The Shifted Wald Distribution for Response Time Data Analysis

Royce Anders, F. -Xavier Alario, Leendert van Maanen

Psychological Methods
2014, Vol. xx, No. x, xx–xx

xxxx-xxxx/xx/$xx.xx DOI: xx.xxxx/xxxxxxxx

The Shifted Wald Distribution for Response Time Data Analysis

Royce Anders and F.-Xavier Alario
LPC UMR 7290
Aix Marseille Université, CNRS

Leendert Van Maanen
Department of Psychology
University of Amsterdam

We propose and demonstrate the shifted Wald (SW) distribution as both a useful measurement tool and intra-individual process model for psychological response time (RT) data. Furthermore, we develop a methodology and fitting approach that readers can easily access. As a measurement tool, the SW provides a detailed quantification of the RT data that is more sophisticated than mean and standard deviation comparisons. As an intra-individual process model, the SW provides a cognitive model for the response process in terms of signal accumulation and the threshold needed to respond. The details and importance of both of these features are developed, and we show how the approach can be easily generalized to a variety of experimental domains. The versatility and usefulness of the approach is demonstrated on three published data sets, each with a different canonical mode of responding: manual, vocal, and oculomotor modes. In addition, model-fitting code is included with the paper.

Keywords: shifted Wald, inverse Gaussian, psychometrics, sequential sampling, evidence accumulation

Introduction

From the earliest inceptions in formal psychology by fundamental figures, such as Wilhelm Wundt and Franciscus Donders, to present day, response time (RT) recording and analysis have continued to remain a major approach in psychological research. Despite greater technologies to also measure correlates of brain activity, such as electroencephalography and magnetic resonance imaging, RT data continues to deliver important information, either side-by-side with the newer data, or as standalone data. Whilst there have been such developments with new measurement technology, also have come developments within the data analysis techniques. A central development to more sophisticated data analysis, is the current movement to consider experimental effects on the full set of observations recorded, rather than on only the central tendencies. Such a movement for RT data, has a strong case rooted within Luce's (1986) extensive mathematical work. The present paper aims to develop further understandings, competencies, and methodologies for the researcher who handles RT data; particularly, we focus on a powerful distributional analysis tool that is not yet in general practice within the psychological community, the shifted Wald model.

We acknowledge funding by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant agreement n° 263575), and the Brain and Language Research Institute (Aix-Marseille Université : A*MIDEX grant ANR-11-IDEX-0001-02 and LABEX grant ANR-11-LABX-0036). We thank the "Féderation de Recherche 3C" (Aix-Marseille Université) for institutional support, as well as Joël Fagot, Mijke Hartendorp, and Françoise Vitu, for making available their data.

The corresponding author, Royce Anders, may be contacted at the address: LPC, Aix-Marseille Université; UMR 7290 Pôle 3 C, Bâtiment 9 Case D; 3 place Victor Hugo; 13331, Marseille Cedex 3, France, and by email: royce.anders@univ-amu.fr.

Distributional RT Analysis

The efficacy of modeling the distributions of RT data to obtain a deeper understanding of experimental effects and underlying processes, rather than only using classical analysis methods, has been well-demonstrated in preceding psychological science literature (Andrews & Heathcote, 2001; Balota & Yap, 2011; Balota, Yap, Cortese, & Watson, 2008; Heathcote, 2004; Laming, 1968; Link, 1992; Luce, 1986; Ratcliff, 1978; Ratcliff, Gomez, & McKoon, 2004; Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999; Staub, White, Drieghe, Hollway, & Rayner, 2010; Stone, 1960; Van Maanen, Grasman, Forstmann, & Wagenmakers, 2012; Van Zandt, 2000, 2002).

There exist quantitative distribution measurement tools for RT data, in which the parameters describe the properties of the observed data distribution. These tools are typically closed-form probability density functions with positive skew and values, such as the shifted Wald (Folks and Chhikara 1978; see Chapter 8.2, Luce 1986; Ricciardi 1977; Wald 1947), ex-Gaussian (Burbeck & Luce, 1982; Heathcote, Popiel, & Mewhort, 1991; Hohle, 1965; Jeansonne & Foley, 1991), Weibull (Fréchet, 1927; Weibull, 1951), log-

normal (Crow & Shimizu, 1988), gamma (Lukacs, 1955) and Gumbel (Gumbel & Lieblein, 1954). Due to their direct quantification and description of the RT distribution, and their straightforward application to data, these techniques are often called *measurement models*.

