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 August 16, 2009

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 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. 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 multiple families of distributions that have been proposed to account for RTs, indicating our theory provides a significantly better account of experimental data. For this purpose, we provide comparisons with four large datasets across tasks and modalitities. Finally, we show how the theory links to neurobiological models of response latencies.

Keywords: Drift Diffusion Model; Ex-Gaussian; Ex-Wald; LATER; Power-Law; Ratio Distribution; RT Distribution

Since its introduction by Donders (1869), reaction time (RT) has 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 account for by most general probability distribution families. This has in many cases led to

---

This work was partially supported by the European Commission through a Marie Curie European Reintegration Grant (MERG-CT-2007-46345).
The author wishes to thank Anna Montagnini for having seeded the thoughts contained in this paper, and Xavier Alario, Borís Burle, Laurie Feldman, Yousri Marzouki, Jonathan Grainger, and Xavier Waintal for discussion and suggestions on these ideas. Correspondence can be addressed to:
`fermin.moscoso-del-prado@univ-provence.fr`

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 to put 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.

We will discuss this problem in four stages. First, we provide 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. 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 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. 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 several lexical processing datasets in across tasks and modalities. Finally, 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

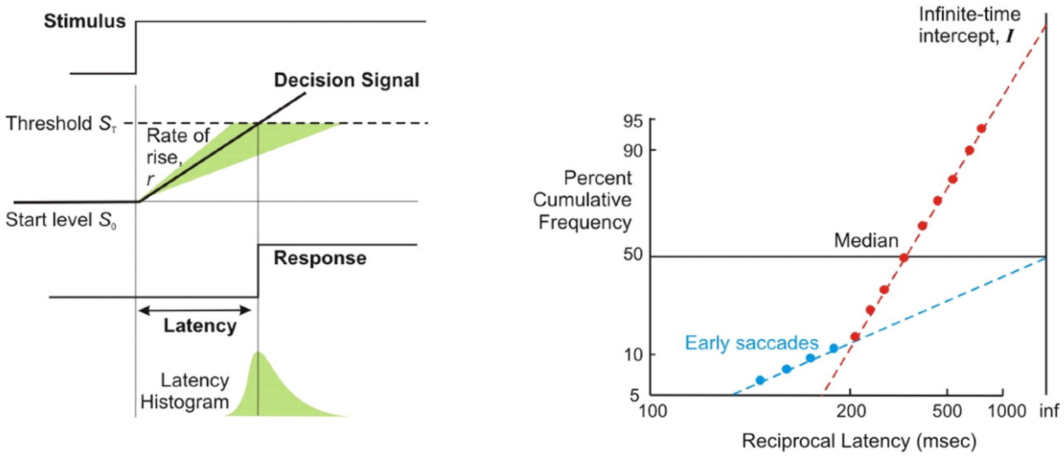

Figure 1. *Left panel*: Schema of the LATER model. Evidence accumulates from an initial state (S_0) to a decision criterion (θ). The rate (r) at which the evidence accumulates varies according to a normal distribution with mean μ_r and variance σ_r^2, giving rise to the typical skewed distribution of 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.*

going from button presses to ocular saccades (*e.g.,* Carpenter, 1981, 1988, 2000, 2001; Carpenter & McDonald, 2007; 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 distribution of the times is the reciprocal of a normal distribution:

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

This distribution is what Carpenter terms a *Recinormal 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:

$$S(T) = \log \frac{P(H|E)}{1 - P(H|E)} = \log \frac{P(H)}{1 - P(H)} + \int_0^T \log \frac{P(E|H)}{1 - P(E|H)} \mathrm{d}t = S(0) + rT. \quad (3)$$

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.

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:

$$\frac{1}{T} = \frac{r}{\theta - S(0)} = \frac{r}{\Delta}. \quad (4)$$

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:

$$\mu = \frac{\mu_r}{\Delta},\ \sigma^2 = \frac{\sigma_r^2}{\Delta^2}, \quad (5)$$

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

$$\text{slope} = \frac{1}{\sigma} = \frac{\Delta}{\sigma_r}. \quad (6)$$

$$\text{intercept} = \frac{\mu}{\sigma} = \frac{\mu_r}{\sigma_r}. \quad (7)$$

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 intercept 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). They 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. However, Montagnini and Chelazzi (2005) provide evidence that manipulations in urgency can also affect te intercept of the reciprobit plot. 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; but see also Johnston & Everling, 2008 for evidence that these express responses might not be of subcortical origin).

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

[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.

($\Delta = \theta - S_0$) remains fairly constant. For several reasons, it is difficult to assume that 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 participants 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:

$$T = \tfrac{\Delta}{r}, \quad r \sim \mathrm{N}(\mu_r, \sigma_r^2),\, \Delta \sim \mathrm{N}(\mu_\Delta, \sigma_\Delta^2) \ . \tag{8}$$

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. 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

[2]This can be reduced to three free parameters if we are only interested in the shape of the distribution and not on its scale (i.e., median value) which is given by the κ parameter. In this case, RTs that have been normalized (i.e., divided) by their median value require only the λ_1, λ_2, and ρ parameters

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:

$$\frac{X_1}{X_2} \sim Fieller(\kappa, \lambda_1, \lambda_2, \rho)$$
$$\kappa = \frac{\mu_1}{\mu_2}, \quad \lambda_1 = \frac{\sigma_1}{|\mu_1|}, \quad \lambda_2 = \frac{\sigma_2}{|\mu_2|}. \tag{9}$$

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.

Special 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 that 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. These four particular cases of Fieller's distribution are summarized in Table 1.

These particular cases of Fieller's distribution can be safely extended to the threshold shown in parentheses in Table 1. Independently of the value of λ_1, if $\lambda_2 < .22$, the distribution is in all respects normal. In what follows we refer to this as the normal zone. Conversely, if $\lambda_1 < .22$, the distribution is in indistinguishable from a recinormal, thus we refer to this as the recinormal zone. As soon as both λ_1 and λ_2 rise above around .443 the distribution approaches Cauchy distribution so we refer to this as the Cauchy zone. When either λ_1 or λ_2 lie between .22 and .4, there is a linear, rapidly growing deviation from (reci-)normality towards the Cauchy distribution. We refer to this area of the plots as the linear zone. In sum, 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.

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

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

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:

$$h(t) = -\frac{\mathrm{d}\log\left(1 - F(t)\right)}{\mathrm{d}t} = \frac{f(t)}{1 - F(t)}, \tag{10}$$

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, they 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).

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. In particular, as λ_2 approaches zero, the peak location goes to infinity, ultimately becoming a monotonically increasing function – a Gaussian hazard.

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,

Table 2: Classification of RT 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 (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). The 'Cocktail' model refers to the piecewise distribution recently proposed by Holden et al. (2009).

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 'Cocktail' model	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

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. Table 2 classifies several common RT distributions according to Clauset and colleagues' taxonomy.[3] The classification has been performed by considering the dominant term in the probability density functions of each distribution. 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

[3] We have added a class to accommodate the Gaussian (Clauset and colleagues consider only thick-tailed distributions

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.

Some theories have proposed RT distributions whose tails are heavier than exponential. Among these, the Instance Theory of Automaticity proposes that RTs should have Weibull (stretched exponential) tails (Colonius, 1995; Logan, 1988, 1992, 1995). Stretched exponential distribution can show – for certain values of its shape parameter[4] – thicker than exponential right tails. Heavy tails can also result from log-normal type distributions which have been proposed as models of RTs (Luce, 1986; Woodworth & Schlosberg, 1954). Few models predict even heavier power-law right tails. Recently, Holden, Van Orden, and Turvey (2009) have proposed a 'Cocktail' model which provides a piece-wise description of the RT distribution as a combination of a log-normal distribution describing the core of the data, and a thick power-law becoming dominant from a certain point in the right tail. Also, although not stated explicitly by Carpenter and colleagues, the LATER model would also predict a power-law type of right tail, which is also characteristic of the extension we propose here. More precisely, the right tail of Fieller's distribution (subsuming LATER's recinormal) converges – from a certain variable value $t_{\min}$ – to a power-law with a scaling paramater value (α) of exactly two (Jan, Moseley, Ray, & Stauffer, 1999; Sornette, 2001). The value of $t_{\min}$ is determined by the values of the CoV parameters λ_1 and λ_2, with a limiting case of $t_{\min}$ going to infinity as λ_2 goes to zero (which would correspond to a plain normal distribution).

As we have seen, our theory predicts much thicker right tails than would be predicted by most current theories, except for the few heavy-tailed distributions mentioned above. By definition, events in the right tail are very rare, but still we are predicting that they should happen much more often than one would expect in most 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, Ratcliff, Gomez, & McKoon, 2008) or even a short as 1,500 ms. (*e.g.*, Balota, Yap, Cortese, & Watson, 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 & Carpenter, 2008; Carpenter, 2001; Carpenter & Williams, 1995; Reddi & Carpenter, 2000). Figure 2 reproduces some results of Reddi and Carpenter (2000) in this respect. Notice that, specially in the time

[4] In particular, Weibull distributions will exhibit heavy tails when their shape parameter has a value smaller than one.

