things are noteworthy though. First, these responses are truly a minority. Most of the short responses that Carpenter and colleagues attribute to separate processes are in fact part of the general RT distribution and there is therefore no reason to believe they came from a different process. The second issue is that these responses are in fact not completely random. That is, even though they are very short, they are still more accurate than one would expect by chance. There are a total of 7437 correct responses and 4701 erroneous ones below 250 ms. This is a significant difference ($\chi^2_1 = 616.72, p = .0000$).

The data presented here correspond to the words in the ELP lexical decision dataset. The above-chance level of correctness of the very short responses could be due to the participants having an overall bias favoring 'yes' responses, even if the experiments had

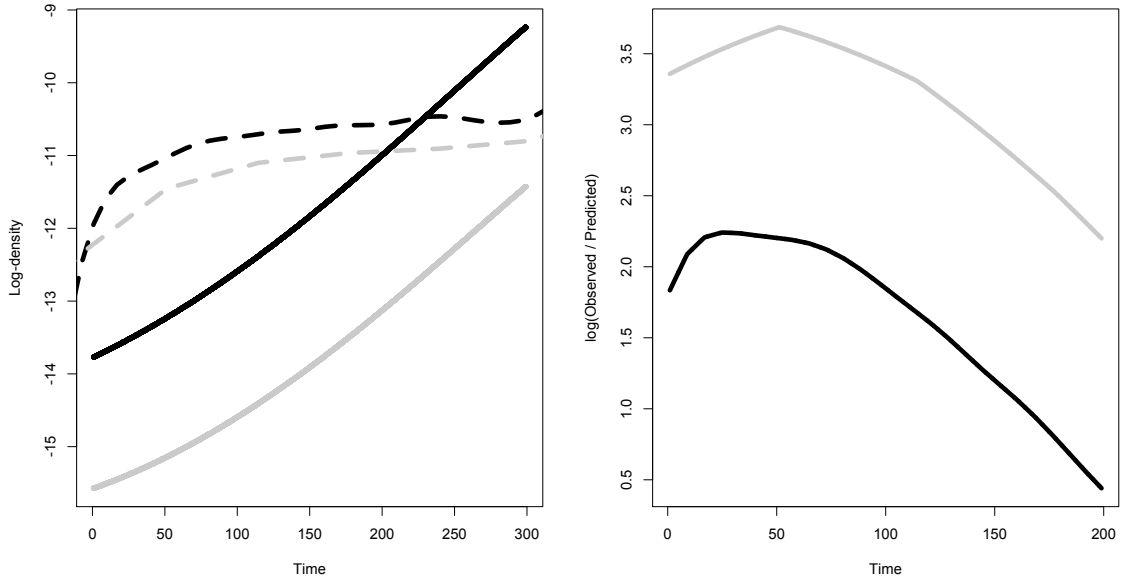

Figure 21. Early visual lexical decision responses.

been balanced in the number of words and pseudo-words that were presented. In fact, analyzing the pseudo-words together with the words one finds that there was indeed a bias: participants responded 'no' significantly more often than they responded 'yes' (1,329,459 'yes' responses *vs.* 1,423,209 'no' responses; $\chi^2_1 = 3192.92, p = .0000$ across the whole dataset). This completely discards the possibility that the significant correctness of the very short responses is due to a bias in favor of 'yes' responses.

The left panel of Figure 21 zooms into the very early visual lexical decision responses of Figure 20. The solid lines plot the predicted log-densities of correct (black) and incorrect (grey) responses, and the solid lines represent the observed log densities (estimated by KDE). The first thing that becomes apparent is that, although the number of erroneous responses is notably increased with respect to the rest of the distribution, there are still significantly more correct than incorrect responses all the way through the interval. The prior expectation for words and non words was even in this experiments. Therefore, this advantage for correct responses can only be due to influence from the actual presentation of the words. The synaptic and conduction delays between optical presentation of a stimulus and the performance of a manual response, have been estimated to lie between 180 ms. and 260 ms. in monkeys, and an additional increase of one third is suggested to account for these times in humans (cf., Ledberg, Bressler, Ding, Coppola, & Nakamura, 2007; Thorpe & Fabre-Thorpe, 1991). This would estimate of non-decisional task component in humans to lie between 240 ms. and 350 ms. However, as can be seen in the figure, even much earlier than this, participants are providing responses that are influenced by the stimulus. This suggests that this time is also variable. This is not very surprising, one would expect

that the neural processes involved also reflect stochastic rise to threshold mechanism for triggering the final motor response and on the perceptual side. The cases when the non-decisional task components were shorter than usual should then be characterized by the general distributions of correct and error responses, which are depicted by the solid lines in the figure. These could explain around 9,756 of the total of 12,138 responses below 250 ms. In addition, as we discussed in the section of anticipations, and additional very small percentage would correspond to the cases where the process was above the response threshold before the presentation of the stimulus. Thus, these will be fully random responses. We can estimate their number using (16), and this would predict around 576 additional random responses which correspond to true anticipations. Putting these two together, there remain around 1,806 responses that cannot be accounted for neither by the general distributions nor by the predicted anticipations. This is approximately 15% of the very short latencies, and 0.1% of all responses, and they can indeed correspond to Carpenter's express responses of sub-cortical origin. The correctness of these responses will be at random, resulting in a stronger increase in the number of erroneous responses. The right panel in Figure 21 illustrates this. When one compares the log-ratios of observed to predicted short responses, on finds that there is a much more marked increase in errors than in correct ones, and the difference between these log-ratios is constant in time.

Separating effects: Picture naming

In order to illustrate the methodology on how to disentangle top-down effects from bottom-up ones, we return to the picture naming dataset. Overlooking the multiple variables that were of interest in this experiment (see Moscoso del Prado, Gabris, and Alario (in prep.) for more details) we will concentrate on two variables that we can predict will have different top-down and bottom-up components. On the one hand, the complexity of an image is likely to have a bottom-up, perceptual component, independently of prior expectation factors. An objective measure of the image's complexity can be obtained by the mere size of the JPEG file that stores it (Székely & Bates, 2000). As the JPEG format uses compression, the size of the resulting file can be regarded as an estimate of the complexity of the image in the Kolmogorov sense (Li & Vitányi, 1993).[14] Therefore, following Reddi *et al.*, (2003) we can predict that this property should influence the intercept of the Reciprobit plot, as do bottom-up effects.

On the other hand, the lexical frequency of the the word chosen to name it is not a property of the stimulus or of how it is perceived, but rather a measure of the prior probability of the response. Although there is probably some correlation between the frequency of an object and the frequency of the word(s) that can be used to name it, the fact that participants used different words to name the same object will enable us to separate effects. As shown by Carpenter and Williams (1995), this should be reflected in a variation of the slope of the Reciprobit plot.

[14]Strictly speaking, a correct approximation of Kolmogorov complexity would be one using lossless compression, instead of the lossy compression used by JPEG. However, if we assume that the relevant information loss is minimal – as it must be since the images can be identified without notice of the information loss – the JPEG file size can still approximate Kolmogorov complexity in the restricted sense that is relevant for our purpose.

Materials and methods.

This is a relatively simple dataset in which 20 French participants had to name 520 line drawings of objects that were presented in a computer screen. The participants did not receive any instruction of what were the "correct" names for the pictures. Instead, for each picture, they were free to choose the name that they considered most adequate. The line drawings were presented at the center of the monitor, after a fixation cross, and remained in the screen until the participant responded up to a maximum of three seconds, and a further period of one second was recorded to allow for very late responses. Responses were measured using a voice key, and an additional correction of the voice-key estimates was performed manually to ensure that the times correspond to the onset of the word as accurately as possible. Trials that elicited multi-word utterances, and trials containing coughs or false starts were removed from the dataset. Furthermore, in order to alleviate the computational cost of the analyses below, in this study we consider only a random sample of 290 of those pictures.

Analysis.

The RTs in this dataset fall well into the Recinormal zone with an estimated CoV of the numerator of $\lambda_1 = .02$, and an acceptably straight Reciprobit plot save for the very slow responses (see Figure 11). Inspection of the subject- and item-specific Reciprobit plots did not reveal any significant departures from normality (see Figure 22 and Figure 23). Therefore, it is safe to analyze these data using the Recinormal model (*i.e.*, to analyze the reciprocal of the RTs under normality assumptions).

To investigate the effects of picture complexity and word frequency on picture naming latencies, we fitted a GAMLSS regression model using the normal family for the response distribution, and a log link function for the scale parameter σ. As dependent variable we used the reciprocal of the RTs (scaled by a factor of 1000 to a Hertz unit scale in order to avoid numerical error from very small numbers) and we considered possible main effects of word frequency (in logarithmic scale) and JPEG file size (in logarithmic scale) both on the mean and on the standard deviation of the normal distribution. In order to analyze individual trial data, it is advisable to take explicitly into account the random effects that the experimental stimuli and the participants themselves introduce in the data (cf., Baayen, *et al.*, 2008). Therefore we included additional random effects of individual subjects and pictures on the standard deviation formula of the model (so that the random effects could have an influence on both the slope and the intercept of the reciprobit plots). As the GAMLSS implementation that we use does not by itself optimize the degrees of freedom of the random effects (as would happen for instance in a linear mixed-effect model), these two parameters were optimized separately using a Nelder-Mead optimization on the AIC of preliminary fits of the model (without completing the optimization process in order to speed up the analysis). In this way we obtained an estimate of 22 for the degrees of freedom of the random effect of individual picture, and an estimate of 19 for the degrees of freedom of the random effect of individual participants. Once the optimal degrees of freedom for the random effects were found, a more precise regression (re-setting the optimization control parameters to its default values so that the optimization process would be completed) was run with the chosen values.

Results.

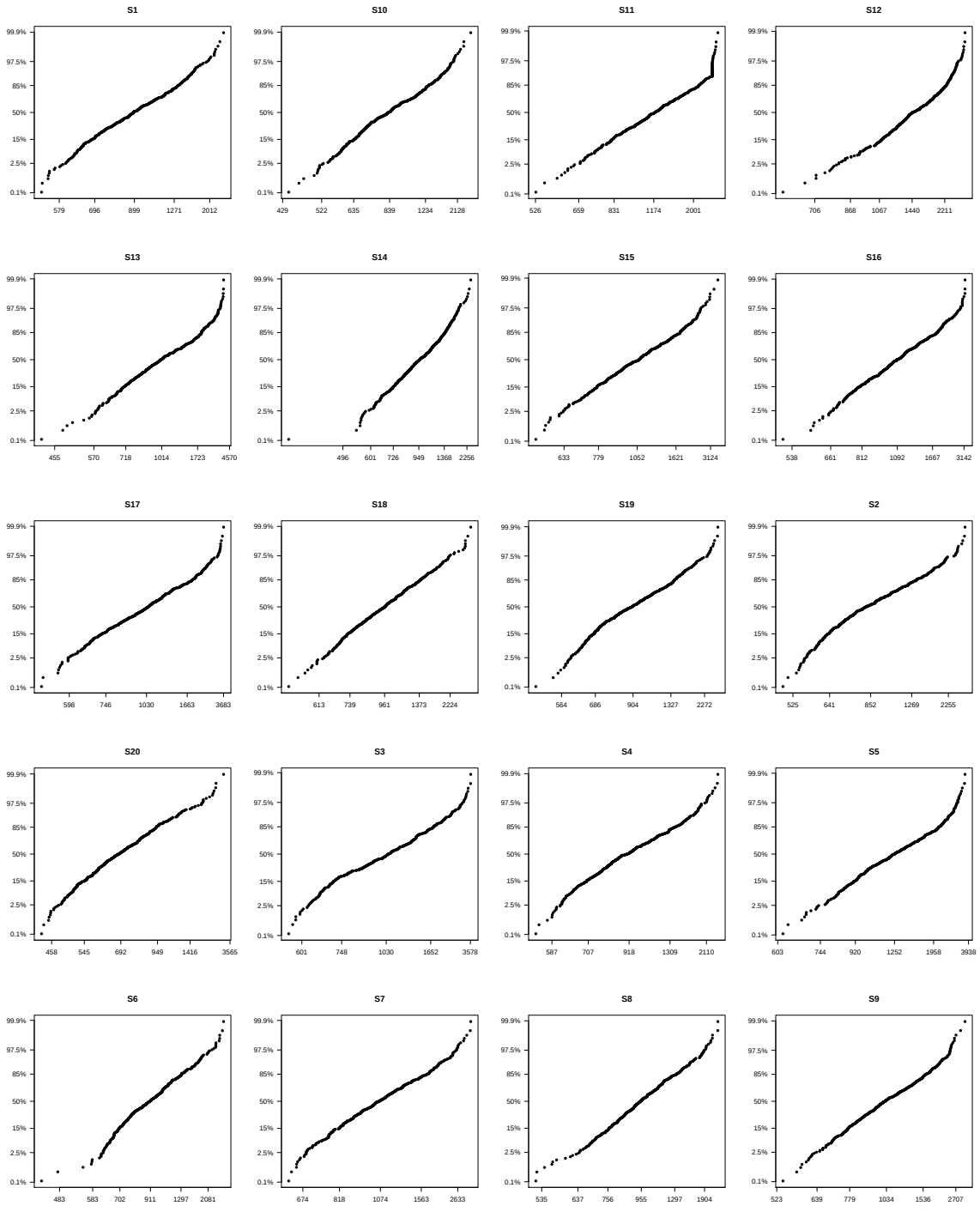

Figure 22. Individual reciprobit plots for each of the 20 participants in the picture naming experiment.