Then there are more complicated models of RT data that aim to model a process that underlies the data, by modeling signal accumulation, such as the LATER (Carpenter, 1981) and E-LATER models (Nakahara, Nakamura, & Hikosaka, 2006), the Linear Ballistic Accumulator (LBA, Brown & Heathcote, 2008), the race model (LaBerge, 1962), and the Drift Diffusion Model (DDM, Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Murdock, 1976). However, these model parameters do not directly describe the distribution of RT data. While the full RT distribution is used to fit these models, the parameters rather indirectly quantify the RT distribution by instead describing the data in the context of an intra-individual process, in which a signal accumulates over time, toward a threshold that must be reached in order to respond. In doing so, these kinds of models have been termed an extension of signal detection theory (Green & Swets, 1966) to the time domain in psychology (p. 268, by Ratcliff et al. 1999; see also Pike 1973). Though due to the greater complexity of how these models describe the data, they are more often difficult to fit to data than simple measurement models, and are rather known as *process models*.

We bring to attention that there is an interesting model of those previously-mentioned that is simultaneously both a measurement and intra-individual process model in the ways just described: the shifted Wald (SW) model. The SW is a simple and concise model that may provide notable advantages in RT distribution analysis. However, the SW is not yet routinely considered a potential option in the psychological community, and so we aim to advance knowledge of this unfamiliar model. Particularly, it appears that understanding of the following aspects is not well-developed in the psychological community. Firstly, that the SW is a strong statistical tool in which the parameters directly quantify the three important aspects needed to fully describe an RT distribution: its onset, deviation around the mode, and tail magnitude. Secondly, that the SW simultaneously provides a simple activation accumulation model for the observed response in a given paradigm. Furthermore, the SW can be more useful when it is specified per design cell in an experiment; and that in this way, the model can be easily generalized to a number of different experimental domains. In the course of the paper, these accounts are clarified and demonstrated.

The paper is organized into a number of sections that work to provide a concise tutorial and methodology for the SW approach. In doing so, there are also demonstrations of the model and the proposed fitting method on simulated and real data, as well as explanations for when one might consider applying the model. The SW is also further discussed in relation to alternative distribution models for RT analysis. Finally, code is provided as a supplementary file to easily apply the model to data. The SW is a strong statistical tool, and a simple process model of aggregate response activation; however its simple process model account does not replace more complex process models that can be validated on RT data, but it can rather serve to provide an aggregate summary of a more complex underlying process, that more complex models might aim to further quantify (e.g., see Zandbelt, Purcell, Palmeri, Logan, & Schall, 2014).

The Shifted Wald

As mentioned previously, the SW has two major forms in which it may be used: as a simple cognitive process model, or alternatively, as a distributional measurement tool. The process model form of the SW shares the same accumulation model (AM) process that is at the heart of other popular AMs, such as the DDM, race, LATER, and E-LATER/LBA models. The distributional measurement form of the SW is simply given by its probability density function (pdf) for RTs, such as in other popular distribution measurement pdfs: for example the ex-Gaussian, Weibull, lognormal, gamma, and Gumbel. The following two sections describe these two principal usages of the SW.

As a Distribution Measurement Tool

The SW is characterized by a probability density function that can be applied to any positively-valued data with a degree of right skew. It is well known that observed RT data from psychological experiments typically involve this form, as depicted in the left plot of Figure 1. Important observations can be gleaned from the plot: firstly, it is clear that the mode, median, and mean of such distributions are not at the same value, and in addition, the sample mean and standard deviation (s.d.) are over-estimated by values in the right tail of the distribution. Furthermore, the s.d. cannot describe the shape of the distribution around any of the central tendency measures (mean, median, mode); and the mean does not provide an indicative location (onset) of the distribution.

Instead, it is useful to distinguish that regular RT distributions are fully identified by three specific quantifiers, and usually not by two general ones. The three standard pieces of information needed are (*i.*) an onset of the distribution after an initial empty interval (e.g. 0-300 ms) where the respondent cannot appropriately complete the task this quickly; (*ii.*) a central tendency area with deviation centrally around the mode value (e.g. 350-750 ms), this is where most RTs lie; and (*iii.*) the long tail area in which the slower RTs occupy (e.g. 800-1300 ms). RT distributions are hence more complex than Gaussian (or non-skewed distributions), and their shapes cannot be appropriately identified by only a mean value and standard deviation. This is because multiple combinations of (*i.*), (*ii.*), and (*iii.*) can all produce the same