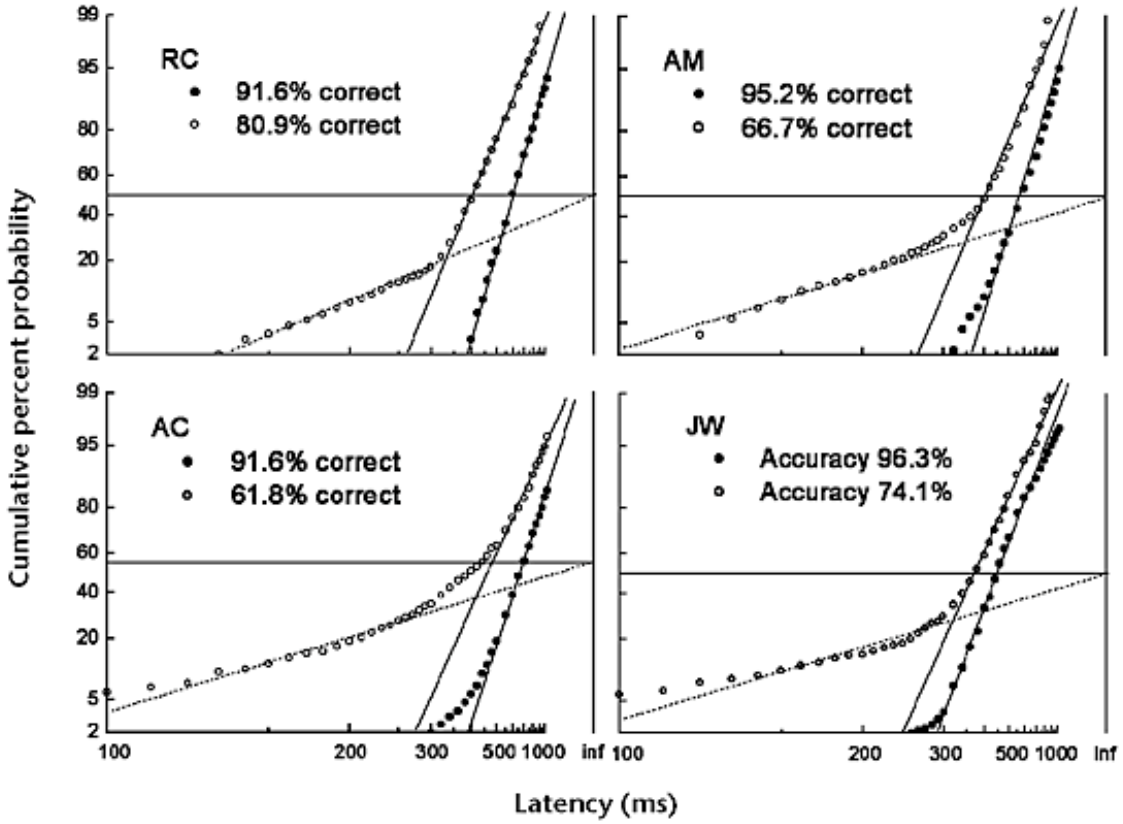

Figure 2. Evidence for the presence of a separate population of express responses. Notice that each of these Reciprobit plots can clearly be fitted by two straight lines, one for a minority of very fast responses, and one for the bulk of experimental responses. The open circles represent a condition 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.*

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. The fast sub-population arises more frequently in some conditions than others. First, as it is evident from Figure 2, the differentiated fast responses arise more clearly in participants 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.

Although the neuro-physiological mechanism that is argued to justify the very short latencies is very plausible, there is some indication that make it difficult to believe that this mechanism is responsible for the greater part of these short latencies. Following Carpenter's argument, one would 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; Carpenter & McDonald, 2007; Montagnini & Chelazzi, 2005).

What the high proportions of fast responses seem to suggest, is that those fast responses actually belong to the same distribution that generates the slower ones. In this direction, Nakahara et al. (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).

Figure 3 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. We can see that a small modification of the LATER model predicts that the majority of fast responses belong to *the same* population as the slower ones.

'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. Taking this into account, we 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 with little variability, 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:

$$T = T_n + T_d = T_n + \frac{\Delta}{r} = \frac{\Delta + T_n \cdot r}{r}, \tag{11}$$

which is also an instance of Fieller's distribution, enabling us to perform the analysis without its explicit consideration.

Errors and Alternative Responses

An issue that has become crucial when comparing theories of RTs in choice tasks is the success with which they 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

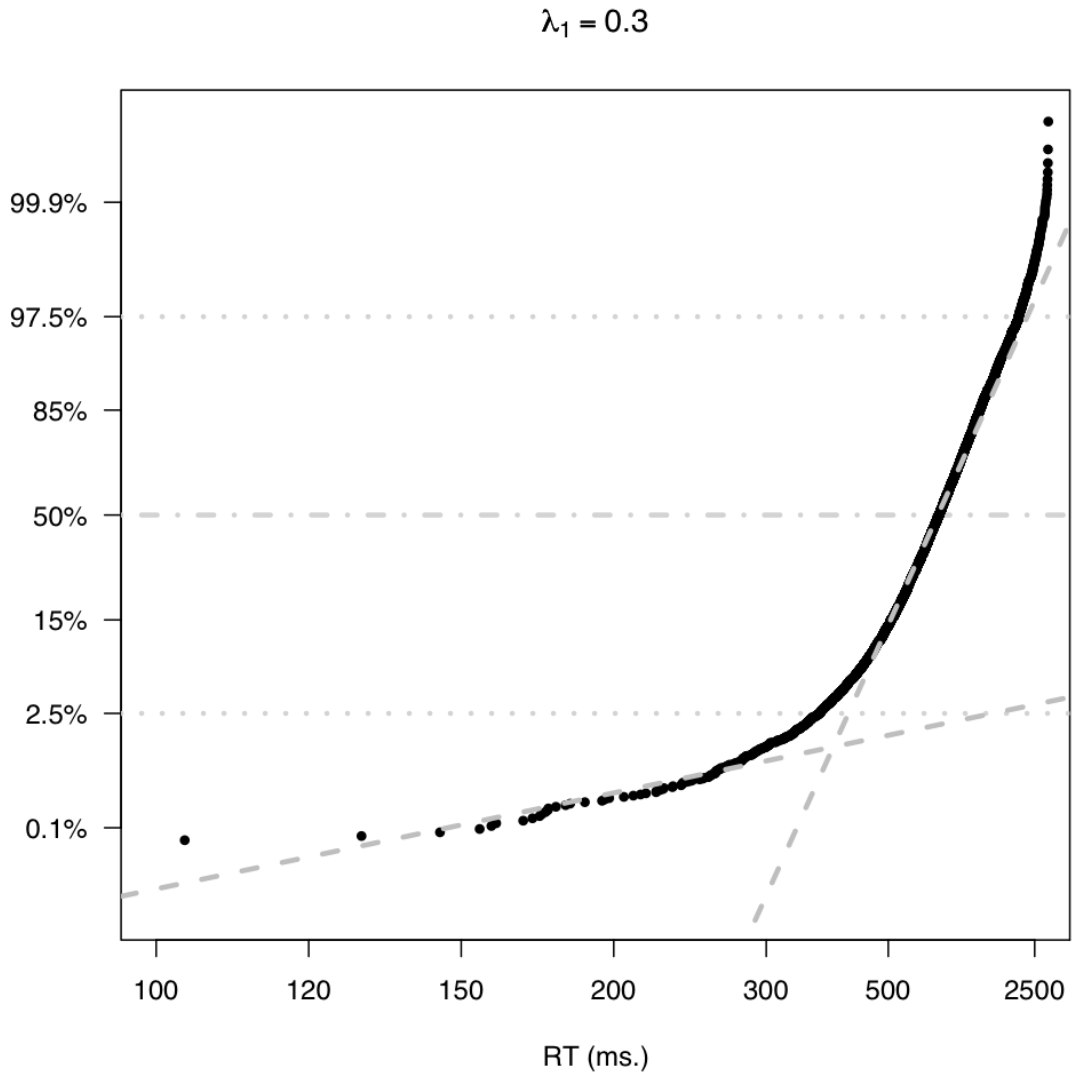

Figure 3. Typical Reciprobit plot of Fieller's distribution bordering the Recinormal zone. The data were sampled from a Fieller's distribution with parameter $\lambda_1 = .3$, that is, outside the Recinormal zone, but not yet reaching the Cauchy zone. The horizontal lines mark the median and 95% interval. 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.