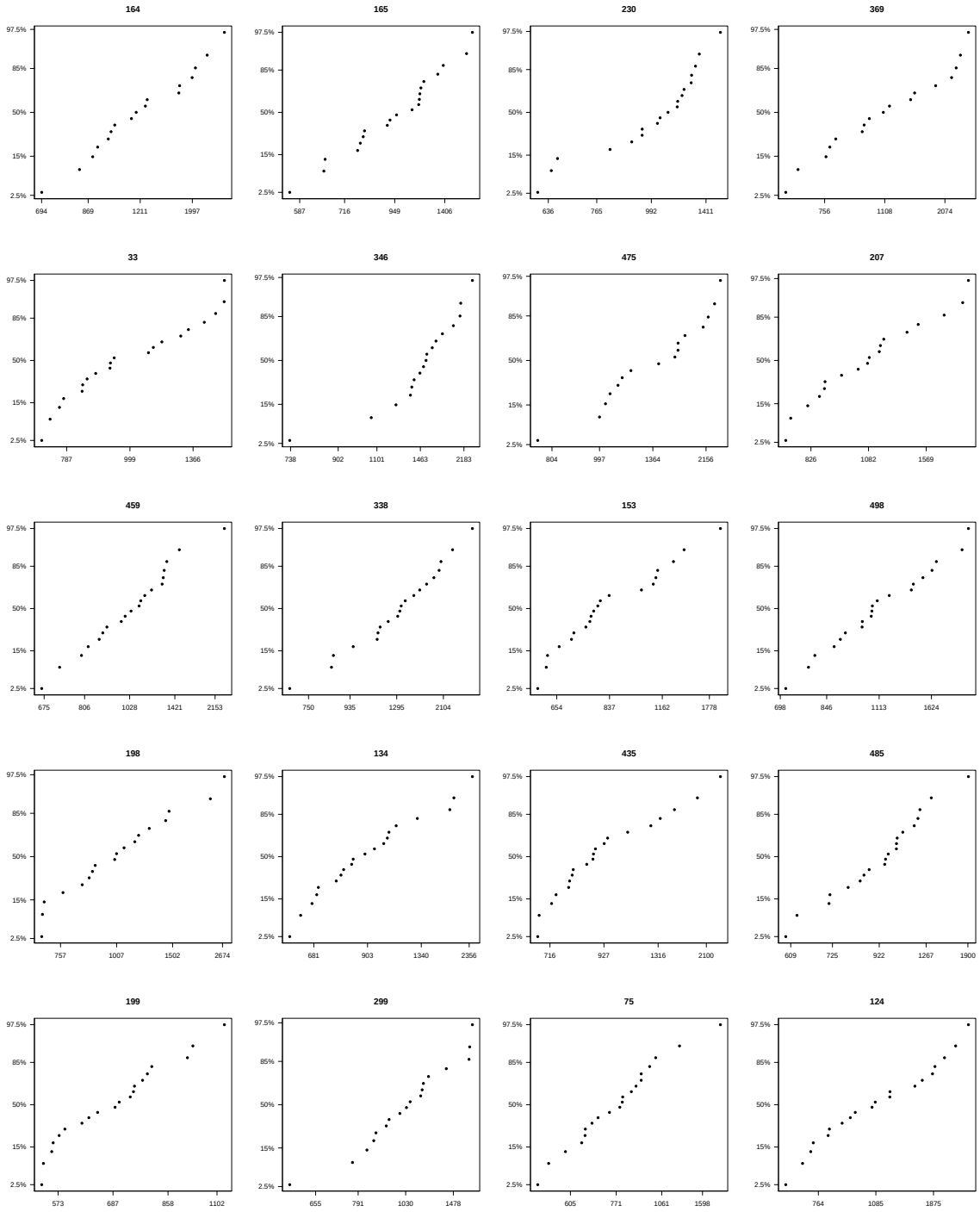

Figure 23. Individual reciprobit plots for a random sample of 20 items (*i.e.*, pictures) in the picture naming experiment.

As we had predicted, we observed that word frequency and picture complexity had different effects on the means and standard deviations. On the one hand, picture complexity only affected the mean component of the model ($\widehat{\beta} = -.07836 \pm .00872, t = -8.990, p < .0001$) while it did not have any significant contribution to the standard deviation component ($\widehat{\gamma} = -.02557 \pm .01913, t = -1.337, p = .1813$). On the other hand, word frequency significantly affected both the mean component of the model ($\widehat{\beta} = .02608 \pm .003408, t = 7.653, p < .0001$) and its standard deviation component ($\widehat{\gamma} = .01917 \pm .007396, t = 2.592, p = .0096$). The estimated intercept components of the model were $\widehat{\beta}_0 = 1.7755$ and $\widehat{\gamma}_0 = -.9105$.

From this analysis we can conclude that picture complexity affects the intercept of the reciprobit plot, but not its slope. This is to say that, everything else being equal, if we we plotted the reciprobit plots corresponding to naming latencies to two pictures of different complexity, we would observe two parallel lines in the plots.

The result for word frequency is a bit less clear. As we discussed before, if something affects the mean component of the GAMLSS regression, then we cannot be sure of whether that factor affects the slope of the reciprobit plot, or it actually influences both the slope and the intercept. The only thing we can discard is an effect only on the intercept, because if that were the case, we would expect to observe the same limitation as we saw in picture complexity. This is the case of word frequency. Therefore we conclude that word frequency affects the slope of the reciprobit plot, and possibly also its intercept.

Figure 24 confirms this interpretation. The upper panels plot the differences on the estimated values of μ_Δ (top left) and μ_r by fitting separate instances of Fieller's distribution to subsets of the data split either by the median frequency or by the median picture complexity. It is clear from the bar-plots that while word frequency has strong effects both on the estimated Δ and r means, picture complexity seems to affect only the mean of r (right bar-plot) but not the mean of Δ (left bar-plot).

We can thus interpreted these results as picture complexity having a bottom-up, perceptual, effect, while word frequency seems to affect both the bottom up and the top down components of the system. This may reflect the contribution of the measure to both the object's and word's frequency.

The GAMMLSS regression enabled us to reach this same conclusion in a single step, without the need to o median splitting (which could be problematic for non-linear effects). In addition, it enables us to explicitly consider random effects and interactions.

General Discussion

The central piece in the theory that we have proposed is the distribution of the quotient of two correlated normal variables, Fieller's distribution. The empirical evidence that we have examined seems to support this distribution as a description of RTs across tasks and modalities.

Power-law right tails

We have presented evidence in support of an RT distribution with a power-law type of right tails. RT distributions, whether individually computed for single participants in a given task and condition, or aggregated across subjects and experimental stimuli have significantly thicker tails than one would predict by any of the distributions that have traditionally been put forward to describe RTs. This includes distributions with exponential

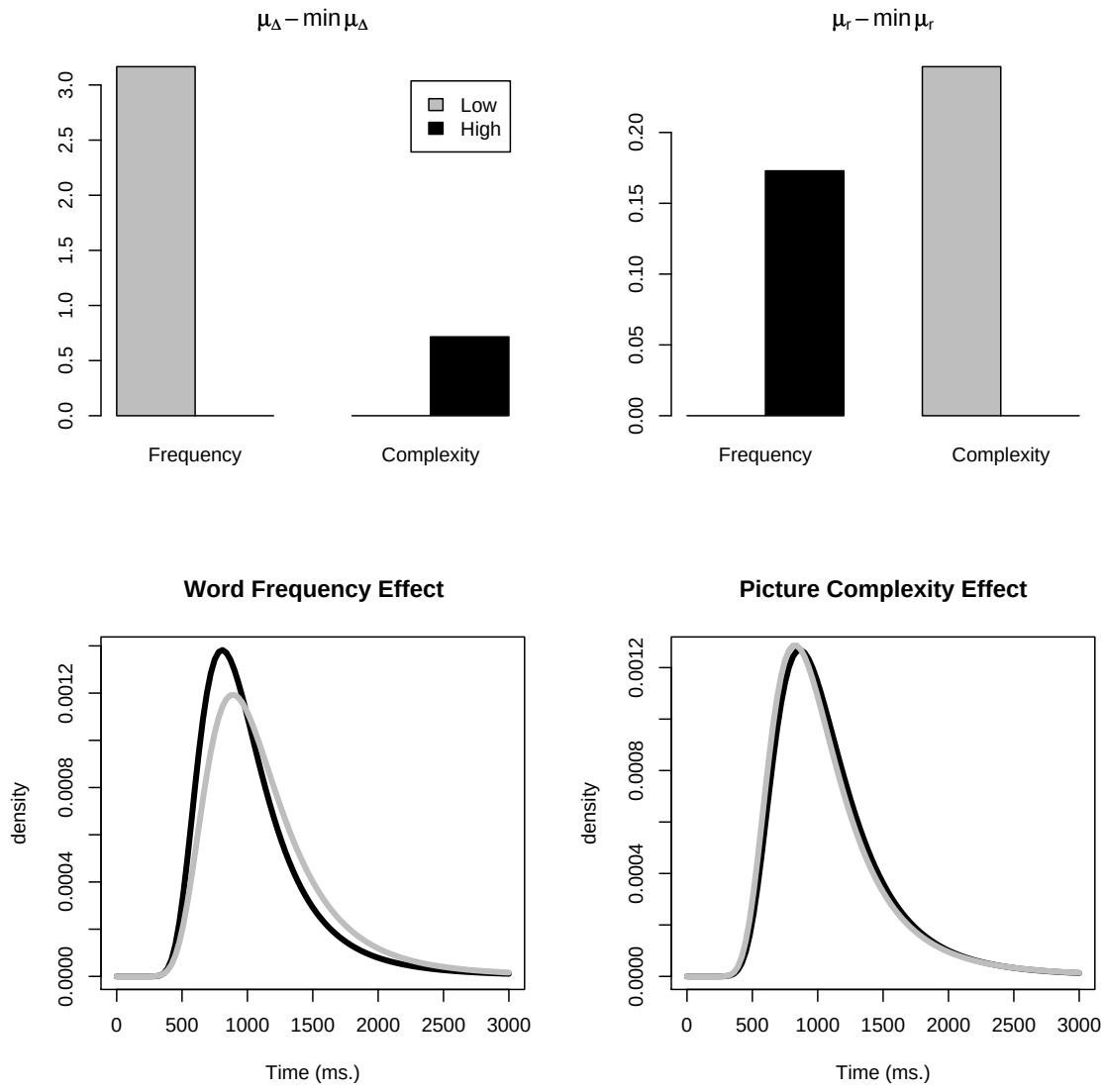

Figure 24. Effects of word frequency and picture complexity on the RT distributions of the picture naming experiment. Both measures have been split into High/Low contrasts using a median split, and separate instances of Fieller's distribution have been to each half of the data. The upper panels illustrate the estimated differences on the means of Δ (*upper left* and r (*upper right*) from the estimated Fieller's fits. The lower panes plot the different RT distributions.

tails such as the Ex-Gaussian (*e.g.*, Balota *et al*, 2008; Hohle, 1965; McGill, 1963; Ratcliff & Murdock, 1976; Ratcliff, 1978), the Ex-Wald (Schwarz, 2001), the Inverse Gaussian (*e.g.*, Lamming, 1968; Stone, 1960), the Gamma (*e.g.*, Christie, 1952; Luce, 1960; McGill, 1963), and the distributions that describe the first passage times through a threshold of (linear versions of) the DDM (Ratcliff, 1978 and subsequent studies) whether in exact forms (*e.g.*, Luce, 1986; Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002; Smith, 2000; Tuerlinkcx, 2004) or in approximate forms (*e.g.*, Lee *et al.*, 2007; Navarro & Fuss, 2008), or by distributions with slightly heavier stretched exponential types such as the Weibull (Colonius, 1995; Logan, 1988; 1992; 1995) or the distributions that arises from the 'ballistic' models recently proposed by Brown and Heathcote (2005, 2008). None of these distributions can show the type of power-law – straight in log-log scale – right tails that we have observed in the data. Admittedly, the evidence for a power law should be taken with care, since a full 'demonstration' of power-law behavior would require to have data spanning at least one more order of magnitude beyond what we have available. However, the qualitative evidence from the visual inspection of the tails, and the quantitative evidence such as the goodness of fit statistics (AIC and BIC), both point in this direction when sufficiently large datasets are examined. Of the distributions that have been proposed to account for RTs, only the Recinormal distribution proposed in LATER (Carpenter, 1981, and subsequent studies) and the generalization by Fieller's normal ratio distribution that we have introduced, are capable of producing this type of tails.