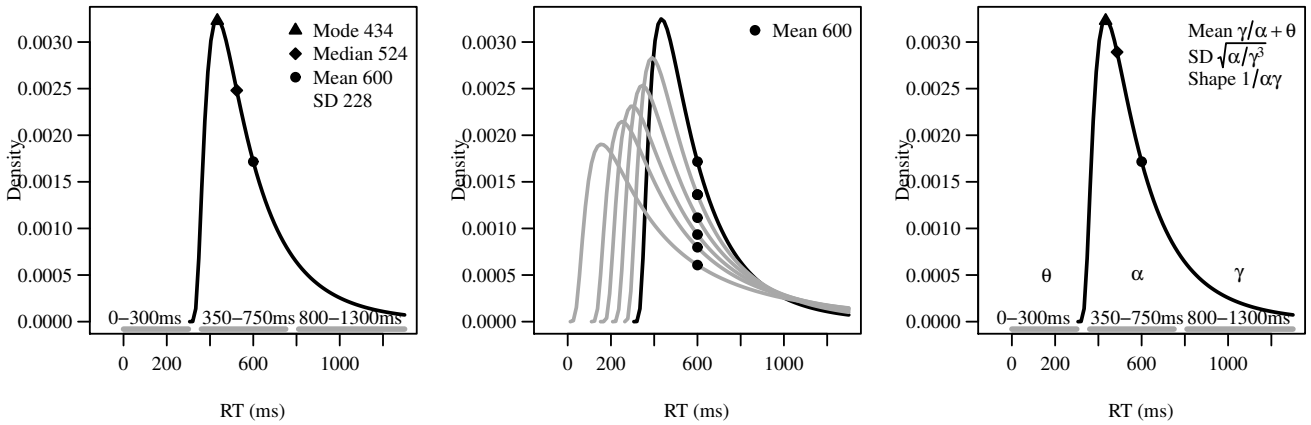

Figure 1. Illustration of RT distribution trends. Left, a standard RT distribution which has an onset, a central tendency area, and a tail thickness; middle, many RT distributions with very different shapes and onsets, but with all sharing the same mean; right illustration of the SW that has a parameter to account for the three RT distribution characteristics.

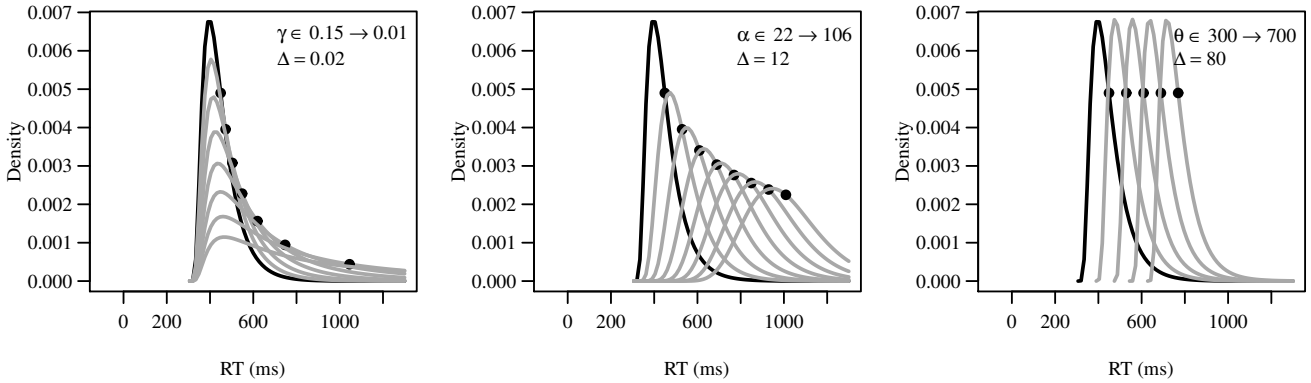

Figure 2. An illustration of how the SW distribution shape changes as one manipulates each parameter. Left to right, changes in γ, α, and θ, each in a direction that produces larger mean values. In the left plot, the black distribution starts with $\gamma = 0.15$ and each successive grey distribution is a reduction of 0.02 units, until γ reaches 0.01.

mean, and/or the same standard deviation. This is easy to show, such as in the middle plot of Figure 1, where the distributions despite having strongly different onsets and shapes, all have the same mean.

Therefore, as a resolution to the disadvantages of the mean and s.d. as RT metrics (see also Balota & Yap, 2011; Balota et al., 2008; Rouder, 2005), one can instead analyze RT data more sophisticatedly with a three-parameter account of the distribution, that directly identifies and describes the RT distribution aspects: (*i.*), (*ii.*), and (*iii.*), and how they change with experimental factor effects. Furthermore, one can calculate the mean and s.d. with these parameters, making them less-likely to be over-estimated in a sample of RTs. These are the primary advantages of the SW as a distribution measurement tool, and these quantifiers are illustrated according to the RT distribution in the right plot of Figure 1.

The SW can serve as such a distribution measurement tool by use of its standard probability density function (pdf),

$$f(X \mid \gamma, \alpha, \theta) = \frac{\alpha}{\sqrt{2\pi(X-\theta)^3}} \cdot \exp\left\{-\frac{[\alpha - \gamma(X-\theta)]^2}{2(X-\theta)}\right\}, \tag{1}$$

with expected value $\alpha/\gamma + \theta$, and variance α/γ^3, for $X > \theta$. The RT mode can also be calculated parametrically, as

$$\text{Mode}(X) = \frac{\alpha}{\gamma}\left[\left(1 + \frac{9}{4\alpha^2\gamma^2}\right)^{1/2} - \frac{3}{2\gamma}\right]. \tag{2}$$

Thus the SW describes a unimodal distribution.