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 description 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. The Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2008) 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 (1999) had proposed, LBA relies on a race between competing accumulators, and errors are produced when this race is won by the "wrong" accumulator. Importantly, the LBA model assumes that the separate accumulators are independent, that is, they are not bound by any inhibitory mechanism. In the theory that we propose, errors also arise from the competition between multiple accumulators. However, in contrast with the LBA model, we propose that some inhibition mechanism binds the accumulators 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, Holmes, & Cohen, 2006). Different accumulators simultaneously integrate evidence, and the first one to reach a threshold triggers a response.

In terms of number of errors, the predictions of our model do not differ much from the predictions of the LBA, and the same tools can be used to predict the number of errors. However,the distribution of RTs predicted by our model is significantly different than that predicted by the LBA. In the former case, by the Theory of Extreme Values, the competition of independent accumulators will give rise to a distribution that asymptotically converges to a Weibull distribution (see Colonius, 1995 and Logan, 1995 for a detailed mathematical discussion of this point). However, in our case, the inhibitory mechanisms that bind the accumulators break the independence pre-condition for application for the distribution of minima as predicted by the Theory of Extreme Values, and the distribution will be – as we argue above – more related to a power-law type.

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. The development of the theory for the two-choice case is valid without significant alterations for the general multiple choice or 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.

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) show that a race between two accumulators would be able to explain error responses. 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, Vidal, Tandonnet, & Hasbroucq, 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).

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. 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.

Fortunately, despite its explicit linear assumption, the predictions of LATER do not depend on the linear trajectories (Kubitschek, 1971). 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:

$$S(t) - S_0 = f(rt). \tag{12}$$

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

$$t = \frac{f^{-1}(S(t) - S_0)}{r} \tag{13}$$

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:

$$S(t) = S_0 \mathrm{e}^{rt}. \tag{14}$$

Then we could use the logarithmic transformation to obtain:

$$t = \frac{\log\left(S(t)\right) - \log(S_0)}{r}. \tag{15}$$

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

$$s(t) = s_0 + rt. \tag{16}$$

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 describes 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):

$$\begin{cases} \mathrm{d}S(t) &= r\,\mathrm{d}t \,+\, \sigma_r\sqrt{2t}\,\mathrm{d}W(t), \\ S(0) &= S_0 \end{cases} \tag{17}$$

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:

$$\begin{cases} \mathrm{d}S(t) &= \upsilon\,\mathrm{d}t \,+\, s\,\mathrm{d}W(t), \\ S(0) &= S_0. \end{cases} \tag{18}$$

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,

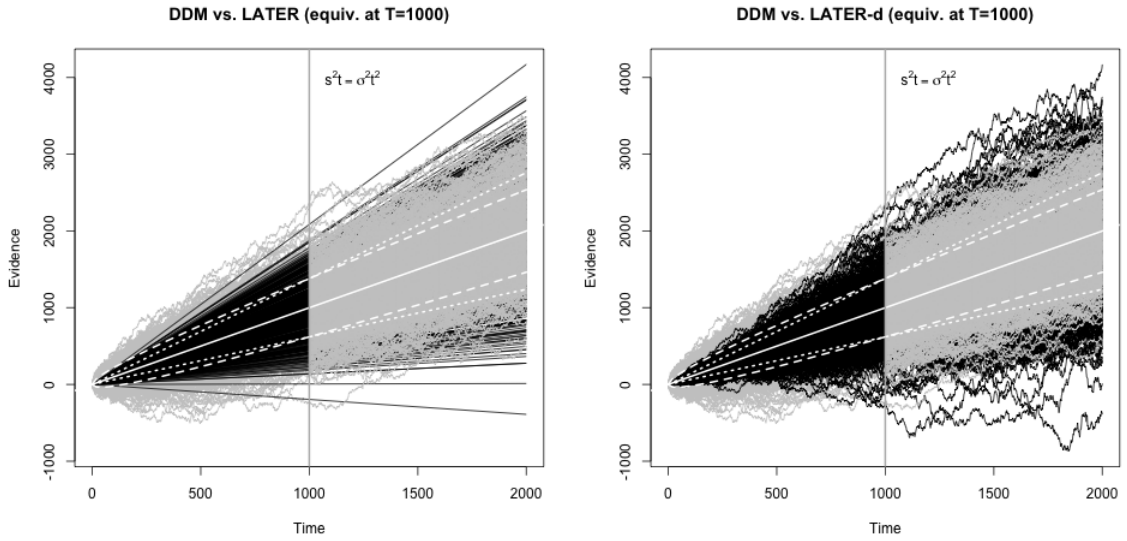

Figure 4. Comparison of LATER and DDM. The left panel overlaps 500 trajectories of the DDM (grey paths; $\upsilon = 1, s = 12.02$), with 500 trajectories of a LATER model (black paths; $r = 1, \sigma_r = .38$). The right panel plots the same DDM trajectories (grey paths), with trajectories sampled from 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.

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 4. The left panel compares 500 trajectories randomly sampled from a DDM with 500 "ballistic" trajectories sampled from a LATER 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 5 illustrates the effects of 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

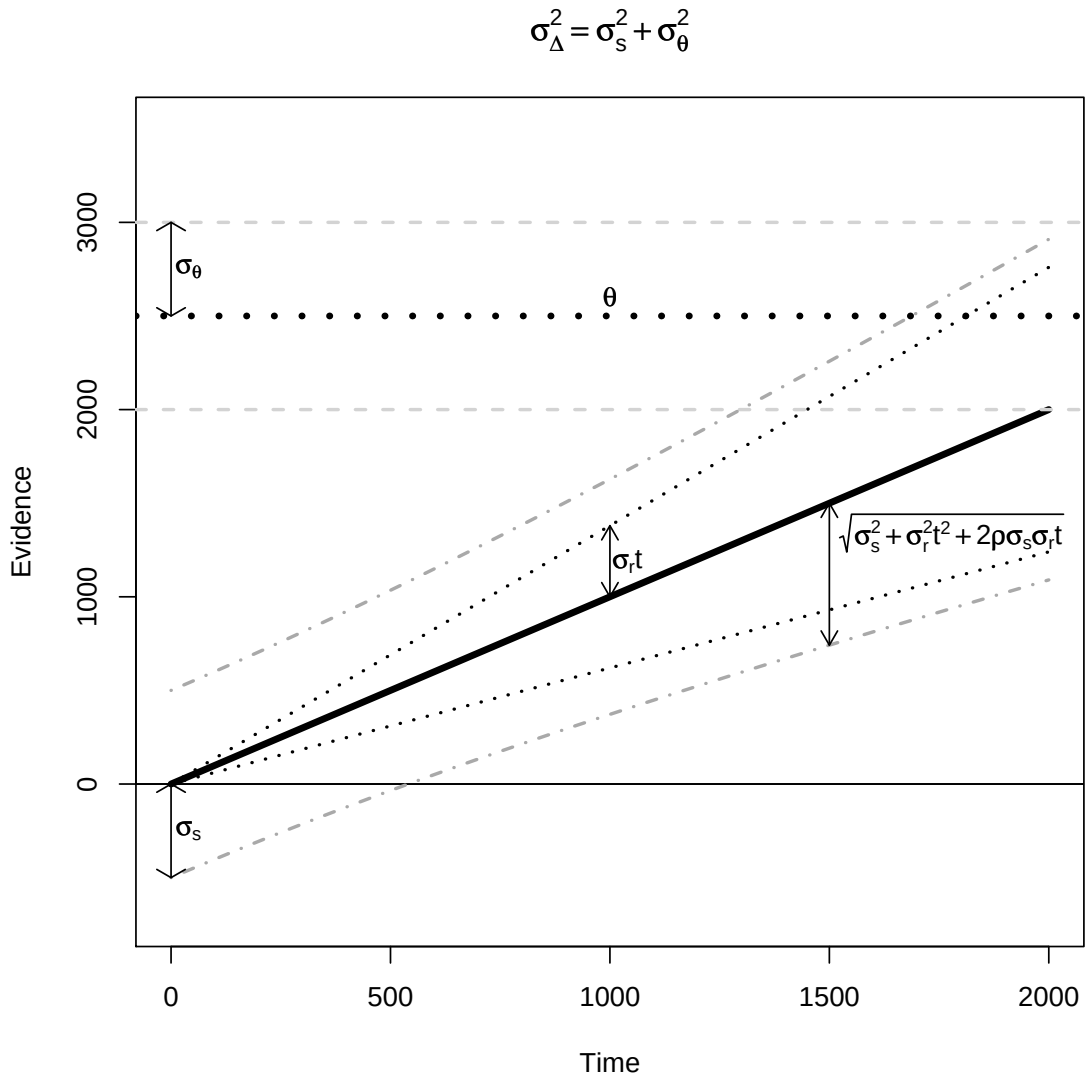

Figure 5. Adding variation in the threshold (σ_θ) and/or starting level (σ_s) to LATER-d.

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	816
word naming (Balota *et al.*, 2007)	English	visual	vocal	recognition	40,481	450
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

(*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 distribution) the accumulated evidence follows a distribution $N(S_0 + rt, \sigma_s^2 + \sigma_r^2 t^2 + 2\rho\sigma_s\sigma_r t)$, 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 5 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 task, 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.