Flexible hazard functions

The shapes of the hazard functions of RT distributions (Burbeck and Luce, 1982; Luce, 1986) provide further evidence in support of Fieller's. We have shown that, in the tasks that we have examined, hazard functions are of the peaked type. In addition, after the peak, the functions seemed to take the monotonically decreasing shape of what is commonly termed an 'infant mortality' type of process. In contrast, Burbeck and Luce showed that responses to low intensity auditory stimuli can give rise to monotonically increasing hazard functions, perhaps with a final plateau. Taken together, these pieces of evidence discard most existing distributions as candidates for a general model of RT distribution. On the one hand, distributions like the Ex-Gaussian or the Weibull can only give rise to monotonic patterns (restricted to increasing in the case of the Ex-Gaussian), and are thus uncapable of accounting for any of the datasets we have analyzed. On the other hand, most other RT distributions such as the Inverse Gaussian, Ex-Wald, and Log-normal are restricted to peaked hazard functions. This makes them unsuitable to account for Burbeck and Luce's low signal intensity data. The distribution that is central to the theory that we are proposing, Fieller's, is characterized by a relatively flexible shape of the hazard function. Strictly speaking our distribution is of a peaked hazard type, followed by a linear decreasing phase (corresponding to its power-law right tail). However, as the value of the λ_2 parameter approaches zero, the distribution converges on a normal distribution, which is characterized by a monotonically increasing hazard rate. This is to say, as λ_2 goes to zero, the location of the peak goes to infinity. This enables the distribution to account for both monotonically increasing hazards and for peaked ones.

Larger number of fast responses

Most current models of RT distributions take the fast guesses to originate in a separate distribution from that of the general RTs. These responses enter in a 'race' with the stimuli-elicited responses. Yellott (1971) proposes a "deadline model" in which the fast guesses are fully independent of the actual responses (*i.e.*, a 'pure race'). In contrast, the models proposed by Ollman (1966) and Tiefenau, Neubauer, von Specht, and Heil (2006), the fast guesses and the responses elicited by the stimuli are mutually exclusive. The original model developed by Carpenter and collaborators (Anderson & Carperter, 2008; Carpenter, 2001; Carpenter & Williams, 1995; Reddi & Carpenter, 2000) is of this later type: the sub-cortical 'express responses' exclude (up to a certain limit) the cortical ones. Their motivation to posit the necessity of these responses comes from the observation of an additional line in the Reciprobit plots in some experiments, indicating that the number of fast responses is far above what a Recinormal distribution would predict. Similarly, Ratcliff (2001) includes a separate subpopulation of uniformly sampled random responses, to account for the secondary line in Reddi and Carpenter's Reciprobit plots.

Nakahara *et al.* (2006) noticed that adding normal variation in the threshold level of LATER could give rise to a slight deviation in the lower part of the reciprobit plot. Our studies of Fieller's distribution have confirmed the intuition of Nakahara and collaborators. Normal variation in either the starting level or the threshold level can give rise to exactly the type of deviations from recinormality that Carpenter and colleagues attribute to sub-cortical responses. In Carpenter's experiments, the faster conditions elicited more of the express responses. Notice that this seems rather counterintuitive. If these fast guesses were in a race with the actual cortical responses, one would expect that the longer the delay of the cortical response, the higher the chances of the sub-cortical having time to reach the threshold, opposite to what Carpenter and collaborators observed. Our theory provides the tools to predict when and how this population will arise. In Figure 8 we illustrated that the effect of the variation in Δ on the overall distribution is attenuated for longer RTs. For these, the accumulated variation converges to the same that would be produced by variation in r alone. Therefore, as Carpenter and his colleagues repeatedly observed, conditions that on average elicit longer responses, will tend to show less of this deviation from recinormality. Another property of the express responses is that variability in the order of stimuli increases their proportion (Carpenter, 2001). This type of variability would be reflected in variation in the predictability of stimuli. As we have argued, this type of variation is reflected in the value of the λ_1 parameter, the greater the variation the greater λ_1 and the larger the deviation from recinormality. Generally, the values of the CoV parameters of Fieller's distribution (λ_1 and λ_2) provide a compact way of predicting the detailed shape of the distribution. For instance, the deviation from recinormality is fully accounted for by the value of the λ_1 parameter.

An important issue is that most of these fast responses are fully accounted for by the general RT distribution. The implication of this is that they are not more random than the rest of the responses. As illustrated in Figure 21, we have seen that performance is above chance up to the very early times below 100 ms. In addition to these fast responses that are part of the overall distribution, we have seen that there are two additional types of fast responses. First, there are the very few that are predicted by the distribution to

correspond to cases where the response threshold has been accidentally surpassed before any information about the stimulus arrived. These are indeed at random and correspond to a tiny fraction of responses. Finally, after discounting for the two types above, there is still one third type of fast responses, amounting to approximately one per thousand, that are fully at random. These can indeed correspond to the sub-cortical 'express responses' of LATER, but are in all cases a truly small proportion.

Need for inhibition

Bogacz *et al.* (2006) demonstrated that different versions of linear accumulator models can all, under certain conditions, be reduced to the classical linear DDM, as long as some inhibition mechanism is present in the system. Importantly, they find that 'pure race' models without any significant contribution of inhibition produce different predictions from those of the DDM. As noticed by Colonius (1995) and Logan (1992, 1995), this type of race models necessarily lead to Weibull type RT distributions. As discussed above, Weibull-type show too thin right tails. The presence of inhibition attenuates the general speed-up caused by the competing accumulators, resulting in a higher number of long responses than would be predicted by models such as that of Logan (1988) or the ones recently proposed by Brown and Heathcote (2005; 2008).

Hick's Law.

In their recent study, Brown and Heathcote (2008) noticed that a setback of their LBA model is that it cannot account for Hick's Law (Hick, 1952): The fact that the time to choose among a number of candidates is directly proportional to the log number of possible alternatives. In their view, under the pure – inhibition free – race model, increasing the number of accumulators would lead to faster responses (as the probability of one of them crossing the threshold at any point would increase), and larger error rates (as there are more accumulators that can possibly win the race). To solve this problem, they refer to some parameter adjustments that could eventually address this problem.

As discussed above, the prior probability of stimuli affects either resting levels or response thresholds in a logarithmic way. For instance, Carpenter and Williams (1995) showed how the logarithm of the log prior probability of a stimulus has a very clear correlation with the distribution of the time it takes for the stimulus to be recognized. Similarly, Oswal *et al.* (2007) found that modifying the instantaneous expectation of the stimuli gave rise to the same changes in the shape of the distribution. In both cases, changes in prior probability are reflected in changes in slope of the reciprobit plot, indicating that it is affecting either the decision threshold or the starting level. This was also confirmed by Reddi and Carpenter (2000). Similar results consistently show up in 'ideal observer' models of decision and recognition tasks (*e.g.*, Moscoso del Prado, 2008; Norris, 2006; 2009). In fact, the correspondence between logarithmic prior probability and resting level (or threshold) is the only one that enables a probabilistic formulation of the model at a computational level. It is optimal (*cf.*, Jaynes, 2003).

By definition, as already noticed by Hick (1952) and Stone (1960), the prior probability is reciprocally related to the number of alternatives. For instance, if we assume all alternatives to be equiprobable, then the probability of each of them is $1/N$, with N being the number of choices. On average, this inverse proportionality extends into all non-

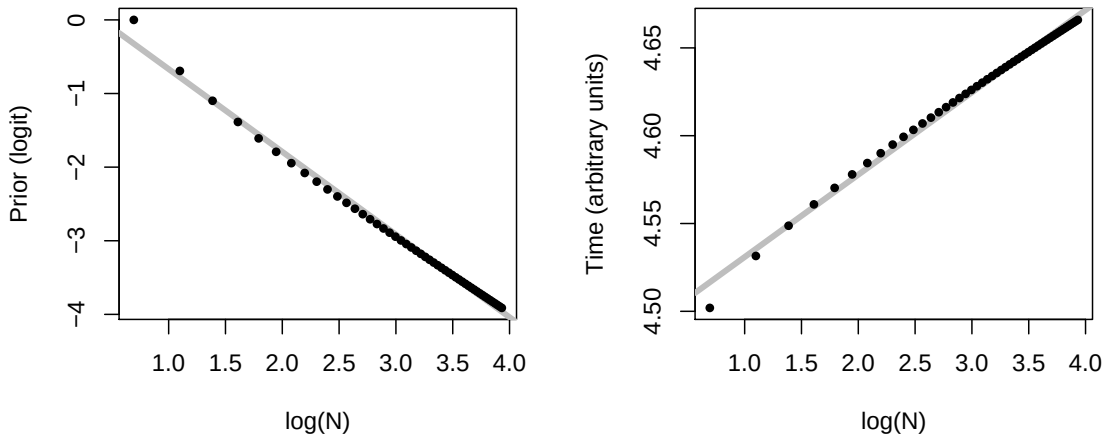

Figure 25. Hick's Law. *Left panel.* Log number of options versus the corresponding logit prior probability in a uniform situation. *Right panel.* Expected effect of number of choices on correct RTs if resting levels of accumulators are proportional to the logit prior probability.

degenerate distributions of probability that one could choose. The log prior probability of the correct response will naturally decrease logarithmically with the number of choices. If this prior probability affects the Δ factor (either the resting level or the threshold itself), this will make the latency of the correct response proportional to the minus log of the number of choices. The expected consequences are shown on Figure 25. There is an (approximately) logarithmic relationship between the number of alternatives and the average RT

Notice that the probabilistic interpretation of the model, with starting levels interpreted as prior odds and the incoming evidence interpreted as a Bayes factor is in itself dependent on the presence of inhibitory mechanisms: if the probability of one option grows, on average the probability of the others need to decrease.

Levels of explanation and 'optimality'

As argued by Marr (1982), there are three levels at which models of cognitive function can explain cognitive phenomena. A computational level presents a formal description of the problem, an algorithmic level, in which a description of the method used to solve it is described, and an implementational level which describes how such computations can be performed in term of neural structures. An important point that was also made by Marr is often overlooked. There needs to be an explicit link between the explanations offered at the different levels.

In this sense, the model that we have introduced constitutes a description of the origin of RTs at an algorithmic level. In addition, we have also explicitly linked the model to computational and implementational descriptions. On the one hand, as has been noticed by proposers of LATER and of the DDM, our theory fits into a general Bayesian inference

framework. This is hardly surprising, in the words of Jaynes (2003) when discussing an example from visual perception:

> We would expect Darwinian natural selection to produce such a result; after all, any reasoning format whose results conflict with Bayesian inference will place a creature at a decided survival disadvantage. Indeed, as we noted long ago [...], in view of Cox's theorems, to deny that we reason in a Bayesian way is to assert that we reason in a deliberately inconsistent way; we find this very hard to believe. Presumably, a dozen other examples of human and animal perception would be found to obey a Bayesian reasoning format as its "high level" organizing principle, for the same reason. [...] (*p. 133*)

A consequence is that different models, making slightly different predictions, claim to describe the behavior of the optimal decision maker, the ideal observer. This could seem like a contradiction. In our opinion, this is not a very informative question. The issue is not whether the decision process is optimal. As we have argued above, it *must* at least approach optimality. The crucial point is to find what is it that is being optimized and under which conditions. For instance, despite their different formulations and predictions, both LATER and the classical DDM are optimal. In both cases, the assumption is that the optimized function – the cost function – is a function of time. In the case of the DDM, the quality of a response is directly proportional to the time it took. In the case of LATER the cost function is non-linear with respect to time. In addition, both models make different assumptions on how the evidence becomes available, either at a constant rate or at a randomly changing one.