The distributional effect that occurs by changing each of the following parameters, γ, α, and θ is illustrated in Figure 2. As shown in the figure, changes in γ affect mass in

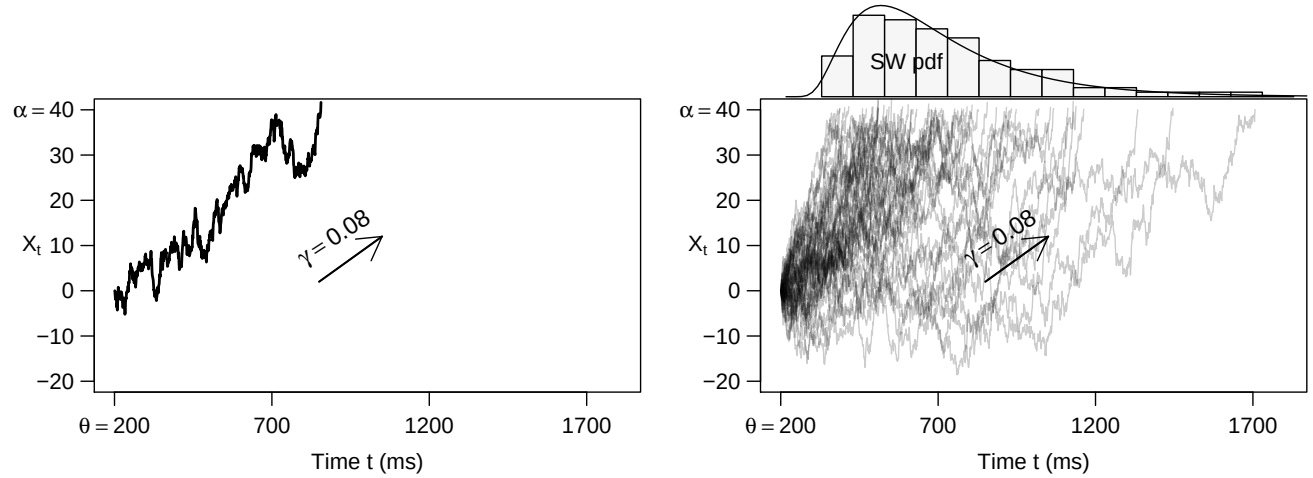

Figure 3. The SW as a cognitive-behavioral model, describing the RT data in the context of a latent quantity (e.g. signal) accumulating to threshold, α, at rate, γ, where θ accounts for the time lapsed outside of (around) this process. Left, a single trial is modeled with the parameters. Right, many trials (e.g. an experimental design cell) are modeled with the same parameter values, and these ultimately form a SW distribution shaped with the same signal accumulation parameters.

the tail; changes in α affect deviation around the mode, and determine normality of the distribution; and changes in θ determine the onset of the distribution (location), but not the distribution shape (distribution of mass). Furthermore, one can notice that the black points in the plots of Figure 2 indicate the mean of each distribution, and it is quantitatively clear that any of these three distribution shape or location effects, will likewise have an effect on the mean RT.

The advantage of the SW as a distribution measurement tool is indeed its ability to quantify the full RT distribution, and parse the data to further locate the specific effect of an experimental manipulation on the RT distribution. For instance, is an experiment effect solely on (*i.*), (*ii.*), or (*iii.*), or a combination of them? The SW has a three-parameter *decomposition of the central tendency*, in which as noted before with the pdf, the

$$E(X) = \alpha/\gamma + \theta \tag{3}$$

$$\mathrm{SD}(X) = \sqrt{\alpha/\gamma^3}\,. \tag{4}$$

Principally in quantifying the mean RT value by (3) distributionally, the SW can for example, detect when a mean-difference in RTs by an experimental manipulation, is specifically explained by one or two kinds of shaping or location effects on the RT distribution. Furthermore, it is possible that shaping or location effects may counteract each other in (3), and hence produce the same means as illustrated in the middle plot of Figure 1, despite the distributions having strikingly different shapes and locations. The SW can detect this pattern and therefore better protect against making a Type II error (erroneously accepting the null hypothesis) in statistical analyses of RT data. This is because the SW can detect significant differences between the RT distribution forms of condition levels, in spite of when these forms may compensate in order to result in insignificantly different mean RT values.