Empirical Evidence

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 it compares to other proposals for RT distributions. 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.

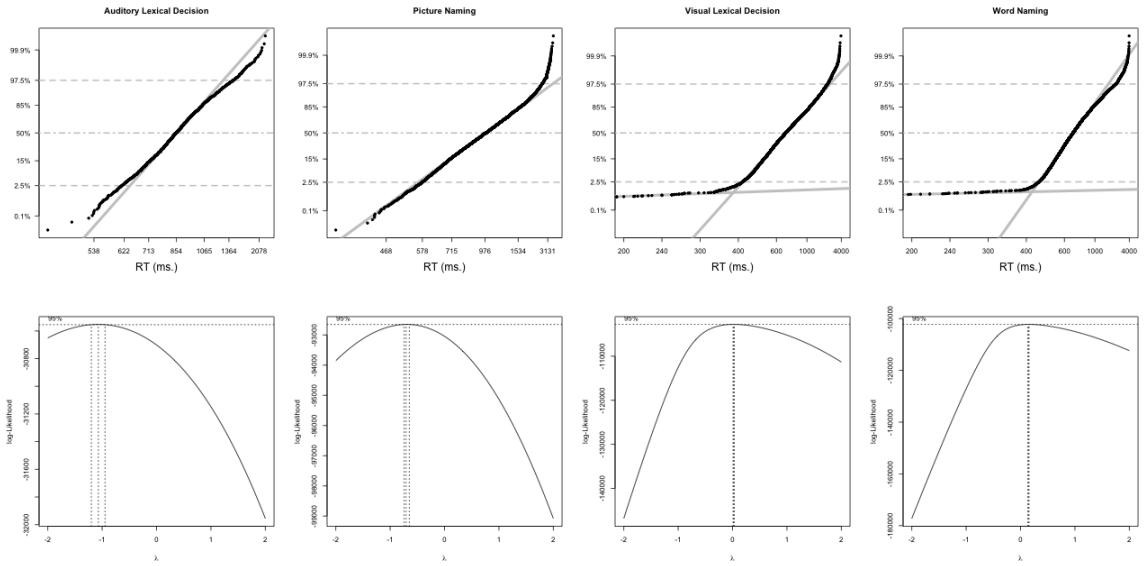

Figure 6. Reciprobit plots (*upper panels*) and log-likelihood (*lower panels*) of the power parameter in the Box-Cox transformation for each of the four individual trial aggregated datasets. The horizontal lines in the Reciprobit plots represent the median and 95% intervals of each dataset. Recinormal 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.

Aggregated 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 the aggregated data should also be described by an instance of Fieller's.

We begin our aggregated analyses by inspecting the Reciprobit plots and the estimated parameter of the Box-Cox power transformation (Box & Cox, 1964) . These are presented in Figure 6 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 (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

Table 4: 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	AIC	43,873	16,374,800	14,287,700	106,242
	BIC	43,886	16,374,824	14,287,724	106,256

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 typically log-normal value of zero. 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 different candidate distributions fit these aggregated data. Table 4 compares the quality of the best fits in terms of Akaike's Information Criterion (AIC) and Schwartz's "Bayesian" Information Criterion (BIC; *cf.*, Liu & Smith, 2009). The table compares fits using the Ex-Wald distribution (*i.e.*, an Inverse Gaussian distribution convoluted with an Exponential to allow for a variable shift; Schwarz, 2001). 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 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 ELP 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-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

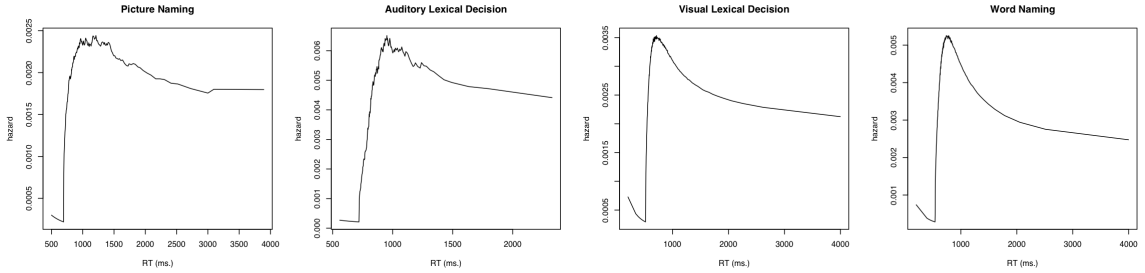

Figure 7. Estimated hazard functions for each of the aggregated datasets.

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 Burbeck & Luce, 1982 and Luce, 1986 for details).

Figure 7 presents the estimated hazard rates (using the method described by Burbeck & Luce, 1982) 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 favoring 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 power-law pattern (*i.e.*, linear in log-log scale), rather than the log-linear decrease that would be predicted by distributions with exponential tails.

Figure 8 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. 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

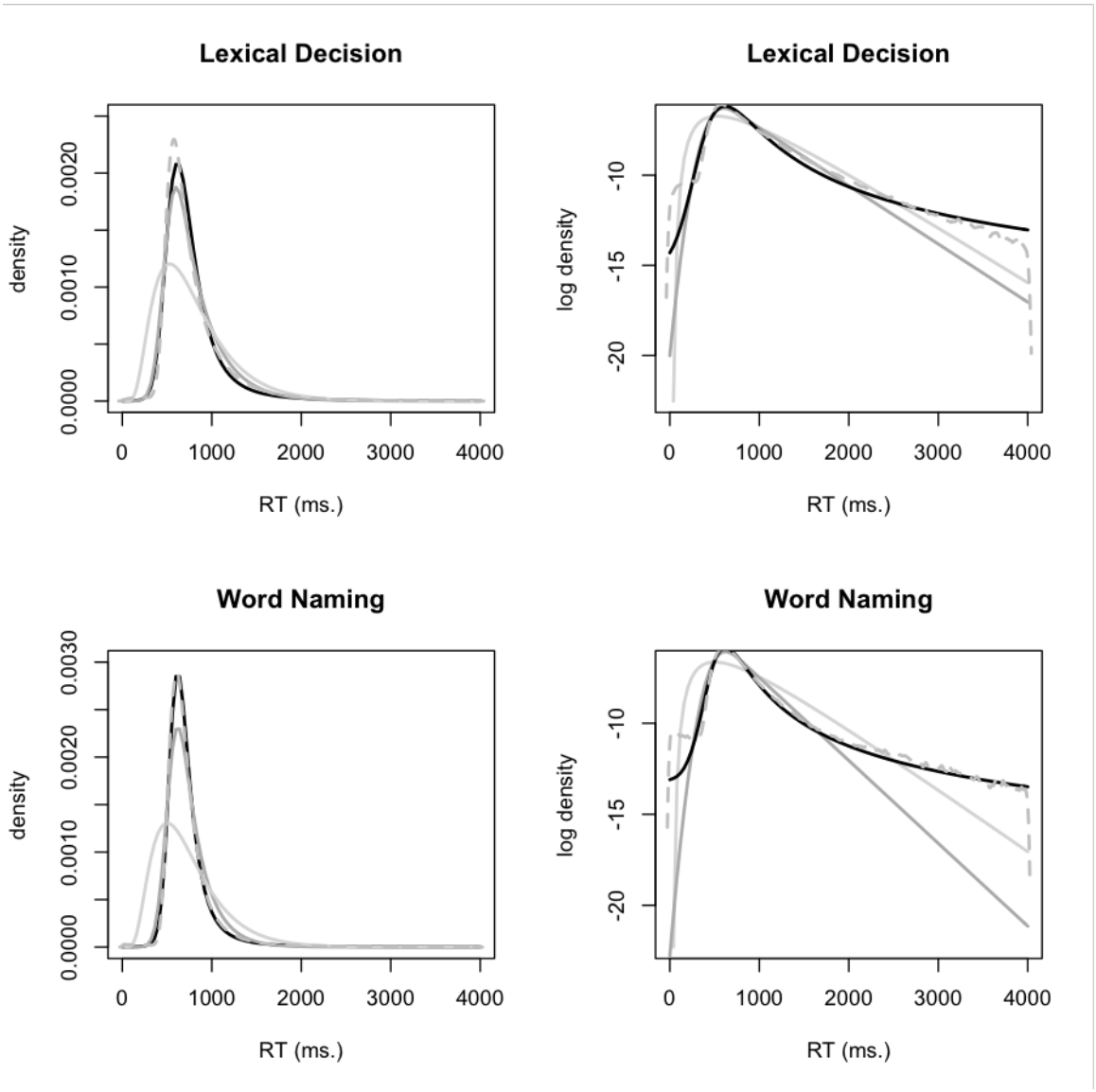

Figure 8. Comparison of the fits provided by Fieller's distribution (black solid lines), the Ex-Gaussian distribution (grey solid lines), and the Ex-Wald distribution (light grey solid lines) to Kernel density estimates (KDE; grey dashed lines) of the aggregated visual lexical decision (*top panels*) 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.