On the other hand, we have seen that the model can be reduced to a non-linear instance of the DDM family of models. In our formulation, we have introduced the non-linearity by making the volatility rate or diffusion coefficient proportional to time. Working from the opposite direction, that is, building up from the known properties of neurons, Roxin and Ledberg (2008) have reached similar conclusions. They show that the behavior of realistic neural network models can be reduced to a one dimensional non-linear diffusion equation. In particular, they arrive at a diffusion equation in which the drift rate has a cubic dependence on the value of the accumulator at any point in time. It remains to be seen whether the distribution we have proposed can be generated by such type of equation, but a general need for non-linearity is apparent from both our theory and Roxin and Ledberg's neural network models. Bogacz, Usher, Zhang, and McClelland (2007) have also suggested that extending the Leaky Competing Accumulator model (LCA; Usher & McClelland, 2001) to include the nonlinearities that are observed in neural populations might lead to a better account of experimental data by the LCA model.

The inclusion of LATER into the DDM family also enables our model to inherit some of the known properties of the DDM. Importantly, the DDM has proven of great value to account for a large set of experimental phenomena on which LATER has not been explicitly tested. Most salient among these phenomena are speed-accuracy trade-offs. Our model being a particular instance of the DDM enables us to take advantage of the DDM ability to explain such phenomena.

Recognition vs. *decision*

In their response to Ratcliff (2001), Carpenter and Reddi (2001) argue that LATER is a model that applies to different processes than the DDM. Whereas the former would describe processes dominated by a recognition component, the later would describe the RTs in processes that are dominated by decisional components. In our opinion, this is not a satisfactory difference. For instance, as argued by Ratcliff (2001), the DDM has in fact been most applied to decisional processes such as the lexical decision task (Ratcliff, Gomez, & McKoon, 2004), or same/different two choice decisions (Ratcliff, 1985; Ratcliff & Smith, 2004). Furthermore, the difference between "recognition" and "decision" seems to us a rather vague one. One can think of any recognition process as a plain decision, in which evidence is accumulated until a threshold is reached. In that sense, we have seen that, as Ratcliff (2001) suggested, LATER can be regarded as a non-linear version of the DDM. We think that Carpenter might have underestimated the power of LATER to account for all types of processes.

In our view, a more consistent difference would refer to the number of accumulation processes in a sequence. As we discussed, having a sequence of LATER-style accumulators will result also in an RT distribution which is also an instance of Fieller's. However, the convolution of instances of Fieller will have the effect of increasing the value of the λ_1 parameter, which controls the deviation from recinormality. Therefore, the more accumulated stages that a process has, the less Recinormal the distribution will tend to be. This will result in the more complex processes deviating more and more from the original LATER model (in which $\lambda_1 = 0$), and may account for some of the intuitions of Carpenter and Reddi (2001).

Implications of the power-law

The power-law signature of the right tail of Fieller's distribution does not come without implications. Power-law distributions occur in a very diverse range of natural phenomena. The origin of this type of distributions has attracted a fair amount of interest from physicists. Generally speaking, power-laws are the typical footprint of systems in a state of "self-organizing criticality" (SOC; *cf.*, Bak & Paczuski, 1995). These are complex systems in which the behaviour of any part is dependent on the whole, so that perturbations (*e.g.*, presentation of stimuli) affect the whole system. It is not surprising that the brain may be one of such systems. Indeed, recent work in neurophysiology has shown that brain oscillations also show patterns that are indicative of a complex SOC system (*cf.*, Buzsáki & Draguhn, 2004). Furthermore, Gilden, Thornton, and Mallon (1995) observed that the noise in RTs also exhibits $1/f$ "pink noise" characteristics, which are another typical mark of SOC systems. We have not explored further ther SOC implications of the power-law, but this may provide a useful way of linking properties of RT distributions with the neurophysiology of the brain. In addition, further predictions on the properties of RT data could hypothetically be derived from the properties of complex systems.

Large datasets, long responses, and data trimming

The power-law properties of the right tails also stress the importance of the *size* of datasets that are used to compare theories. The most conclusive evidence that is contrastive

among theories comes from these tails. The greater part of the advocated RT distributions are sufficiently flexible as to be able to replicate the patterns shown around the distributional mode, giving rise to the model mimicry problem discussed by Ratcliff and Smith (2004), Van Zandt and Ratcliff (1995) and Wagenmakers *et al.* (2004). As we have seen, comparing models using relatively small datasets – up to somewhere over 1,000 responses per participant – gives an unrealistic bias in favor of exponential-tailed distributions. Very late responses happen very rarely, and without those, exponential tails appear to give the best fits to the data. As soon as a sufficient number of these responses has appeared, the picture changes drastically. Power-law type distributions begin to offer by far the best fits. Proportionally, the differece in favor of the power law found in large datasets is substantially larger than the equivalent advantage of exponential tails in the smaller datasets, thus the positive average values for both information criteria in Table 7.

This also speaks to the damage resulting from truncating long and short responses as 'outliers'. This has been both the recommended technique (*e.g.*, Luce, 1986; Ratcliff, 1993; Van Zandt, 2002; Whelan, 2008) and the 'standard practice' in the field. As we have argued, trimming the long responses results in the loss of crucial information and should therefore be avoided in as much as possible (a certain amount of trimming will remain from the fact that the measurement of RTs stops after some deadline in most experiments). This problem is in fact not exclusive to the analysis of RT data. As discussed by Bak and Paczuski (1995), Newman (2005), and Mandelbrot (1983) these 'contingent' events are also erroneously discarded by attribution to 'special' causes in areas such as market fluctuations or earthquakes. However, they are but consequences of the power-law that governs these phenomena. As our analyses show, very long RTs are not events from some other distributions, but plain events in the general one. This is to say, long RTs are just long, not 'weird' at all but rather their frequency (but not their actual occurrence) is well predictable. These very large rare events are the hallmark of self-organizing – emergent – systems, that are governed by power-laws.

Analyzing RT distributions rather than means

In their recent study, Balota *et al.* (2008) argued in favor of considering the full shape of the RT distribution rather than just mean RTs when analyzing data. In their study, it becomes clear that different factors might affect the distributions in ways different than just changing the mean RT. Similar points on the importance of taking RT distributions fully into consideration have been made by several researchers in the field (*e.g.*, Heathcote, Popiel, & Mewhort, 1991; Luce, 1986; Ratcliff, 1979; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Van Zandt, 2002). This importance of the distribution is also implicit in the reciprobit analysis techniques adopted by Carpenter and colleagues (Carpenter & Williams, 1995; Oswal *et al.*, 2007; Reddi *et al.*, 2003; Reddi & Carpenter, 2000). We fully subscribe this view. As we have seen, the ways in which the RT distributions change are indeed informative of the processes that may be generating them. Furthermore, we have seen the importance of jointly analyzing data coming from different participants and items. However, as argued by Baayen *et al.* (2008), doing this requires explictly taking this variation into account by means of including random effects in regression analyses. In our observations, the need for random effects is made more accute, as evidenced in the word naming dataset, RT distributions may depart from recinormality, but reducing inter-subject and inter-stimuli

variation might put them back into the recinormal area, where the reciprobit analysis may be safely performed. We have offered a technique that formally integrates the random effect with the reciprobit analysis techniques of Carpenter and hi colleagues. This tool enables us to have an objective way of evaluating – further than by visual inspection of the plots, which may become very difficult for continuous variables and non-linear effects – the contributions of a factor towards the slopes and intercepts of the reciprobit plots.

Conclusion

We return to the statement advanced in the introduction. RTs are directly proportional to the difficulty of the task, and inversely proportional to the rate at which information becomes available to solve it. Both task difficulty and rate of information income are normally distributed.

References

Anderson, A. J., & Carpenter, R. H. S. (2008). The effect of stimuli that isolate S-cones on early saccades and the gap effect. *Proceedings.of the Royal Society B*, *275*, 335-44.

Aroian, L. A. (1947). The probability function of a product of two normally distributed variables. *The Annals of Mathematical Statistics*, *18*, 256-271.

Aroian, L. A., Taneja, V. S., & Cornwell, L. W. (1978). Mathematical forms of distribution of the product of two normally distributed variables. *Communications in Statistics: Theory and Methods*, *A7*, 156-172.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*. (In press)

Baayen, R., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290-313.

Bak, P., & Paczuski, M. (1995). Complexity, contingency, and criticality. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 6689-96.

Balling, L., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*. (In press)

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445-59.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*. (In press)

Basso, M. A., & Wurtz, R. H. (1997). Modulation of neuronal activity by target uncertainty. *Nature*, *389*, 66-9.

Basso, M. A., & Wurtz, R. H. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *Journal of Neuroscience*, *18*, 7519-7534.

Bernard, J. B., Moscoso del Prado Martín, F., Montagnini, A., & Castet, E. (2008). A model of optimal oculo-motor strategies in reading for normal and impaired visual fields. In L. U. Perrinet & E. Daucé (Eds.), *Proceedings of the 2n French Conference on Computational Neuroscience* (p. 97-102). Marseilles: Societé Française des Neurosciences.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*, 700-765.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 1655–1670.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211-252.

Brillouin, L. (1956). *Science and Information Theory*. New York, NY: Academic Press.

Brody, J. P., Williams, B. A., Wold, B. J., & Quake, S. R. (2002). Significance and statistical errors in the analysis of dna microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 12975–12978.

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117-128.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153-178.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*, 117-33.

Burle, B., Vidal, F., Tandonnet, C., & Hasbroucq, T. (2004). Physiological evidence for response inhibition in choice reaction time tasks. *Brain and Cognition*, *56*, 153-164.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*, 1926-1929.

Carpenter, R. H. S. (1981). Oculomotor procrastination. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye Movements: Cognition and Visual Perception* (p. 237-246). Hillsdale, NJ: Lawrence Erlbaum.

Carpenter, R. H. S. (1988). *Movements of the Eyes* (2nd ed.). London: Pion Ltd.

Carpenter, R. H. S. (2000). The neural control of looking. *Current Biology*, *10*, R291-R293.

Carpenter, R. H. S. (2001). Express saccades: is bimodality a result of the order of stimulus presentation? *Vision Research*, *41*, 1145-1151.

Carpenter, R. H. S., & Reddi, B. A. J. (2001). Reply to 'Putting noise into neurophysiological models of simple decision making'. *Nature Neuroscience*, *4*, 337.

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*, 59-62.

Christie, L. S. (1952). The measurement of discriminative behavior. *Psychological Review*, *59*, 89-112.

Clauset, A., & Erwin, D. H. (2008). The evolution and distribution of species body size. *Science*, *321*, 399-401.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2007). Power-law distributions in empirical data. *arXiv:0706.1062v1*. (`http://arxiv.org/abs/0706.1062`)

Colonius, H. (1995). The instance theory of automaticity: Why the Weibull? *Psychological Review*, *102*, 744-750.

Comon, P. (1994). Independent Component Analysis, a new concept? *Signal Processing*, *36*(3), 287-314.

Cooper, S. (1982). The continuum model: statistical implications. *Journal of Theoretical Biology*, *94*, 783-800.

Cousineau, D., Brown, S., & Heathcote, A. (2004). Fitting distributions using maximum likelihood: Methods and packages. *Behavior Research Methods, Instruments, & Computers*, *36*, 742-756.

Craig, C. C. (1928). The frequency of function of y/x. *The Annals of Mathematics*, *30*, 471-486.

Craig, C. C. (1936). On the frequency function of xy. *The Annals of Mathematical Statistics*, *7*, 1-15.

Donders, F. C. (1869). On the speed of mental processes. *Attention and Performance*, *2*, 412-431.

Feller, W. (1968). *An introduction to probability theory and its applications (vol. 1, 3rd ed.)*. New York, NY: Wiley.

Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, *24*, 428-440.

Freeman, J., & Modarres, R. (2006). Inverse Box-Cox: The power-normal distribution. *Statistics & Probability Letters*, *76*, 764-772.

Geary, R. C. (1930). The frequency distribution of the quotient of two normal variates. *Journal of the Royal Statistical Society*, *93*, 442-446.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). $1/f$ noise in human cognition. *Science*, *267*, 1837-9.

Gondan, M., & Heckel, A. (2008). Testing the race inequality: A simple correction procedure for fast guesses. *Journal of Mathematical Psychology*, *52*, 322-325.

Green, D. M., & Luce, R. D. (1971). Detection of auditory signals presented at random times: III. *Perception & Psychophysics*, *9*, 257-268.