Other Parameterizations. There is also an alternative three-parameter description of the SW using parameter names μ, λ, τ rather than γ, α, θ, that is worth clarifying to defuse possible misunderstandings. Here, the three parameters instead describe the distribution by its pre-shifted mean, μ, and tail length, λ; the third parameter describes the shift equivalently as θ, but typically has name τ. It relates to the previous parameterization as follows:

$$\begin{aligned} \mu &= \alpha/\gamma \\ \lambda &= \alpha^2 \\ \tau &= \theta\,. \end{aligned} \tag{5}$$

This parameterization of the SW on the left, is more often referred to as the shifted inverse Gaussian (IG) distribution, or three-parameter IG; although it is exactly the equivalent distribution.

Also of note, is another parameter combination that may be used to compare distribution results. We assign

$$\beta = \mu/\lambda = 1/\alpha\gamma, \tag{6}$$

to denote the parameter combination, which strictly describes the SW distribution's shape; β will be involved later in our estimation approach.

Other Distribution Measurement Models. As mentioned previously, the SW is among a number of other measurement models, such as the ex-Gaussian, lognormal,

Weibull, gamma, and Gumbel, that can also serve to improve the sophistication of RT analysis; they each have different parameterizations that describe the distribution shape more specifically, and can also be improvements over simple mean and standard deviation comparisons. The current work does not aim to provide an extensive comparison exercise between all six distributions, and prior work has found that other distributions, such as the lognormal, Weibull, and gamma (with a shift added, see Rouder 2005) may fit similarly well to the same kinds of positively skewed data (Folks and Chhikara 1978; see also Palmer, Horowitz, Torralba, and Wolfe 2011); and rather the fundamental difference that may exist between such models is simply in the way one describes the data with the particular distribution's parameter meanings.

Indeed the principal motivation for the SW focus herein is based on the fact that it is the only distribution with parameters that also describe the RT data in the context of an activity accumulation process model; and secondly, because the three SW parameters provide a complete and clear interpretation of the RT data: specifically in terms of distribution (*i.*) onset, (*ii.*) central deviation around the mode / normality, and (*iii.*) tail thickness, which are aspects importantly recommended for RT data analysis by Rouder (2005). One can note that some of the previously-mentioned distributions' parameters do not provide a direct interpretation of such aspects like (*i.*) and (*ii.*). Also some of them are less consistent to fit with maximum-likelihood estimation (MLE) methods, e.g. the lognormal, Weibull, and gamma, since the likelihood is unbounded (Cheng & Amin, 1981; Koutrouvelis, Canavos, & Meintanis, 2005). However fortunately for the SW, the likelihood is bounded and it can thus provide more consistent parameter values when fit to data in this way. That is in being bounded, the maximum value of the SW likelihood function (1), for a given $\theta > 0$, is a finite value (Section 2, Cheng & Amin).

As an Accumulation Model

The same distribution measurement parameters γ, α, and θ, discussed in the previous section, also directly describe the data in the context of a continuous time-stochastic accumulation process, where a single latent quantity, X, continuously accumulates until it reaches a threshold. Such an accumulation process is also known as a type of Brownian motion process (BMP), and is at the base of all other popular accumulation models, e.g. DDM, LATER, E-LATER/LBA; where elementary changes in the accumulation process rules easily define one model or another. The SW is hence a close family member of the other popular accumulation models, which have been well-supported in prior literature to provide useful cognitive process models (Mulder, Van Maanen, & Forstmann, 2014; Ratcliff & Smith, 2004); and their exact mathematical relationships to the SW will be discussed in more detail later.

In the case of the SW, more specifically, X, accumulates at a given rate, γ, with noise until it reaches a threshold, α; and θ (the shift) is the minimal time lapsed outside of the process, which can be distributed before and after this accumulation process; the total time lapsed, T, is the data fit by the SW. This latent accumulation process provides a potential model for any data that involves a quantity accumulating over time that eventually reaches a value (or threshold). The SW thus provides the opportunity for a potentially-useful signal accumulation model, analogous to the hypothesized signal-to-response threshold event of behavior.

In the context of RT data and the appropriate experimental task, this kind of underlying accumulation process that we note is similarly shared (by elementary adjustments) with the other aforementioned accumulation models, has been well-supported to correspond to a signal-to-response threshold, neuro-behavioral event (for examples, see Gerstein & Mandelbrot, 1964; Mulder et al., 2014; O'Connell, Dockree, & Kelly, 2012; Smith & Ratcliff, 2004, and simulation work by Zandbelt et al., 2014 who show that in many cases, a single accumulator like the SW can often efficiently explain the result of a large ensemble of accumulators). In the case of the signal-to-response threshold interpretation of the SW: γ corresponds to the accumulation rate of the internal signal X, α to the threshold needed to initiate the physical response, and θ to the time distributed before and after this process (the time lapsed external of signal accumulation, abbreviated as TEA). Thus the total time lapsed, that is θ plus the accumulation time, is the RT recorded.