Table 5: 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	$\mathbf{1}9,000 \pm 77$	**1**9,158	$\mathbf{3}0,086 \pm 98$	**3**0,182
	BIC	$\mathbf{1}9,021 \pm 77$	**1**9,179	$\mathbf{3}0,109 \pm 98$	**3**0,205
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	

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 9. 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 implies).

Individual participants analyses

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 participant in the ELP datasets. In these datasets, each participant responded to a relatively large number of words (an average of 1,414 correct responses to words per participant in the lexical decision dataset, and of 2,323 correct responses per participant in the word naming dataset), thus enabling separate fits to each participant. Table 5 summarizes the results

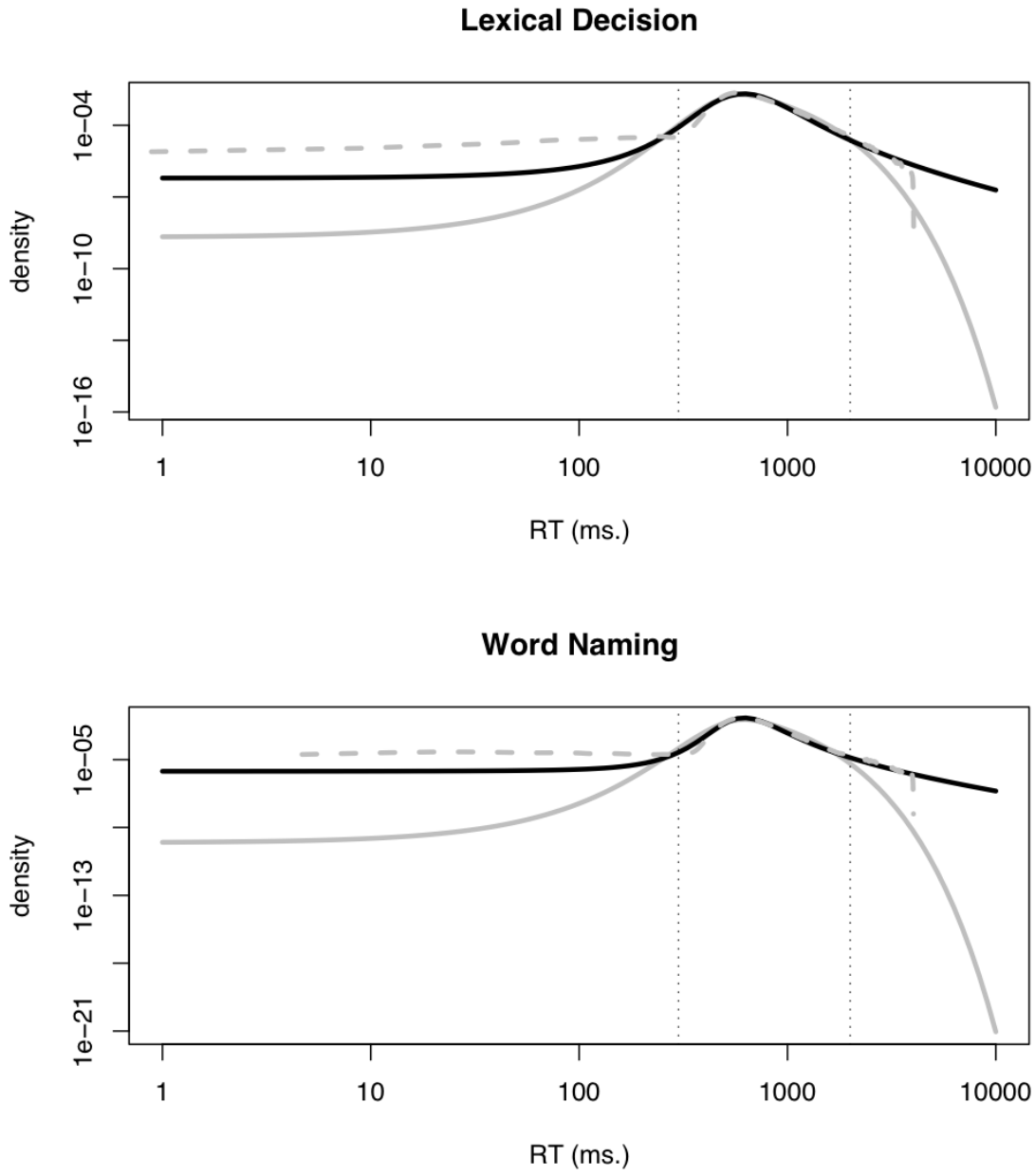

Figure 9. Log-log scale comparison of the fits provided by Fieller's distribution (black solid lines) and the Ex-Gaussian distribution (grey solid lines) to the KDE estimates of the aggregated visual lexical decision (*top panel*) and picture naming (*bottom panel*) latencies from the English lexicon 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.

of fitting distributions individually to each participant. 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 participant-specific differences between the information criteria for both fits. While in the picture naming dataset there was a clear preference for the Fieller's fit in most participants, in the lexical decision datasets there was a huge inter-participant variability on the differences between estimated fits. 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 participant identity, and a possible mixed-effect interaction between the distribution and the participant. In the picture naming data there was a significant advantage for Fieller's fits ($\widehat{\beta} \simeq 265$, $t = 8.4, p < .0001$, $\widehat{p}_{\text{mcmc}} = .0234$) 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 ($\widehat{\beta} \simeq 151$, $t = 2.38$, $p = .0174$, $\widehat{p}_{\text{mcmc}} = .2150$) 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 participant identity and preferred distribution ($\chi^2_{6,2} = 73.55$, $p < .0001$).[5]

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 participants. This is not difficult to understand. The lexical decision datasets included much fewer responses than does the word naming one, and it is thus less likely for a participant to elicit relatively long responses than it is in the larger samples of the word naming dataset. As the main difference between Fieller's distribution and the Ex-Gaussian is found in the heavier right tails, only participants that showed some of the very rare long RTs would be better accounted for by Fieller's. An additional issue that needs to be considered in this respect is that the method for data collection used in the ELP included truncation of the responses at 4,000 msec., thus significantly reducing the information on the right tails (and thus favoring the lighter-tailed distributions such as the Ex-Gaussian).[6]

Figure 10 illustrates the estimated RT distributions of an ideal 'prototypical' partic-

[5]We report both t-based (p) and Markov Chain Montecarlo estimates ($\widehat{p}_{\text{mcmc}}$) 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, Davidson, & Bates, 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.

[6]To perform our analyses we excluded all responses at 4,000 msec. or above, as these corresponded either to measurement or coding errors in the file, or to truncations of slower responses.

ipant in each of the tasks.[7] These are plotted by the solid black lines in the figures. 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 participant 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 11). 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 participants 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.

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 12. The figure displays the estimated values of the λ_1 and λ_2 parameters in the separate individual participant 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

[7] To obtain these curves, we estimated the cumulative density functions of the RTs individually for each participant 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 participant's range of responses. Estimation without extrapolation would have overestimated the densities at the right tail, as these would be estimated only from the participants that produced them, ignoring that most participants in fact did not. This would exaggerate the power-law appearance, biasing in favor of Fieller's distribution. 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.