Grice, G. R. (1968). Stimulus intensity and response evocation. *Psychological Review*, *75*, 359-373.

Grice, G. R. (1972). Application of a variable criterion model to auditory reaction time as a function of the type of catch trial. *Perception & Psychophysics*, *12*, 103-107.

Hanes, D. P., & Carpenter, R. H. S. (1999). Countermanding saccades in humans. *Vision Research*, *39*, 2777-2791.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*, *274*, 427-430.

Hays, W. L. (1973). *Statistics for the Social Sciences (2nd ed.)*. New York, NY: Holt, Rinehart & Winston.

Hayya, J., Armstrong, D., & Gressis, N. (1975). A note on the ratio of two normally distributed variables. *Management Science*, *21*, 1338-1341.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, *9*, 394-401.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin*, *109*, 340–347.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11-26.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, *56*, 635-639.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*, 382-6.

Iacus, S. M. (2008). *Simulation and Inference for Stochastic Differential Equations with R Examples*. NY: Springer.

Jan, N., Moseley, L., Ray, T., & Stauffer, T. (1999). Is the fossil record indicative of a critical system? *Advances in Complex Systems*, *2*, 137-141.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kennedy, C. A., & Lennox, W. C. (2001). Moment operations on random variables, with applications for probabilistic analysis. *Probabilistic Engineering Mechanics*, *16*, 253-259.

Kubitschek, H. E. (1966). Normal distribution of cell generation rates. *Nature*, *209*, 1039-1040.

Kubitschek, H. E. (1971). The distribution of cell generation times. *Cell and Tissue Kinetics*, *4*, 113-22.

Lamming, D. R. J. (1968). *Information Theory of Choice Reaction Times*. London, UK: Academic Press.

Leach, J. C. D., & Carpenter, R. H. S. (2001). Saccadic choice with asynchronous targets: evidence for independent randomisation. *Vision Research*, *41*, 3437-3445.

Ledberg, A., Bressler, S. L., Ding, M., Coppola, R., & Nakamura, R. (2007). Large-scale visuomotor integration in the cerebral cortex. *Cerebral Cortex*, *17*, 44-62.

Lee, M. D., Fuss, I. G., & Navarro, D. J. (2007). A bayesian approach to diffusion models of decision-making and response time. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 809–816). Cambridge, MA: MIT Press.

Li, M., & Vitányi, P. (1993). *An Introduction to Kolmogorov Complexity and its Applications*. New York, NY: Springer.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492-527.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883-914.

Logan, G. D. (1995). The Weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review*, *102*, 751-756.

Luce, R. D. (1960). Response latencies and probabilities. In K. A. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical Methods in the Social Sciences, 1959*. Stanford, CA: Stanford University Press.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Luce, R. D., & Green, D. M. (1972). A neural timing theory for response times and the psychophysics of intensity. *Psychological Review*, *79*, 14-57.

Mandelbrot, B. B. (1983). *The fractal geometry of nature*. New York, NY: Freeman.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association*, *60*, 193-204.

McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Busch, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (p. 309-360). New York, NY: John Wiley & Sons.

Miller, D. R., & Singpurwalla, N. D. (1977). *Failure rate estimation using random smoothing* (Tech. Rep. No. AD-A040999/SST). National Technical Information Service.

Moscoso del Prado Martín, F. (2008). A fully analytical model of the lexical decision task. In B. C. Love & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (p. 1035-1040). Austin, TX: Cognitive Science Society.

Nakahara, H., Nakamura, K., & Hikosaka, O. (2006). Extended later model can account for trial-by-trial variability of both pre- and post-processes. *Neural Networks*, *19*, 1027-1046.

Navarro, D. J., & Fuss, I. G. (2008). *Fast and accurate calculations for first-passage times in Wiener diffusion models.* Manuscript, School of Psychology, University of Adelaide. (`http://www.psychology.adelaide.edu.au/personalpages/staff/danielnavarro/papers.html`)

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*, 323-351.

Nolan, J. P. (2009). *Stable Distributions – Models for Heavy Tailed Data*. Boston: Birkhäuser. (In progress, Chapter 1 online at `http://academic2.american.edu/~jpnolan`)

Norris, D. (2006). The bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, *113*, 327-357.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction time distributions. *Psychological Review*. (In press)

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*, 155-166.

Oswal, A., Ogden, M., & Carpenter, R. (2007). The time course of stimulus expectation in a saccadic decision task. *Journal of Neurophysiology*, *97*, 2722-2730.

Pearson, K. (1910). On the constants of index distributions, etc. *Biometrika*, *7*, 531-541.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446-461.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*, 212-225.

Ratcliff, R. (1987). More on the speed and accuracy of positive and negative responses. *Psychological Review*, *94*, 277-280.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-32.

Ratcliff, R. (2001). Putting noise into neurophysiological models of simple decision making. *Nature Neuroscience*, *4*, 336.

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of Neurophysiology*, *90*, 1392-1407.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal Neurophysiology*, *97*, 1756-1774.

Ratcliff, R., & Murdock, B. B. (1976). *Retrieval processes in recognition memory.*

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333-367.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438-81.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review April 1999*, *106*, 261-300.

Reddi, B., Asrress, K. N., & Carpenter, R. (2003). Accuracy, information, and response time in a saccadic decision task. *Journal of Neurophysiology*, *90*, 3538-3546.

Reddi, B. A. J., & Carpenter, R. H. S. (2000). The influence of urgency on decision time. *Nature Neuroscience*, *3*, 827-830.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, *54(3)*, 507-554.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195-223.

Roxin, A., & Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Computational Biology*, *4*.

Scharf, B. (1978). Loudness. In E. Carterette & M. Friedman (Eds.), *Handbook of Perception. IV: Hearing* (p. 187–242). New York, NY: Academic Press.

Schwarz, W. (2001). The ex-wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, *33*, 457-469.

Silverman, B. W. (1986). *Density Estimation.* London, UK: Chapman & Hall.

Sinha, N., Brown, J., & Carpenter, R. (2006). Task switching as a two-stage decision process. *Journal of Neurophysiology*, *95*, 3146-3153.

Smith, P. L. (2000). Stochastic, dynamic models of response times and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408-463.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*, 161-168.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*, 251-260.

Székely, A., & Bates, E. (2000). Objective visual complexity as a variable in studies of picture naming. *Center for Research in Language Newsletter, UCSD*, *12*. (`http://crl.ucsd.edu/newsletter/12-2/article.html`)

Thompson, K. G., Hanes, D. P., Bichot, N. P., & Schall, J. D. (1996). Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *Journal of Neurophysiology*, *76*, 4040-4055.

Thorpe, S. J., & Fabre-Thorpe, M. (2001). Neuroscience: Seeking categories in the brain. *Science*, *291*, 260-263.

Tiefenau, A., Neubauer, H., Specht, H. von, & Heil, P. (2006). Correcting for false alarms in a simple reaction time task. *Brain Research*, *1122*, 99-115.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, *36*, 702-716.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the Leaky, Competing Accumulator model. *Psychological Review*, *108*, 550–592.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 465, 424.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted & H. Pashler (Eds.), *Stevens' Handbook of Experimental Psychology (3rd edition), Vol. 4: Methodology in Experimental Psychology* (p. 461-516). New York, NY: Wiley.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*, 208-256.

Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, *2*(1), 20-54.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S (4th edition)*. New York, NY: Springer.

Vickers, D., Nettelbeck, T., & Willson, R. J. (1972). Perceptual indices of performance: the measurement of 'inspection time' and 'noise' in the visual system. *Perception*, *1*, 263–295.

Voss, A., & Voss, J. (2007). Fast-dm: a free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*, 767-75.

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52*, 1-9.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28-50.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140-159.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., Rijn, H. van, & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*, 332-367.

Wagenmakers, E.-J., Van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*, 1229-35.

Wagenmakers, E.-J., Van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3-22.

Whelan, R. (2008). Effective analysis of reaction time data. *Psychological Record*, *114*, 475-482.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental Psychology*. New York, NY: Holt.

Yellott, J. I. (1971). Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology*, *8*, 159-199.

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

Appendix A
The Recinormal distribution

We define the Recinormal distribution as the distribution of a random variable X whose reciprocal $Y = \frac{1}{X}$ is normally distributed with mean μ and standard deviation σ. If $\phi(y|\mu, \sigma^2)$ is the density function of Y, as Y is monotonically related to X, the density function of X is:

$$f(x|\mu, \sigma) = \phi(y|\mu, \sigma) \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = \phi\left(\frac{1}{x}\middle|\,\mu, \sigma\right) \frac{1}{x^2}. \tag{33}$$

Developing the normal density function and simpliflying the above expression, we obtain the density function of the Recinormal:

$$f_r(x|\mu, \sigma) = \begin{cases} \frac{1}{x^2\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(1-\mu x)^2}{2\sigma^2 x^2}} & \text{if } x \neq 0. \\ 0 & \text{if } x = 0, \end{cases}, \tag{34}$$

where the value at zero has been added by by taking the limits of the general function value.

By integration of (34), we obtain the cumulative distribution function of the recinormal:

$$F_r(x|\mu, \sigma) = \int_{-\infty}^{x} f_r(t|\mu, \sigma)\mathrm{d}t = \Phi(0|\mu, \sigma) - \Phi\left(\frac{1}{x}|\mu, \sigma\right) + \mathrm{I}(x \geq 0), \tag{35}$$

where Φ denotes the cumulative density function of a normal distribution, and I denotes the indicator function.

As can be seen in Figure A1 salient property of this distribution is that it is bimodal: It always has exactly one positive mode and one negative mode, although one of the two halves of the distribution can often cover a very small volume. In addition, the density function is zero valued only at the origin, that is to say, all other points but zero are possible. This has the counter-intuitive consequence that, if used to describe the distribution of response latencies, it will invariably predict a – possibly very small – proportion of *negative* RTs.

For a variable defined on the positive domain only, the Recinormal distribution is a shifted version of the Box-Cox or power-normal family of distributions (Box & Cox, 1964; Freeman & Modarres, 2006) with a fixed value of the power parameter (ν) of -1. For $x > 0$, the density of this distribution is given by:

$$f(x) = \frac{x^{\nu-1}}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}}, \tag{36}$$

where z is the Box-Cox transformation of x with power parameter ν:

$$z = \begin{cases} \frac{x^\nu - 1}{\nu} & \text{if } \nu \neq 0 \\ \log(x) & \text{if } \nu = 0 \end{cases} \tag{37}$$

Substituting $\nu = -1$ in the previous expressions, we obtain:

$$f(x) = \frac{1}{x^2\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-1-x\mu)^2}{2x^2\sigma^2}}, \tag{38}$$

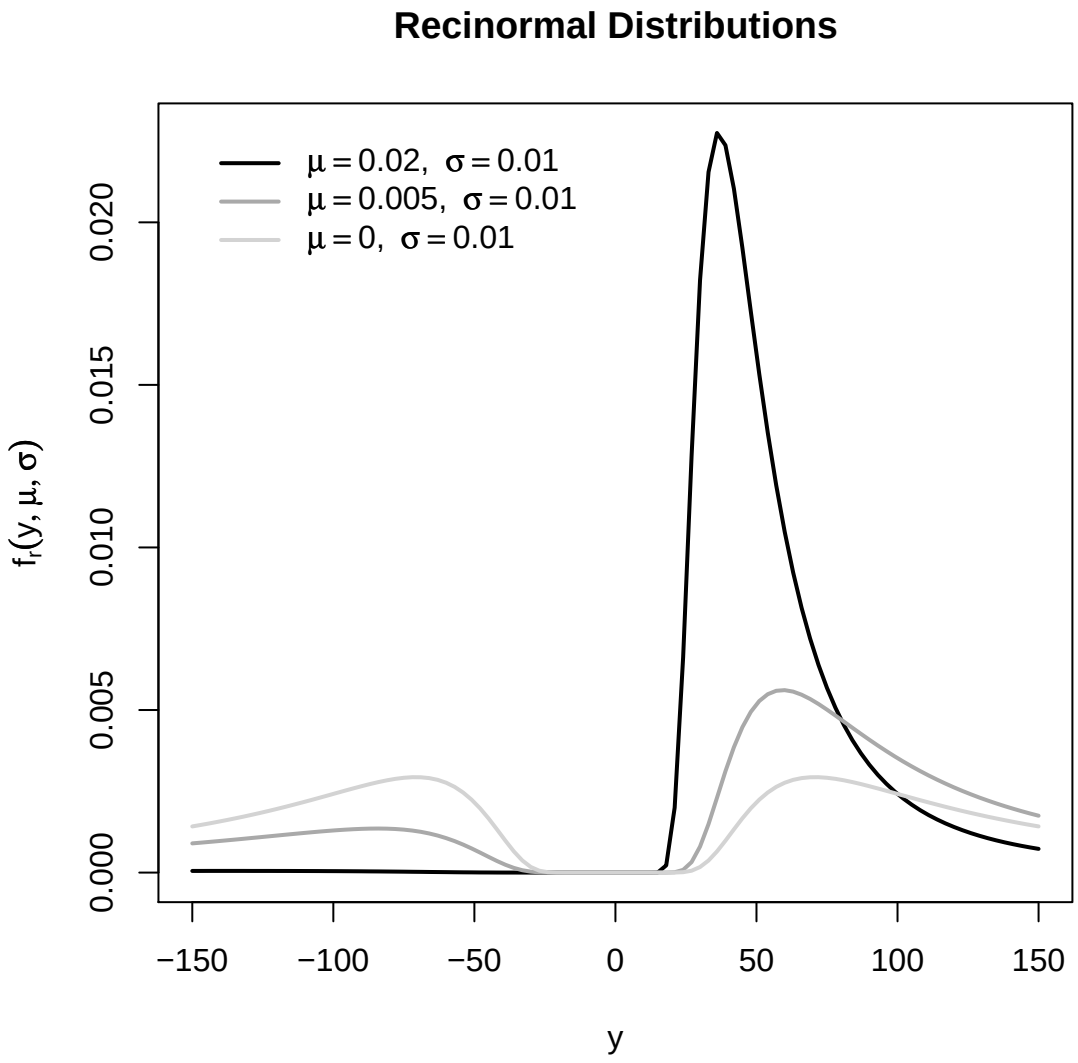

Figure A1. Examples of recinormal distributions with different values for the μ parameter.