This latent accumulation process is illustrated in the left plot of Figure 3 for a single trial, in which a *random walk with drift* (RWD, beginning at $\theta = 200$ ms, and having average slope $\gamma = 0.08$) is simulated. Particularly, X starts at a value of 0, and accumulates with noise over time to eventually intercept the threshold $\alpha = 40$, lapsing a total time of 600 ms to reach the threshold. Then with the 200 ms of external accumulation time, the total RT is $200+600 = 800$ ms. A RWD thus corresponds to the accumulation of X over time, and provides a simulation of the SW intra-individual process model. A random-walk alone concerns movement due only to random noise, and a random-walk with drift concerns movement due to a steady accumulation tendency (γ), with random noise. Then in the right plot of the figure, many of these RWDs are simulated with the same SW parameter values as in the left plot (such as modeling an RT distribution of an experimental design cell), and it is shown that their final finishing times equate to a SW distribution with the same parameters: $\gamma = 0.08$, $\alpha = 40$, and $\theta = 200$, used to simulate the RWDs.

The design of the RWD of the SW, that many other accumulator models share, is the following form:

$$X_t = X_{t-1} + \gamma + \varepsilon\,, \tag{7}$$

where the position of a random variable X at time t, as X_t, is equal to its prior position value, X_{t-1}, plus a movement tendency, $\gamma > 0$ (known as drift), and marginal error, ε (or noise). The noise ε at each time step t, may be simply simulated by random draws from a Gaussian distribution with mean 0 and standard deviation 1.

Then note that any given threshold, $\alpha > 0$, unto which the time process terminates when X_t reaches that value, as $X_t \geq \alpha$, will produce a Wald distribution of data. That is because with probability 1, X_t will reach α; and thus for every RWD, one can expect a finishing time (T). Therefore, since T denotes the time t at which X_t reaches α, for a single RWD process with N runs, then the data is of the form

$$\mathbf{T} = (T_i)_{1\times N}\,, \tag{8}$$

for the N times (e.g. or RT observations) that the SW distribution describes (T is also known as the *first passage time* of the BMP).

Finally as mentioned previously, the shift parameter θ, accounts for all time passed outside of the accumulation event. Thus it accounts for aspects external to the RWD by shifting all values of t by a constant, in which the starting point of the accumulation process, $X_0 = 0$, instead becomes,

$$X_\theta = 0\,. \tag{9}$$

While θ shifts the distribution from the left, note that its effect, mathematically, is equivalent in being able to account for external processes that occur on either side of the accumulation event.

Utilizing the Shifted Wald

Whether one decides to utilize the SW as a distribution measurement tool, or as an accumulation model of the data, the approach is the same: to simply fit the distribution, which is to estimate its three parameters. It is the same approach since the parameters of the SW simultaneously describe both the shape of the RT distribution, and the data in the context of latent accumulation to threshold, using the same values.

Fitting Method

We developed a fitting method that combines techniques of deviance criterion minimization of observed-versus-predicted quantile distance, and maximum likelihood (ML) estimation, to fit the model parameters. The approach is detailed mathematically in the Appendix, and code to apply the method in R software (R Core Team, 2015) is provided as a supplementary file. In summary, the method utilizes an

observed data RT quantiles − model-predicted RT quantiles

minimization search across a single parameter, β from (6), to fit the model. More specifically, for every β, the other parameters of the SW may be directly calculated by closed-form ML estimators (e.g. see Nagatsuka & Balakrishnan, 2013), which are obtained through a method of moments approach. Then the model-predicted quantiles can be calculated, and the parameter set that leads to the minimum distance between the observed data RT quantiles is selected. We have found that the SW in the context of this method, is consistent in the recovery of parameters on simulated data, and satisfies well the model fit checks on appropriate real data (e.g. right-skewed RT distributions). Furthermore, the fitting procedure finishes on the level of minutes using standard personal computing technology. The following sections provide such illustrations of the SW approach on both simulated and real data applications. Then after the data applications, we will discuss ways in which the provided fitting procedure may be customized, as well as other options one may consider for fitting the SW. Before these data application sections however, we discuss three important topics: about the data fitting approach, model fit diagnostics, and data outliers.

Data Fitting Approach. In each data application we demonstrate an approach in which for every unique experimental design cell (combination of factors, by subject), a SW is fit. For example, a 2×2 design with 10 subjects would consist of $2\times 2\times 10 = 40$ SW distributions fit. Therefore, a mixture of SW distributions is accounting for the entire RT distribution.

The result of a full fit is hence a set of SW parameters (γ, α, and θ) obtained for each unique experimental design cell, and then both main effects and interactions can be assessed in the parameters, across levels of each condition. Then in order to provide a mechanism for standard hypothesis testing of such effects, we demonstrate utilizing an ANOVA on the parameters, however other analyses may be considered.