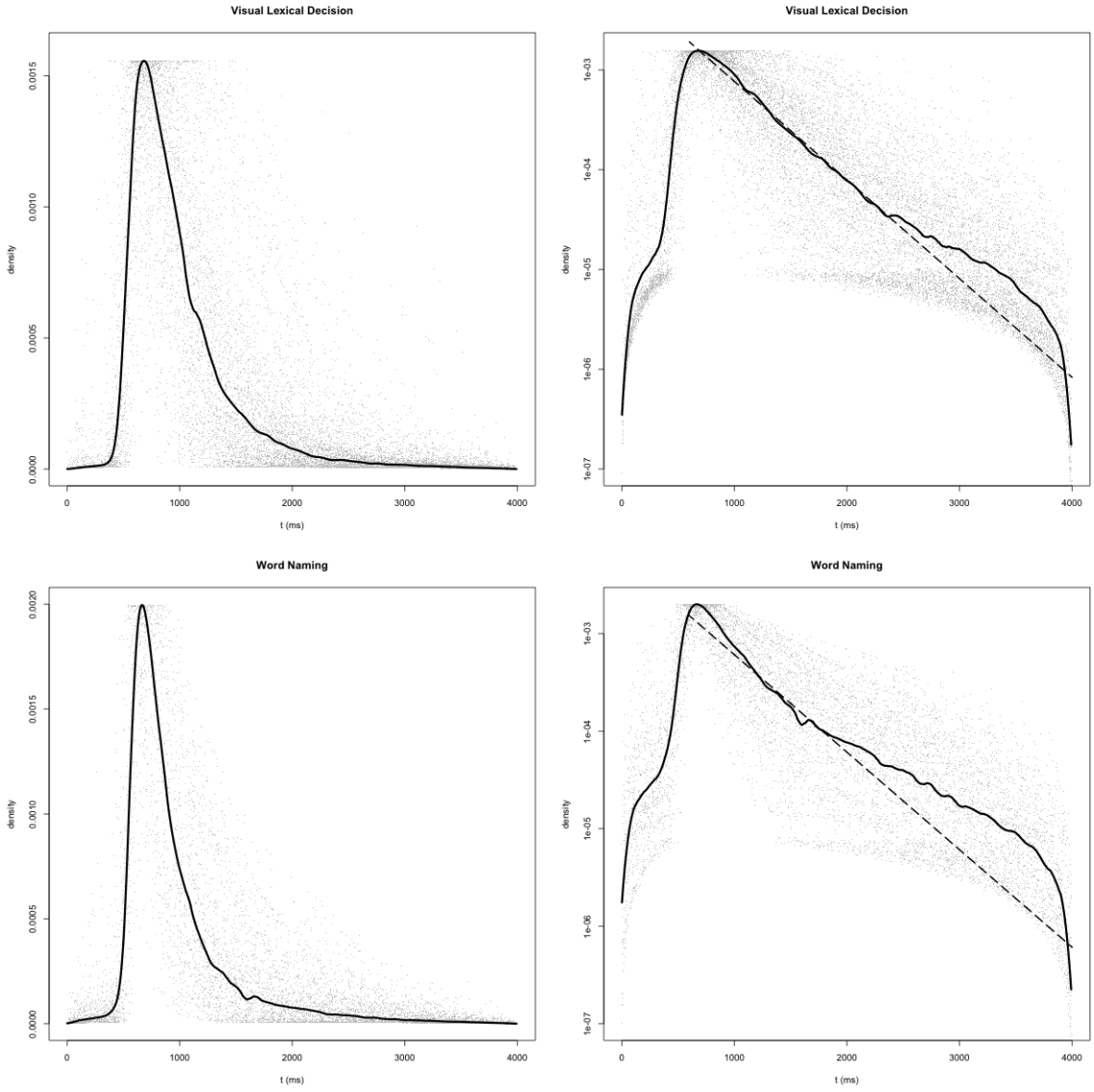

Figure 10. Ideal 'prototypical' participant in the lexical decision (*top panels*) and word naming (*bottom panels*) datasets. The left panels depict the densities, and the right panels are their equivalents in log-scale. The grey points are samples of 50 density points for each participant. The solid 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 what an exponential tailed fit to these data should look like.

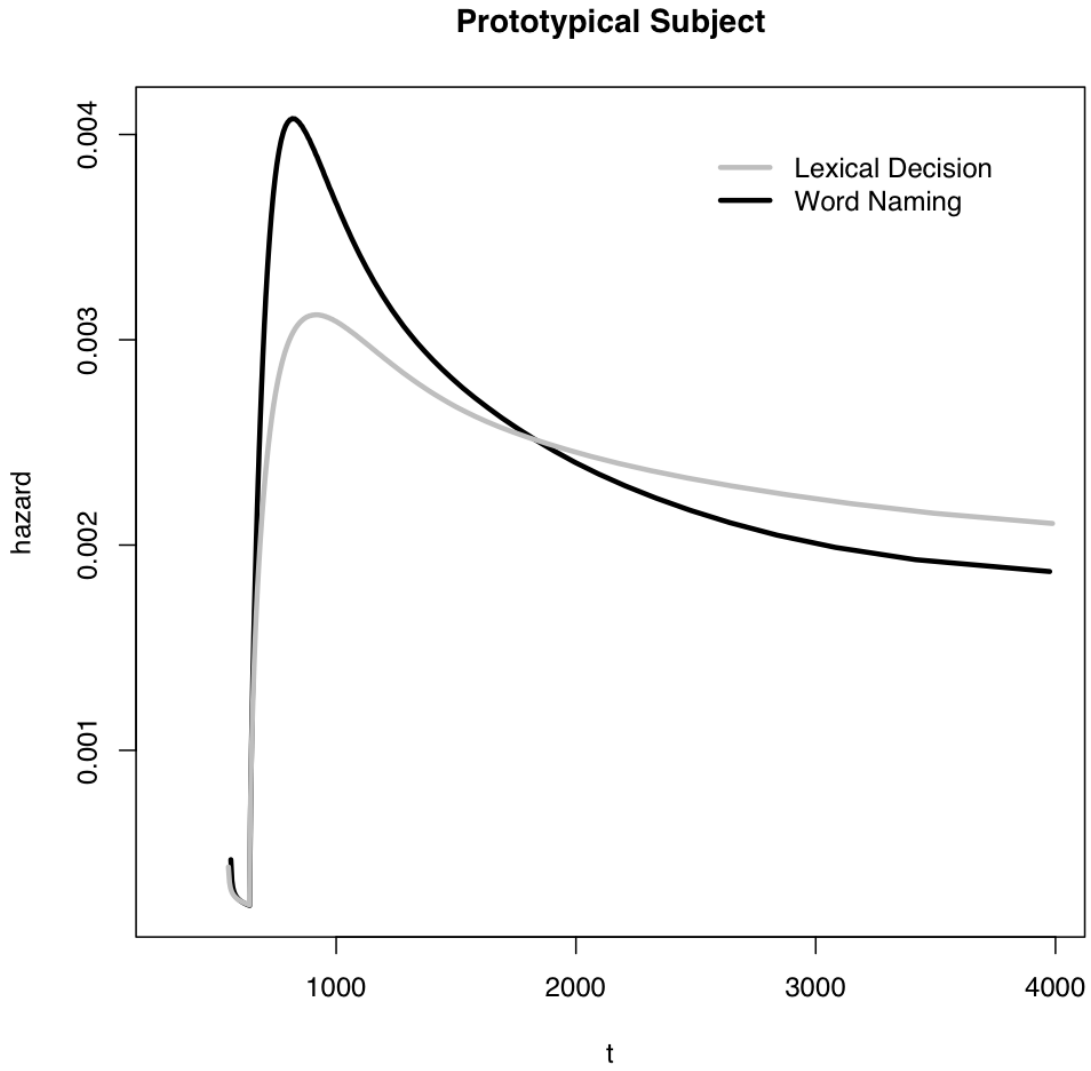

Figure 11. Estimated hazard function for the 'prototypical' participants. The curves where estimated using the non-parametric method described by Burbeck and Luce (1982) on the estimated quantiles of the prototypical distributions.

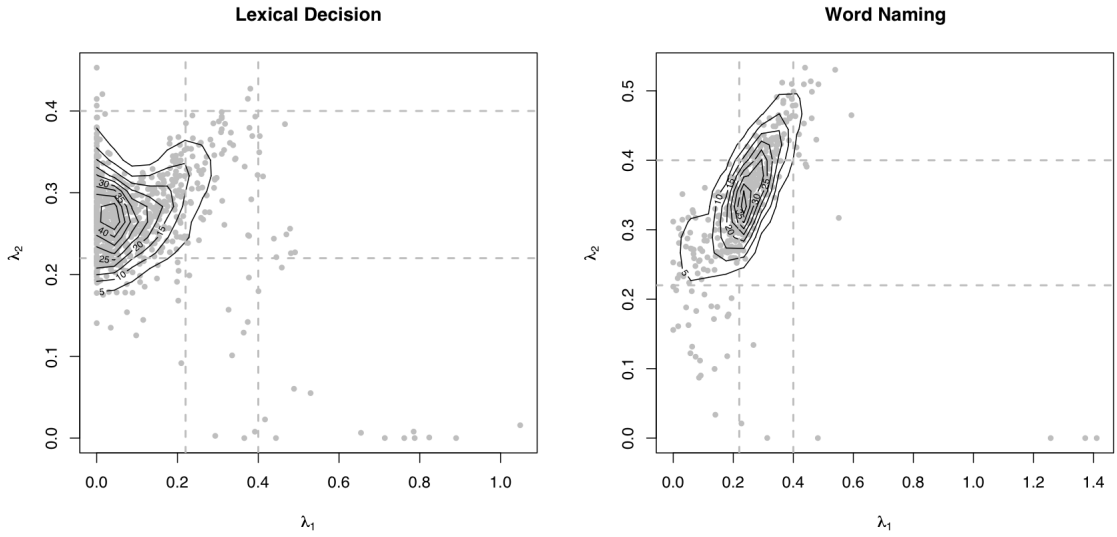

Figure 12. Values of the CoV parameters λ_1 and λ_2 obtained by fitting Fieller's distribution to each individual participant in the visual lexical decision (*left panel*) and word naming datasets (*right panel*) of the ELP. Each point represents the fit obtained for an individual participant. The contours represent a 2-dimensional KDE of the density. 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 value outliers resulting from non-converging fits.