Maximum-likelihood estimation

The maximum-likelihood estimation of the μ and σ parameters of the Recinormal distribution derives naturally from its relationship to the plain normal distribution through the reciprocal. This relation implies that one can estimate the values of both parameters using the unbiased estimators for the normal parameters of the reciprocal. Therefore, if we have observed a series of n RTs $t_1, \ldots, t_n$ with reciprocals $f_i = 1/t_i$, then our estimators are:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f_i = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{t_i}. \tag{39}$$

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \widehat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{t_i} - \widehat{\mu}\right)^2. \tag{40}$$

Appendix B
Fieller's Normal Ratio Distribution

Let X_1 and X_2 be normally distributed random variables with respective means θ_1 and θ_2 and standard deviations σ_1 and σ_2 and a Pearson correlation coefficient of ρ. Let W be the random variable resulting from the quotient of X_1 and X_2 ($W = X_1/X_2$). The distribution of W is an old problem whose study, according to Craig (1928), dates back to (and was left unsolved by) Pearson (1910). Craig (1928) provided expressions for its moments, when they do exist. Hinkley (1969) gives the general distribution for W, which he attributes to Fieller (1932). The probability density function for W is:

$$f(w) = \frac{b(w)d(w)}{\sigma_1\sigma_2 a^3(w)\sqrt{2\pi}}\left[\Phi\left(\frac{b(w)}{a(w)\sqrt{1-\rho^2}}\right) - \Phi\left(-\frac{b(w)}{a(w)\sqrt{1-\rho^2}}\right)\right] + \frac{\sqrt{1-\rho^2}}{\pi\sigma_1\sigma_2 a^2(w)}\mathrm{e}^{-\frac{c}{2(1-\rho^2)}}, \tag{41}$$

where

$$\begin{aligned} a(w) &= \sqrt{\frac{w^2}{\sigma_1^2} - \frac{2\rho w}{\sigma_1\sigma_2} + \frac{1}{\sigma_2^2}}, \\ b(w) &= \frac{\theta_1 w}{\sigma_1^2} - \frac{\rho(\theta_1 + \theta_2 w)}{\sigma_1\sigma_2} + \frac{\theta_2}{\sigma_2^2}, \\ c &= \frac{\theta_1^2}{\sigma_1^2} - \frac{2\rho\theta_1\theta_2}{\sigma_1\sigma_2} + \frac{\theta_2^2}{\sigma_2^2}, \\ d(w) &= \mathrm{e}^{\frac{b^2(w) - c a^2(w)}{2(1-\rho^2)a^2(w)}}, \end{aligned} \tag{42}$$

and Φ is the cumulative distribution function of the standard normal distribution.

Hinkley (1969) also provides the cumulative distribution function of W:

$$F(w) = L\left(\frac{\theta_1 - \theta_2 w}{\sigma_1\sigma_2 a(w)}, -\frac{\theta_2}{\sigma_2}; \frac{\sigma_2 w - \rho\sigma_1}{\sigma_1\sigma_2 a(w)}\right) + L\left(\frac{\theta_2 w - \theta_1}{\sigma_1\sigma_2 a(w)}, \frac{\theta_2}{\sigma_2}; \frac{\sigma_2 w - \rho\sigma_1}{\sigma_1\sigma_2 a(w)}\right), \tag{43}$$

where $L(i, j; k)$ is the value at (i, j) of the cumulative distribution function of a standard bivariate normal distribution with Pearson correlation coefficient k.

Although in the original characterization given above this distribution appears to have five free parameters, in effect four parameters are sufficient to fully describe it; the crucial

values that determine the distribution are the correlation coefficient, the ratio between the normal means, and the scale of the variation parameters relative to the corresponding mean. Therefore, we can describe any instance of Fieller's distribution with four degrees of freedom, correspoding to the parameters:

$$\begin{aligned} \kappa &= \frac{\theta_1}{\theta_2}, \\ \lambda_1 &= \frac{\sigma_1}{|\theta_1|}, \\ \lambda_2 &= \frac{\sigma_2}{|\theta_2|}, \\ -1 < \rho < 1. & \end{aligned} \tag{44}$$

Approximation by normal and/or recinormal distributions

Fieller's distribution can either be unimodal or bimodal. Marsaglia (1965) noticed that for small values of λ_2, the number of modes in the distribution is mostly determined by the value of λ_1. If λ_1 is greater than approximately .443066 then the distribution will be unimodal, and otherwise it will be bimodal, having one positive and one negative mode.[15] For instance, the Recinormal distribution is a limiting case of Fieller's distribution when $\lambda_1 \to 0$, and it is thus always bimodal. However, Marsaglia's assymptotic argument has a limited value. Note, for example, that the normal distribution is also a particular limiting case of Fieller's (when $\lambda_2 \to 0$) and it is always unimodal regardless of the value of λ_1 despite the very low value of λ_2.

In addition, both Geary (1930) and Marsaglia (1965) estimated that – in unimodal cases – if the value of λ_2 is smaller than $1/3$, then the distribution will be well approximated by a normal distribution. In a similar direction, Hayya, Armstrong, and Gressis (1975) indicate that, as long as $\lambda_2 < .39$, the distribution can be approximated using a normal. However, our own empirical studies taking intercorrelation into account – see Figure 2 – we find that these two are overoptimistic estimates as values of λ_2 smaller than $1/3$ can sill result is significant deviations from normality

The possibility of approximating Fieller's distribution using a normal distribution was also studied by Hinkley (1969). He proposed that in cases when $0 < \sigma_2 \ll \theta_2$ (*i.e.*, very small values of λ_2 in our notation), the cumulative distribution function of Fieller's distribution is well approximated by:

$$F(w) \simeq F^\star(w) = \Phi\left(\frac{\theta_2 w - \theta_1}{\sigma_1 \sigma_2 a(w)}\right), \tag{45}$$

with Φ being the standard normal cumulative distribution function, and $a(w)$ as defined in (42). Explicit error bounds for this approximation are also provided by Hinkley, with a

[15] Marsaglia (1965) made this argument on the basis of the values of the *means* θ_1 and θ_2, provided that the standard deviations were 1 and thus he placed the threshold at $\theta_1 \simeq 2.257$. We have rephrased this in terms of the coefficients of variation λ_1 and λ_2 to extend it to the general case.

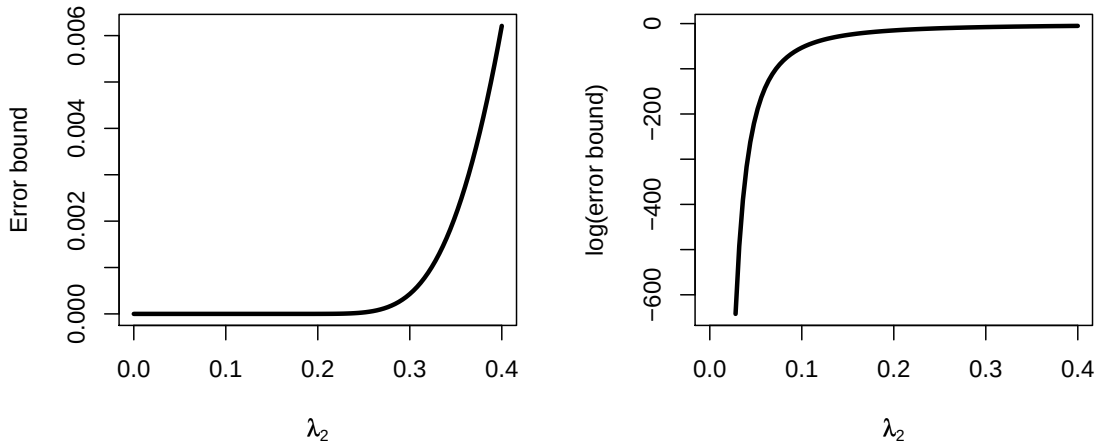

Figure B1. Error bounds (left panel) and logarithmic error bounds (right panel) given by Hinkley (1969) for the normal approximation to Fieller's cumulative distribution as a function of the value of the λ_2 parameter.

simple unrefined version of these bounds being given by:[16]

$$|F(w) - F^{\star}(w)| \leq \Phi\left(-\frac{1}{\lambda_2}\right). \tag{46}$$

The evolution of these error bounds as a function of λ_2 is illustrated in the left panel Figure B1. The figure shows that the error bounds are are insignificant as long as λ_2 is smaller than around .2, but become very serious from there onwards, from where the become rather stable (in logarithmic scale, see right panel). Taken together with our Monte Carlo simulations (Figure 2) investigating the shape of the distribution, we can say that, if λ_2 is smaller than .2, Fieller's distribution is to all effects normal. The above also extends to the reciprocal case. If $\lambda_1 < .2$, the reciprocal of the distribution will be normal, that is to say, it will be a recirnormal distribution.

Approximation by a Cauchy distribution

If both random variables X_1 and X_2 have mean values of zero (or equivalently $\lambda_1 \to \infty$ and $\lambda_2 \to \infty$ in our notation), then the distribution of their ratio is exactly a Cauchy distribution (also known as Lorentz distribution in the physics domain), with probability density function:

$$f(x) = \frac{1}{\pi s\left(1 + \left(\frac{x-m}{s}\right)^2\right)}, \tag{47}$$

[16] A more sophisticated and accurate expression for the bounds is also given by Hinkley (1969; (13) on page 638), but this simpler version suffices for our purposes, and has the advantage of being a function of λ_2 only, irrespective of the value of w.

with m and s being the location and scale parameters of the distribution, respectively indicating its median and half-width.

Therefore the Cauchy distribution is also a limiting case of Fieller's distribution.A remarkable property of the Cauchy distribution is that all its moments are undefined (*i.e.*, the corresponding integrals that define them do not converge). This has the implication that, if responses follow a distribution of this type, measures like empirical means and variances will not be replicated from one experiment to another, no matter how well controlled the experiments are or how large the sample is. It becomes important then to see how close to infinity one needs to be for the distribution to become Cauchy-like.

Figure 2 indicated that the divergence between Fieller's distribution and a normal distribution stabilizes at a negentropy of around 4 nats when λ_2 is greater than approximately .4, this being approximately the value of the negentropy of a Cauchy distribution. In addition, as we saw above, Marsaglia (1965) determined that if the value of λ_2 is greater than .44, Fieller's distribution will be unimodal. Taking these to facts together, we have that if $\lambda_2 > .44$, we will have a unimodal distribution with a negentropy equivalent to that of a Cauchy. We could therefore say that above this value the distribution is in fact well approximated by a Cauchy, but a further caution needs to be taken. An important property of the Cauchy distribution is that its reciprocal is also Cauchy distributed. By the same logic, for the reciprocal to be similar to a Cauchy in terms of the number of modes and the negentropy, it is necesary that λ_1 is also greater than the .44 threshold. Therefore, we can say that Fieller's distribution is well-approximated by a Cauchy distribution when both its CoV parameters are above the threshold.