Model Fit Diagnostics. In each of the data applications, we demonstrate three diagnostics that can assess goodness of model fit, and whether the model is appropriate for the data. As ordered in the rows of Figure 4, the three diagnostics we propose consist of *(a)* a quantile-quantile (QQ) plot in milliseconds (ms) of the observed data deciles $(.1,.2,.3,.4,.5,.6,.7,.8,.9)$ against the model-predicted deciles (with all cells included); *(b)* a by-decile residual distribution plot, that includes residuals (difference in ms between the data deciles and the model-predicted deciles), which are then standardized through dividing by the parametric SD in (4); and *(c)*, a by-cell aggregate residual plot, in which for each cell, provides the sum of these standardized residuals across the cell's nine deciles.

The QQ plot provides an indication of overall trends in systematically misfitting quantiles of the distribution, as well as misfit outliers. In addition, it gives an idea about the scale and range of the data. The summarizing data points (dark

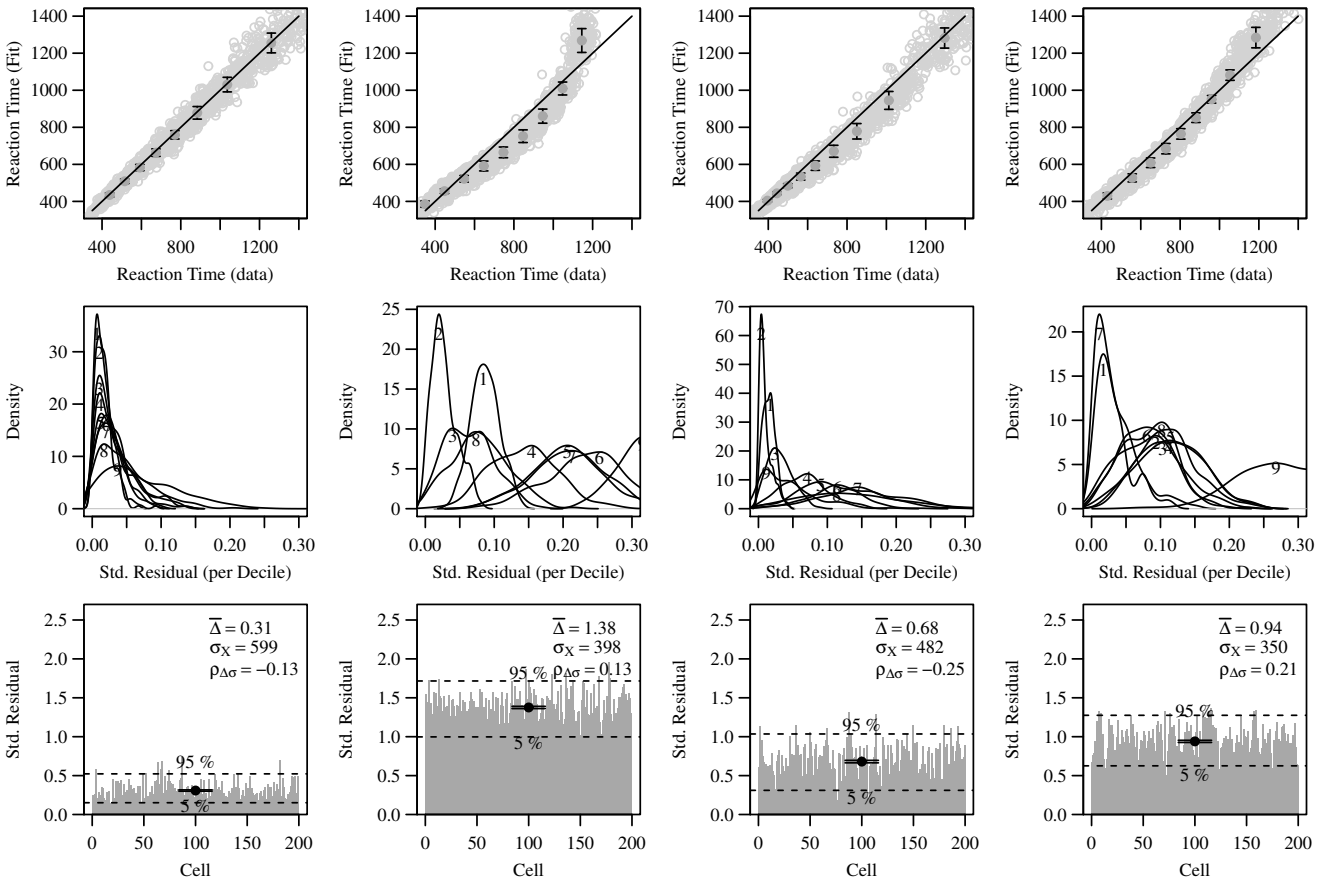

Figure 4. An illustration of appropriate and inappropriate model fit check performance on various simulated data sets, each with 250 observations and 200 cells fit. The checks are explained in detail within the model fit diagnostics section. Column 1 displays satisfactory fit on a SW-simulated data set, columns 2 (uniform distribution) and 3 (exponential distribution) show inappropriate fit, and column 4 (Gaussian distribution) illustrates poorer fit.