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-participant 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 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 stimuli, and thus there will be variation in both top-down and bottom up properties of the stimuli 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.

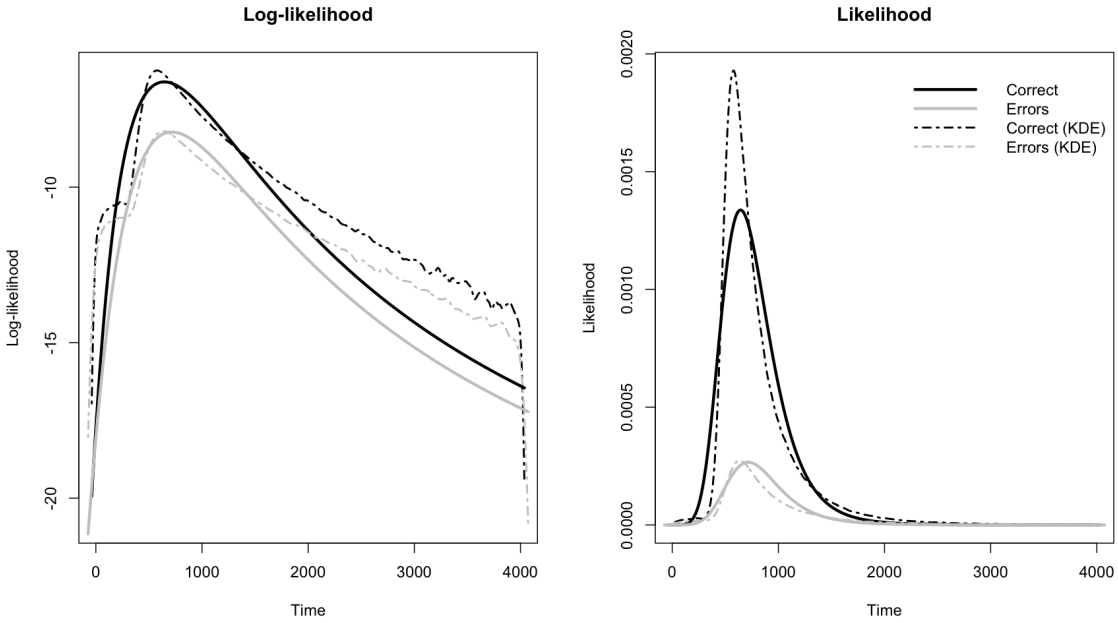

Figure 13. Lexical decision data from the ELP ($\sim$ 1.3 million individual responses) fitted as an inhibition-free competition between accumulators. The solid lines represent the predicted densities (right panel) and log densities (left panel) of the fitted model (the model was fitted with fixed variances 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.

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 better 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. 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. 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 13 illustrates the RT distribution that would be predicted by a race of independent accumulators of the type proposed by Brown and Heathcote (2005, 2008). The right panel shows that a relatively good fit of the density is obtained in comparison with 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 distribution. The

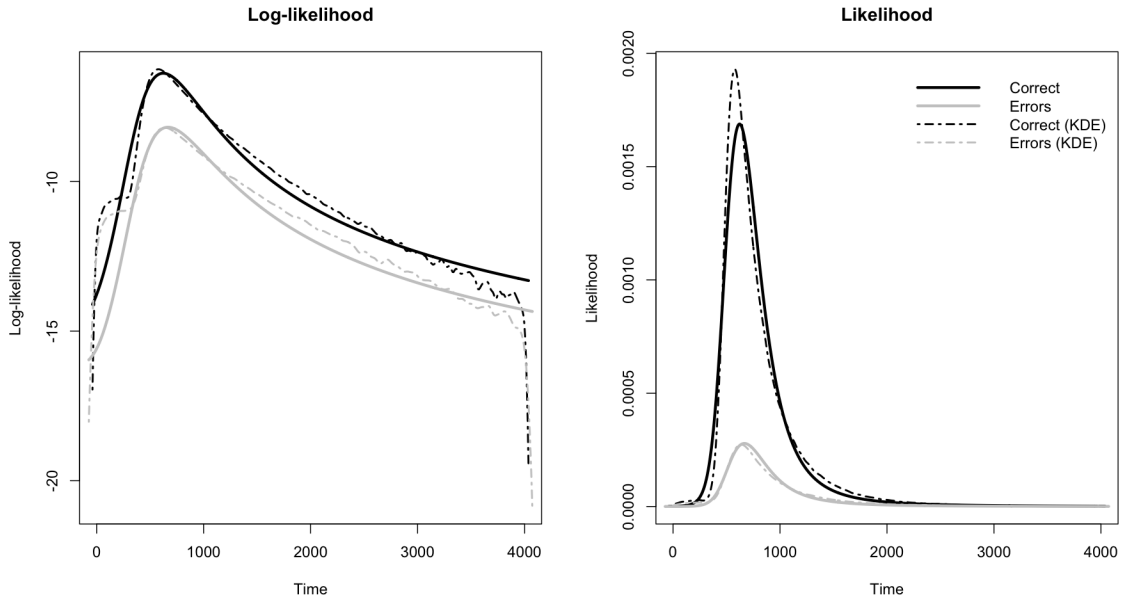

Figure 14. Lexical decision data from the ELP ($\sim$ 1.3 million individual responses) fitted as a competition including inhibition between accumulators. The solid lines represent the predicted densities (right panel) and log densities (left panel) of the fitted model (the model was fitted with fixed 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.

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 14 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 13, 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 8 and Figure 9, neither 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 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 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 15 zooms into the very early visual lexical decision responses of Figure 14. 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 these 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, 2001). This would estimate a non-decisional task component in humans

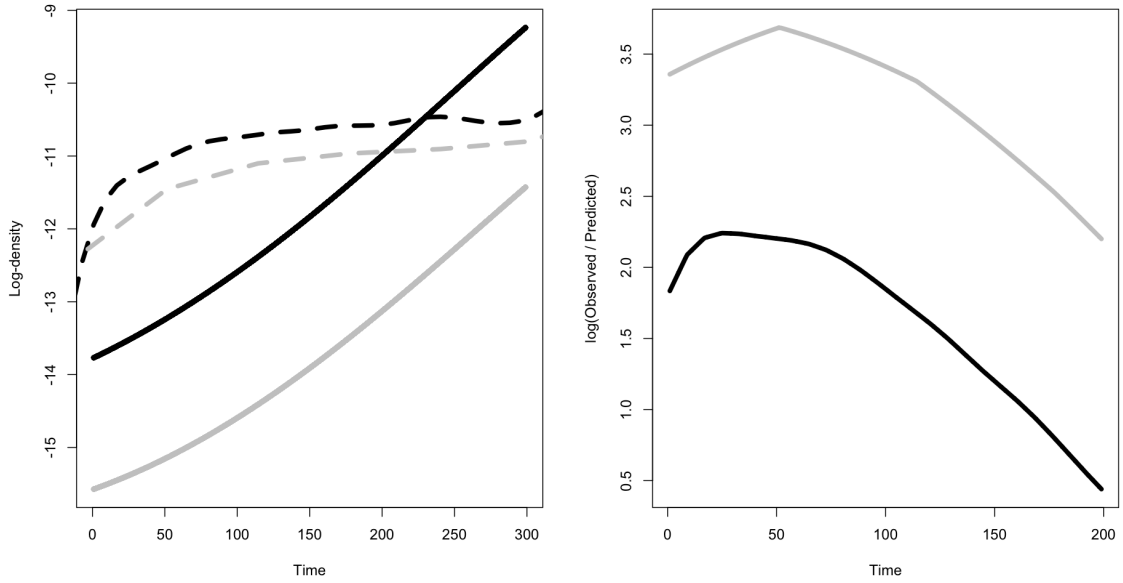

Figure 15. Early visual lexical decision responses.

in the range of 240 ms. to 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 non-decisional times are 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. An additional very small percentage would correspond to the cases where the accumulator was accidentally above the response threshold before the presentation of the stimulus. These would be fully random responses. We can estimate their number at around 576 additional random responses. 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 15 illustrates this. When one compares the log-ratios of observed to predicted short responses, one finds that there is a much more marked increase in errors than in correct ones, and the difference between these log-ratios is constant over time.

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.

Heavy right tails

We have presented evidence in support of an RT distribution with a very heavy right tail. RT distributions, whether individually computed for single participants in a given task and condition, or aggregated across participants 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 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 follow-up studies) whether in exact forms (*e.g.*, Luce, 1986; Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002; Smith, 2000) or in approximate forms (*e.g.*, Lee et al., 2007; Navarro & Fuss, 2008). Although stretched exponential type distribution such as the Weibull (Colonius, 1995; Logan, 1988, 1992, 1995) or the distributions that arise from the 'ballistic' models recently proposed by Brown and Heathcote (2005, 2008) also can give rise to heavy tails, in the data that we have analyzed these seem still too thin.The patterns observed in the data seem more consistent with a power-law – straight in log-log scale – type distribution, such as the one that would be predicted by Holden et al. (2009)'s Cocktail model, or the distributions that are predicted by both LATER (Carpenter, 1981, and follow-up studies) and its generalization as introduced in this study. 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.