If either of the CoV parameters is below the Cauchy threshold of .44, and still both are above the normal thresholds of .2, Fieller's distribution will be in a somewhere in-between a normal or recinormal distribution and a Cauchy distribution. These cases are illustrated by Figure B2. The panels plot the log-densities of different instances of Fieller distribution (we plot the log densities in order to be able to visualize also very small secondary modes). The upper two panels are instances of Fieller's falling into the main normal (left) and recinormal (right) areas. On the other extreme, the bottom panels represent two cases where the distribution has also entered the Cauchy zone. Note that both of them are unimodal, but the plot on the left shows a more pronounced skewness than the plot in the right. In fact the left plot appears as a 'tilted' version of a typical Cauchy log density plot. This is due to the plot in the left being much closer to the .44 threshold ($\lambda_1 = \lambda_2 = .5$) than the plot on the right ($\lambda_1 = \lambda_2 = 2$). Otherwise, both show the typical 'thick' tails of Cauchy distributions, which can be appreciated by the concavity of both log-tails in the two cases. Finally the plots in the middle panels correspond to instances that are in the transitional zones, with the plot in the left already approaching a Normal, while the plot in the right is bimodal and approaches recinormal. However, both of them also have a Cauchy component, which is again reflected in the concavity of the log-tails in the left plot. In fact, both plots are reciprocally related (one is the distribution of the other's reciprocal), thus this Cauchy component affects both of them.

Shift invariance

Fieller's distribution is robust under a constant shift, such as the one that would be introduced by a constant non-decisional component. This is easy to see. Let W be an

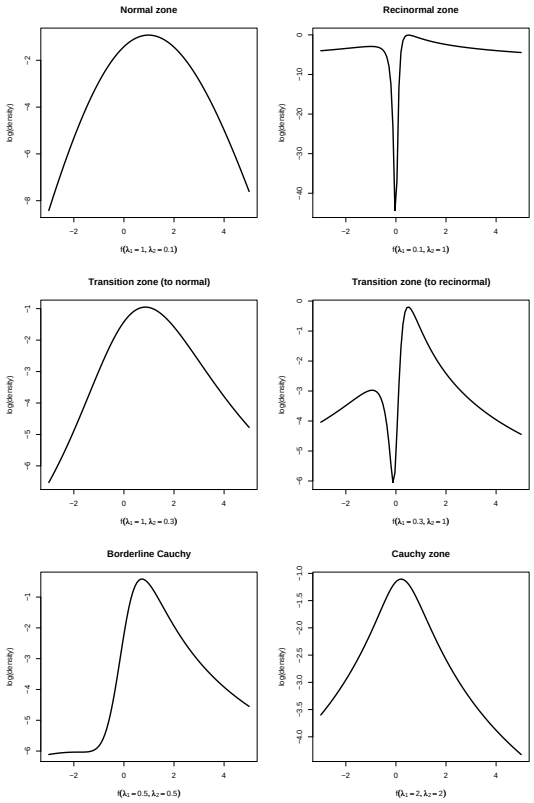

Figure B2. Especial cases of Fieller's distribution for different values of the λ_1 and λ_2 CoV parameters. The curves plot the log densities of Fieller's distribution (with the remaining parameters $\kappa = 1$ and $\rho = 0$ in all cases). The top panels represents a normal (left panel) and a recinormal (right panel) case. The bottom panels plot two instances where the distribution approaches Cauchy. One just a the limit (left panel) and one well into the Cauchy zone (right panel). The middle panels represent transitional distributions closer to the normal zone (left panel) and recinormal zone (right panel), but that are not well approximated by neither normal/recinormal, nor Cauchy distributions.

instance of Fieller resulting from the ratio of normally distributed variables X_1 and X_2, and let β be a constant shift that is added to W:

$$W_s = W + \beta = \frac{X_1}{X_2} + \beta = \frac{X_1 + \beta X_2}{X_2}. \tag{48}$$

The numerator of this expression is a linear combination of normally distributed variables, and it is itself normally distributed. Therefore, W_s is distributed according to Fieller's distribution, but its parameters κ, λ_1 and ρ are different than those of W, while λ_2 remains unchanged. (48) provides a direct way for removing the correlation by substracting a constant β. If we have estimated the values of $\widehat{\kappa}$, $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ and $\widehat{\rho}$ from a dataset. We can get new estimators $\widehat{\kappa}'$, $\widehat{\lambda}'_1$, and $\widehat{\lambda}'_2$ so that $\widehat{\rho}' = 0$. The estimated value of $\widehat{\beta}$ is given by the expression:

$$\widehat{\beta} = \widehat{\rho}\,\frac{\widehat{\sigma}_1}{\widehat{\sigma}_2} = \widehat{\rho}\,\frac{\widehat{\lambda}_1}{\widehat{\lambda}_2}\,|\widehat{\kappa}|. \tag{49}$$

From which the estimation of the rest of the parameters becomes trivial:

$$\widehat{\kappa}' = \widehat{\kappa} - \widehat{\beta}. \tag{50}$$

$$\widehat{\lambda}_1' = \widehat{\lambda}_1\sqrt{1-\widehat{\rho}^2}\left|\frac{\widehat{\kappa}}{\widehat{\kappa}-\widehat{\beta}}\right|. \tag{51}$$

$$\widehat{\lambda}_2' = \widehat{\lambda}_2. \tag{52}$$

A less clear modification takes place when, rather than being a plain constant, the shift is itself normally distributed. Let X_3 be a normally distributed random variable:

$$W_s = W + X_3 = \frac{X_1 + X_2X_3}{X_2}. \tag{53}$$

Unfortunately, the numerator is not a linear combination of normal variables any longer, but instead it contains a normal term (X_1) and a normal product term (X_2X_3). However, as detailed in Appendix C, the distribution of a product of normal variables is itself well-aproximated by a normal distribution of known mean and variance as long as the CoV of at least on the variables remains low or moderate. In fact, the value of the CoV of X_2 is rarely going to be higher than .5, and in these cases the divergence from normality is insignificant. Therefore, although strictly speaking the numerator of (53) may or may not be a linear combination of normally distributed terms, in practice we can assume it is, at least as long as none of the normal variables involved has a very high CoV. This implies that, even if the shift term added is not a constant, bu rather normally distributed, the resulting distribution will be well-described by another instance of Fieller's.

Relative stability

As we have seen above, Fieller's distribution is robust to the addition or substraction of a constant and, at least under certain constraints, is also solid to the addition (or substraction) of a normally distributed amount. A related issue is whether the sum of Fieller distributed variables gives rise to another instance of Fieller's distribution. In other words, the question is whether Fieller's distribution is an instance of a stable distribution.

In strict terms, the question of stability is a complex and interesting mathematical problem in its own right, that goes beyond the scope of our study. As a quick summary, only three distribution families are known to have this property: the normal distribution, the Cauchy distribution, and the Lévy skew alpha-stable distribution (LSASD), which is a generalization of the two others (cf., Nolan, 2009). A possible direction to prove this would be to show that Fieller's distribution is in fact subsumed by or equivalent to the LSASD. Both are four parameter distributions, and both have the Dirac, normal, and Cauchy distributions as special cases. Here we will limit ourselves to see if a sum of Fieller's distributions is at least well-approximated by another instance of Fieller's.

Let X_1, X_2, X_3, and X_4 be normally distributed random variables with respective CoVs λ_1, λ_2, λ_3 and λ_4. Let $\rho_{i,j}$ denote the coefficient of correlation between variables X_i qnd X_j, and $V_1 = X_1/X_2$, $V_2 = X_3/X_4$ be instances of Fieller's distribution independent of each other. The variable W is:

$$W = V_1 + V_2 = \frac{X_1}{X_2} + \frac{X_3}{X_4} = \frac{X_1X_4 + X_2X_3}{X_2X_4}. \tag{54}$$

Both the numerator and denominator are in this case expressed in terms of normal product distributions. Therefore, loosely speaking as long as one of λ_1 or λ_4, and one of λ_2 or λ_3 have a low or moderate value, we can safely expect W to be an instance of Fieller's distribution. In addition, having λ_2 or λ_4 close to zero, puts us back into the constant shift case, which as we saw above is Fieller distributed. Furthermore, if all λ_i have relatively high values (*i.e.*, above .4, then both V_1 and V_2 will in fact be well-approximated by a Cauchy distribution. Fortunately, Cauchy is one of the three distributions which is known to be stable, and thus the sum of two (independent) Cauchy distributions will itself be a Cauchy distribution, and thus an instance of Fieller (cf., Nolan, 2009). In sum, although we have not proved that the sum of independent Fieller ditributed variables is Fieller distributed, we at least see that a Fieller distribution will in many cases provide a good approximation to the actual distribution.

Appendix C
The normal product distribution

The distribution of the product of two correlated normal variables is a generally complex one, given in the form of a long series expansion by Craig (1936) and in integral form by Aroian (1947) and Aroian, Taneja and Cornwell (1978). Interestingly, as is the case with Fieller's, the shape of this distribution is also determined by the correlation coefficient between both variables and their individual CoVs. Fortunately, even though the distribution is clearly non-normal, Aroian (1947) notes that as either (or both) of the CoV parameters of the distribution approach zero, the distribution converges on normality.[17] Furthermore, he noted that this convergence is faster if the correlation coefficient between both variables has the same sign as the product of their means. We find that, in fact, for moderate values of the CoV parameters, it is in fact well-approximated with a normal distribution.

Let X_1 and X_2 be normally distributed variables with respective means θ_1, θ_2, standard deviations σ_1, σ_2, CoVs λ_1 and λ_2, and $V = X_1X_2$. Figure C1 shows the effect of varying the CoV parameters on Montecarlo estimates of the negentropy of a normal product distribution. Note that even for values of the CoVs as high as $\lambda_1 = 3.2$ and $\lambda_2 = 2$, the estimated values of the negentropy remain very low – a negentropy .2 nats reflects a very small deviation from normality; compare with the negentropy values for Fieller's distribution plotted in Figure 2. Furthermore, if either of the CoV is below .2 (vertical dotted lines in the figure) the divergence from normality is to all effects null. Therefore we can safely generalize Aroian (1947)'s theorems on normal convergence of normal product distributions as long as at least one of the two CoV parameters remains relatively low. In addition, as mean and variance of this normal approximation we can use Craig (1936)'s expression for the moments of V:

$$\begin{aligned} \theta_V &= \theta_1\theta_2 \\ \sigma_V^2 &= \sigma_1^2\theta_1^2 + \sigma_2^2\theta_2^2 + \sigma_1^2\sigma_2^2. \end{aligned} \tag{55}$$

[17] In fact both Aroian and Craig worked on the reciprocal CoVs, and thus the theorems were stated with the reciprocal CoVs approaching infinity (both positive and negative).

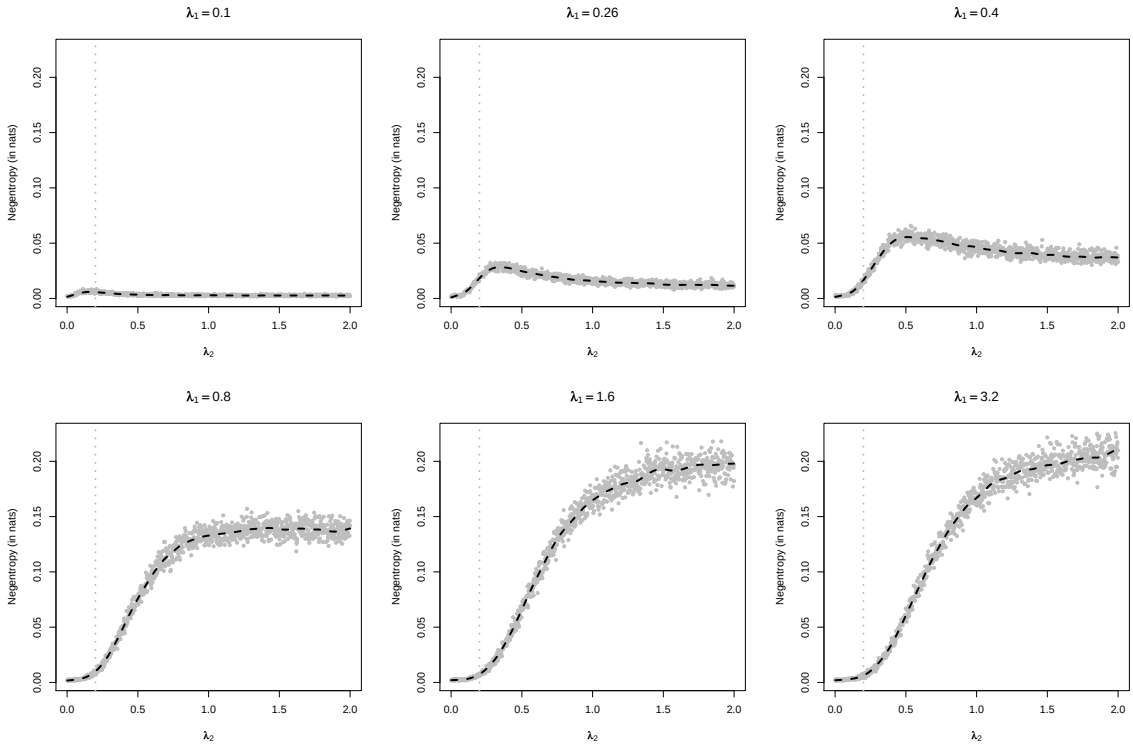

Figure C1. Deviations from normality and recinormality for a normal product distribution for different values of the coefficients λ_1 and λ_2. The plots are Montecarlo estimates of the negentropy for changing values of λ_2 fixing the value of λ_1. Each point represents a Montecarlo estimate of the negentropy based on a sample of 10,000 items from a normal product distribution. As fixed parameters, we used realistic values estimated from visual lexical decision data ($\theta_1 = 450, \theta_2 = 1, \rho = .59$), and the values of σ_1 and σ_2 were calculated to give the apropriate values of λ_1 and λ_2. The vertical dotted lines indicate the approximate locations of the phase change at .2. The black dashed lines are non-parametric regression smoothers.

Appendix D
Distribution Fitting for Large Datasets

Fitting parametric probability density functions to very large datasets can be a complex computationally demanding task. In order to obtain the fits to the English Lexicon Project datasets in this study, we developed a technique that enables very fast and accurate distribution fitting, even for extremely large datasets.

The goal of distribution fitting is to fing a set of paramater values $\widehat{\theta_1}, \widehat{\theta_2}, \ldots$ that optimize the correspondence between a parametric probability distribution family with density function $f(x|\theta_1, \theta_2, \ldots)$ and a list of n observed data points $x_1, x_2, x_3, ..., x_n$. Formally this is equivalent to finding the values of the parameters that minimize the Kullback-Leibler divergence[18] between the parametric distribution f and the real – unknown – probability

[18]The Kullback-Leibler divergence or cross-entropy is an information-theoretical measure of how differ-

distribution function of the data g. Therefore, we need to minimize:

$$D(g||f;\theta_1,\theta_2,\ldots) = \int_{-\infty}^{\infty} g(x)\log\frac{g(x)}{f(x|\theta_1,\theta_2,\ldots)}\mathrm{d}x \quad (56)$$

$$= h(g) - \int_{-\infty}^{\infty} g(x)\log f(x|\theta_1,\theta_2,\ldots)\,\mathrm{d}x \quad (57)$$

$$\propto -\int_{-\infty}^{\infty} g(x)\log f(x|\theta_1,\theta_2,\ldots)\,\mathrm{d}x, \quad (58)$$

where the differential entropy of the data distribution ($h(g)$) is a constant term for all values of the parameters, and can therefore be ignored in the minimization process.

Given that $g(x)$ is actually unknown, most algorithms implicitly use the Montecarlo integration assumption that the data themselves provide a representative sample of their distribution, and thus the concentrate in maximizing the summed log-likelihood of the data as the target function in an iterative multivariate optimization algorithm such as Nelder-Mead or Newton-Raphson. This is in short the principle of the maximum likelihood method discussed by Hays (1973). Variations of this method have proved very efficient for fitting diverse reaction time distribution (see, e.g., van Zandt, 2000). This type of maximum likelihood methods are in general most efficient, at least as long as the RT measurements themselves are relatively accurate so as to allow the explicit continuity assumption (Cousineau, Brown, & Heathcote, 2004).

However, when datasets become very large, unless there exists a tractable analytic solution for the maximum likelihood estimates, this approach becomes computationally very expensive as each iteration of the chosen optimization algorithm involves a full evaluation of the target log-likelihood, and possibly estimates of its Hessian or Lagrangian as well. A possible solution when dealing with very large datasets is to use only a subsample of the data to perform the fits. An alternative solution to this problem is offered by methods that summarize the data prior to fitting the distribution. In these methods, one does not fit the actual datapoints, but rather uses some summary of the data to provide a non-parametric description of g, which is then fitted using a maximum likelihood method. This is the basic idea behind the Quantile Maximum Likelihood (QML) method introduced by Heathcote, Brown, and Mewhort (2002). The method consists in obtaining the location of a set of quantiles from the data, and then using maximum likelihood to estimate the parameters that would produce the most similar quantile distribution, with the important of each quantile point being weighted by its probability density at that point. Heathcote and collaborators show that this method is more accurate than tradition maximum likelihood for several typical RT distribution families, as it is less sensitive to outliers and noise generated by finite datasets (with smaller datasets, the Montecarlo integration assumption becomes less and less warranted). A similar, and allegedly more robust method is the "vicentiles" method introduced by Ratcliff (1979) in which the observations falling within each inter-quantile bin are averaged, and these vincentiles are then used as the summary of the distribution.

---

ent two probability distributions are, being zero *iff* both distributions are identical, and strictly positive otherwise.

Kernel-Mediated Minimum Divergence

In our study, we have used a different strategy to summarize the data in order to obtain an approximation of g. The basic idea is to construct a relative histogram g^* of the data by binning the x_i into $m \ll n$ intervals centered at $d_1, d_2, \ldots, d_m$. This histogram provides a non-parametric approximation to a discrete version of g. Similarly, if the spacing δ between the d_i is sufficiently small, and f can be assumed to be zero beyond certain limits, then we can approximate f by a discrete version f^* so that:

$$f^*(d_i|\theta_1, \theta_2, ...) = \int_{d_i-\delta/2}^{d_i+\delta/2} f(x|\theta_1, \theta_2, ...)\mathrm{d}x \tag{59}$$

$$\simeq \delta f(d_i|\theta_1, \theta_2, ...). \tag{60}$$

Then we can estimate the paramaters by minimizing the corresponding discrete version of (58):

$$J(\theta_1, \theta_2, \ldots) = -\sum_{i=1}^{m} g^*(d_i) \log f^*(d_i|\theta_1, \theta_2, \ldots) \tag{61}$$

$$\simeq -\log\delta - \sum_{i=1}^{m} g^*(d_i) \log f(d_i|\theta_1, \theta_2, \ldots) \tag{62}$$

$$\propto -\sum_{i=1}^{m} g^*(d_i) \log f(d_i|\theta_1, \theta_2, \ldots), \tag{63}$$

where the term $-\log\delta$ in (62) does not depend on the parameters $\theta_1, \ldots$ and can thus be omitted from the optimization, which is equivalently done using (63).

Setting g^* to be a plain relative histogram of the observed counts in each bin d_i provides in many cases a good non-parametric approximation of the density function, and its computation is extremely fast and simple ($\mathcal{O}(n)$). However, using the plain histogram is very sensitive to noise, a problem that is especially marked in the tails of the distributions, were only a few points will occur, leading to many empty bins and the corresponding high frequency noise. In order to attenuate this problem, we can use a smoother on the histogram, so that part of this high frequency noise is removed. A common technique for doing this is to use Kernel Density Estimation methods (KDE; cf., Silverman, 1986). KDE is a family of non-parametric methods to estimate the density of a set of points, without a-priori knowledge of the type of function that generated them. Given a chosen bandwidth σ, and a particular *kernel function* K, the KDE estimate of the probability density function of the data is:

$$g_k(x|\sigma, K) = \frac{1}{n\sigma} \sum_{i=1}^{n} K\left(\frac{x_i - x}{\sigma}\right). \tag{64}$$

There are several typical choices of a kernel function, the most common being the rectangular, triangular, Epanechnikov's, cosine, and Gaussian functions. In general, all of these kernels give rise to good approximations of an unknown density function. In this work, we have used the Gaussian kernel function, as this is the default function in the KDE

function provided by default by the R statistical software.[19] The Gaussian kernel function is given by:

$$K(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}}. \tag{65}$$

Using this kernel amounts to centering a Gaussian density function of standard deviation σ at each observed point, and then taking the normalized sum of the Gaussian components to represent the density of the data. Therefore, the smoothness of the resulting distribution will depend strongly on the choice of bandwidth σ: Higher values of σ will result in smoother density estimates. In order to choose the value of σ we use Silverman (1986)'s "rule of thumb" method which is used by default in the R implementation:

$$\sigma = .3n^{-\frac{1}{5}} \min \left\{ s, \frac{q_{.75} - q_{.25}}{1.349} \right\}, \tag{66}$$

with $q_{.25}$ and $q_{.75}$ being the first and third quartiles of the data and s its sample standard deviation. For instance, in the large lexical decision and word naming datasets, this method suggested a bandwidth of around 5 ms. The process of computing the estimate is speeded up by doing the calculations working on a discretized space using frequency domain convolutions (via the Fast Fourier Transform) and then using linear interpolation to return to the continuous domain. It will therefore suffice for our purposes to use the discrete support points as the bins in our discrete version:

$$g_k^*(d_i) = \int_{d_i-\delta/2}^{d_i+\delta/2} g_k(x)\mathrm{d}x = \delta g_k(d_i) \propto g_k(d_i), \tag{67}$$

where, as was done for f^*, the constant δ term can be omitted.

g_k^* constitutes a smoothed version of the histogram g^*, which – depending on the σ used – is less sensitive to the noise and outliers in the original data. We can now minimize (63) using g_k^* in place of g^* by any standard iterative optimization algorithm. In the current study, we have used the Nelder-Mead algorithm to perform the minization. As we defined $m \ll n$, the process of fitting f^* to g_k^* is much faster than fitting the whole set of data points. Furthermore, the smoothing provided by the KDE method will make the method less sensitive to noise and outliers than the raw maximum likelihood methods, a property that is also shared by the QMLE and vincentizing methods mentioned above. By default, in this study we have used a grid of 512 support points or bins (*i.e.*, $m = 512$) spaced at regular intervals. We have used the default choice of extremes such that $d_1 = \min\{x_1, \ldots, x_n\} - 3\sigma$ and $d_m = \max\{x_1, \ldots, x_n\} + 3\sigma$, except for distributions that are only supported on the positive hemiline (such as the Ex-Wald), for which we have set the lower extreme as $d_1' = \max\{d_1, 1\}$.

In preliminary comparisons with plain ML (using subsampling on the large datasets), QML, and vincentizing techniques, the Kernel-mediated Minimum Divergence (KmD) method that we describe here seemed to perform best, most markedly in very large datasets. A more formal comparison of this new method with the others remains to be done to have a better understanding of the situations in which one should prefer one method or another.

[19] Function `density` in the general `stats` package. *R* Development Core Team and contributors. `http://www.r-project.org`.

In any case, the non-parametric summarization of the data distribution by KDE prior to parametric distribution fitting, presents a clear advantage in terms of robustsness and speed (the whole calculations for datasets greater than 1M points takes less than one second on a two-year old mid-range laptop).

Appendix E
Estimation of Hazard Functions

The estimation of the hazard function from a sample of an unknown distribution can be problematic, particularly so in the tails of the distribution. In our experience, the non-parametric method that was discussed by Burbeck and Luce (1982) and Luce (1986) (and that they attribute to Miller and Singpurwalla, 1977) seems to provide the best results, and it is thus the method we have used to estimate the hazard functions in this study.

The method consist in ordering the RTs in an increasing sequence. If Z_i is the i-th smallest RT, we obtain the sequence:

$$Z_0 = 0 \leq Z_1 \leq \ldots \leq Z_n. \tag{68}$$

If we additionally define a new sequence:

$$S_i = (n - i + 1)[Z_i - Z_{i-1}], \tag{69}$$

then the hazard function is estimated by a step function H whose i-th element is:

$$H_i = \frac{j}{\sum_{k=i-j+1}^{i} S_k},\ j < i < n. \tag{70}$$

j is a smoothing parameter that is set in advance. Higher values of j result in smoother estimates of the hazard function, but come at the cost of loosing information on its actual shape. In our analyses, we have set j to a fixed value of 30, as this produced a good compromise between smoothness and accuracy. In addition, as recommended by Burbeck and Luce (1982), for the first 29 intervals, we have set $j = i$.