Flexible hazard functions

The shapes of the hazard functions of RT distributions (Burbeck & 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. As demonstrated by Holden et al. (2009), the Cocktail model's RT distribution enables this flexibility of hazard functions including both peaked and monotonic types. The distribution that is central to the theory that we are proposing, Fieller's, is also 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

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 5 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 accounted for by the general RT distribution (save for a residual one per thousand). The implication of this is that they are not completely random. As illustrated in Figure 15, we have seen that performance is above chance up to the very early times below 100 ms.

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).

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. In our model, if 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, and this decrease is linear in logarithmic scale, thus naturally accounting for Hick's Law.

Issues of model 'simplicity' and falsifiability

A common argument used to support different models of RT distributions is the relative simplicity of a model against others. In many cases, simplicity in this field is understood as synonimous to 'number of free parameters'. For instance, in their discussion of the LBA model, Brown and Heathcote (2008) argue that their model is simpler than other equally performing theories, since it would require only four free parameters to account for the distribution of responses in a single experimental condition. Likewise, Holden et al. (2009) make a similar argument, their model requires the fitting of six free parameters, in comparison to the seven to nine that would be required to fit a DDM. In this sense, our model is also made of four free parameters, as is the LBA. In our opinion, parameter counting might be a too naïve way of assessing model complexity, and one may need more sophisticated information theoretical tools to make model complexity assessments.

Our theory can in fact be considered a more strictly specified version of Holden and colleagues' Cocktail model. As in theirs, ours exhibits both flexible hazard functions, and power-law behavior from a task dependent point. In addition, the log-normal mixture component that the Cocktail model uses to account for the bulk of the responses, can very well be equivalent to the shape of Fieller's distribution around its mode. Note, that although both models can to a large extent be considered equivalent, in our case we achieve the description of the whole set of latencies using a single function. The values of the CoV and correlation parameters fully determine the shape of the distribution. In addition, while allowing the whole richness of the Cocktail mode, the theory that we have presented also provides a plausible algorithmic level model: a simple rise to threshold mechanism with inhibition linking the accumulators

Furthermore, whereas the the scaling parameter of the power-law in the right tail of

the Cocktail model is a free parameter, our model makes a stricter prediction (and thus falsifiable in the Popperian sense) that the value of this scaling parameter should be exactly two. This precise value is a keystone of ratio distributions: They cannot give rise to any other value. Finding that any particular experimental task or condition elicits a power-law right tail whose scaling parameter is reliably different than two would lead to the complete rejection of our theory. Thus, not only does the theory account for RT distributions but, perhaps more importantly, it also predicts what types of RT distributions are altogether impossible. This is not a common feature of all theories of RT distributions.

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.

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 forcefully argued by Jaynes (2003, p. 133), 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 decisional component, the later would describe the RTs in processes that are dominated by recognition 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.

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), but note that several other mechanisms, including ratio distributions such as Fieller's, can give rise to power-laws without any explicit need for self-organization (Newman, 2005; Sornette, 2001). SOC systems are complex systems, 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 $1/f$ 'pink' noise patterns that are indicative of a complex SOC system (*cf.*, Buzsáki & Draguhn, 2004). Furthermore $1/f$ noise characteristics have also been reported for RT distributions by some researchers (Gilden, 1997, 2001; Gilden, Thornton, & Mallon, 1995; Thornton & Gilden, 2005, 2007; Van Orden, Holden, & Turvey, 2003, 2005; but see also Farrell, Wagenmakers, & Ratcliff, 2006; Wagenmakers, Farrell, & Ratcliff, 2004, 2005; Wagenmakers, Grünwald, & Steyvers, 2006 for views questioning the evidence for $1/f$ noise patterns). 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, Ratcliff, 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 5.

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), Mandelbrot (1983), and Newman (2005), 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.

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*, 335-344.

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)

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., & 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.

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*, 16551670.

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

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., & McDonald, S. A. (2007). LATER predicts saccade latency distributions in reading. *Experimental Brain Research*, *177*, 176-183.

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

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

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

Farrell, S., Wagenmakers, E.-J., & Ratcliff, R. (2006). $1/f$ noise in human cognition: Is it ubiquitous, and what does it mean? *Psychonomic Bulletin & Review*, *13*, 737–741.

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.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, *8*, 296–301.

Gilden, D. L. (2001). Cognitive emissions of $1/f$ noise. *Psychological Review*, *108*, 33–56.

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

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.

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.

Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological Review*. (in press)

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.

Johnston, K., & Everling, S. (2008). Neurophysiology and neuroanatomy of reflexive and voluntary saccades in non-human primates. *Brain and Cognition*, *68*, 271-283.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795. Available from `http://citeseer.ist.psu.edu/539880.html`

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.

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.

Liu, C. C., & Smith, P. L. (2009). Comparing time-accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin & Review*, *16*, 190-203.

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.

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

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

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.

Montagnini, A., & Chelazzi, L. (2005). The urgency to look: Prompt saccades to the benefit of perception. *Vision Research*, *45*, 3391-3401.

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.

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

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience*, *26*, 8965-8975.

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

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

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*, 159182.

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*, *106*, 261-300.

Reddi, B. A. J., Asrress, K. N., & Carpenter, R. (2003). Accuracy, information, and response time in a saccadic decision task. *Journal of Neurophysiology*, *90*, 3538-3546. Available from `http://jn.physiology.org/cgi/content/abstract/90/5/3538`

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

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*. Available from `http://dx.doi.org/10.1371%2Fjournal.pcbi.1000046`

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

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.

Sornette, D. (2001). Mechanism for powerlaws without self-organization. *International Journal of Modern Physics C*, *13*, 133-136.

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

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.

Thornton, T. L., & Gilden, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, *12*(3), 409–41.

Thornton, T. L., & Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review*, *114*, 71–103.

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

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*, 550592.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*, 331–50.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and $1/f$ scaling. *Journal of Experimental Psychology: General*, *134*, 117–23.

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

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f$ noise in human cognition. *Psychonomic Bulletin & Review*, *11*, 579615.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2005). Human cognition and a pile of sand: A discussion on serial correlations and Self-Organized criticality. *Journal of Experimental Psychology: General*, *134*, 108–116.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

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.

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.

Appendix A
The Recinormal distribution

We define the Recinormal distribution as the distribution of a random variable X whose reciprocal $Y = 1/X$ 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:

$$f_r(x|\mu,\sigma) = \phi(y|\mu,\sigma^2)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = \phi\left(\frac{1}{x}\middle|\,\mu,\sigma^2\right)\frac{1}{x^2}. \tag{19}$$

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

$$f_r(x|\mu,\sigma) = \begin{cases} \frac{1}{x^2\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(1-\mu x)^2}{2\sigma^2x^2}} & \text{if } x \neq 0. \\ \\ 0 & \text{if } x = 0, \end{cases}, \tag{20}$$

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

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 given by the probability density function (Fieller, 1932; Hinkley, 1969):

$$f(w) = \frac{b(w)d(w)}{\sigma_1\sigma_2a^3(w)\sqrt{2\pi}}\left[\Phi\left(\frac{b(w)}{a(w)\sqrt{1-\rho^2}}\right) - \Phi\left(-\frac{b(w)}{a(w)\sqrt{1-\rho^2}}\right)\right] + \frac{\sqrt{1-\rho^2}}{\pi\sigma_1\sigma_2a^2(w)}\mathrm{e}^{-\frac{c}{2(1-\rho^2)}}, \tag{21}$$

where

$$\begin{aligned} a(w) &= \sqrt{\frac{w^2}{\sigma_1^2} - \frac{2\rho w}{\sigma_1\sigma_2} + \frac{1}{\sigma_2^2}}, \\ b(w) &= \frac{\theta_1 w}{\sigma_1^2} - \frac{\rho(\theta_1+\theta_2 w)}{\sigma_1\sigma_2} + \frac{\theta_2}{\sigma_2^2}, \\ c &= \frac{\theta_1^2}{\sigma_1^2} - \frac{2\rho\theta_1\theta_2}{\sigma_1\sigma_2} + \frac{\theta_2^2}{\sigma_2^2}, \\ d(w) &= \mathrm{e}^{\frac{b^2(w)-ca^2(w)}{2(1-\rho^2)a^2(w)}}, \end{aligned} \tag{22}$$

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

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:

$$\begin{aligned} \kappa &= \frac{\theta_1}{\theta_2}, \\ \lambda_1 &= \frac{\sigma_1}{|\theta_1|}, \\ \lambda_2 &= \frac{\sigma_2}{|\theta_2|}, \\ -1 < \rho &< 1. \end{aligned} \tag{23}$$