grey) are the mean data deciles against the mean model fit deciles, and the error bars are the standard deviation of the decile's residuals. The importance of this check is to observe critically any curvatures in the plot, which is a strong sign of misfit. For example, the second plot in Figure 4 shows the SW systematically underestimates the middling deciles and overestimates the final deciles; this is data simulated from a uniform distribution. Then in data simulated from the exponential distribution in the third plot, the model underestimates the middling deciles but captures the last decile (large right tail RTs). Then in the fourth plot with simulations from the Gaussian distribution, the model systematically overestimates the tail deciles; hence one can see that larger values of parameter α in Figure 2 brings the SW closer to normality, though the SW distribution may still have a tendency to produce larger right-tail values than a Gaussian distribution.

The decile residual distribution plot verifies an important observed property of SW-simulated or positive right-skewed data, in which residual magnitudes have a tendency to increase with data variance and magnitude. Such a trend can be observed in the decile distribution orderings in the first plot of Figure 4 (second row) and the right column of Figure 5 (note that each decile label is located at its respective distribution mode). These trends are exemplary of a strong fit. In contrast, very poor or inappropriate fits will be signaled by notably outlying decile distributions, such as in plots 2-4 of Figure 4 (second row). Then as for middle-range fits, which may also be due to right-skewed data that is more noisy, has small sample sizes, smaller numbers of cells, and/or an alternative underlying process model that is more compatible, one may observe no outlying distributions but instead a coarser ordering, such as in plots 2-4 of Figure 11 (second row).

Finally note that in these plots, the residuals are standardized by (4) in order to render this model fit diagnostic less-biased to differences between cells in overall data variance and magnitude. One may also consider other variation statistics for this standardization, such as the quantile standard errors of normal- or log-normal-transformed quantiles (see

Chapter 3.5, Wilcox, 2012).

Thirdly, the by-cell summed residual plot indicates a goodness-of-fit measure, Δ, for each cell. The lower and upper dotted lines respectively denote the 5% and 95% quantile range of these Δ values. The average cell goodness-of-fit is indicated by $\bar{\Delta}$, and the mean standard deviation (from Equation 4) of the data cells is σ_X. The standardization of these residuals by the SD in (4) aims to balance for the tendency of cells which have a large σ, to also have a large Δ; in which case comparisons of Δ between cells or data sets has less meaning. A measure for the efficiency of (4) as a standardization statistic is given by $\rho_{\Delta\sigma}$, which is the Pearson correlation between Δ and this statistic. In calculating $\rho_{\Delta\sigma}$, the statistic in (4) has been found to be a more stable measure than the raw data standard deviation, which may be selectively overestimated by values in the right tail. Hence since the effectiveness of (4) may depend on the goodness of model fit, larger values of $\rho_{\Delta\sigma}$ may also possibly signal poorer model fit such as in plots 3-4 of Figure 4 (third row).

Since as in our simulations, SW-simulated data fit with the SW generally result in: an in-line QQ plot, ordered or non-outlying standardized residual decile distributions (subject to data noise and adequate cell numbers), and lower $\bar{\Delta}$ and $\rho_{\Delta\sigma}$ values, while non-SW simulated data do not generally satisfy all of these three features, we suggest these diagnostics to assess goodness of model fit. Researchers may also consider examining other kinds of model fit checks.

Handling Data Outliers. It is important to note that very large RT values are not necessarily outliers (contaminant RTs), but may simply be large or extreme observations (or samples) of the 'true' underlying RT population distribution (e.g., see Barnett & Lewis, 1994). Therefore, these observations may be quite informative or essential for appropriately determining the shape of the distribution modeled. For example, parameter γ as in the first plot of Figure 2, accounts for tail thickness, and is indeed the signal accumulation rate parameter in the cognitive process model. Thus cutting out large RT values will consequently affect/reduce the information for determining appropriate γ values. This would also be the case if one is fitting an alternative accumulation model such as the DDM; cutting out the tail values and trying to "normalize" the RT distribution will likewise warp drift and threshold values—hence the RT tail is an important part of the data that should not be filtered off.

Therefore, it is recommended that only *contaminant* RT values should be removed from the data before being fit. Contaminant values may consist of RT values that arise from other processes or distributions which are foreign to the intended experimental control, such as pertaining to recording errors, lack of sincere participant effort, or spasmodic responses. Therefore, only uncharacteristically large RT tail values should be removed from the data before being fit. Likewise, inappropriate values before the leading edge of the distribution should be removed: such as recording machine mishap RTs, or spontaneous button pressing, where the respondent did not appropriately complete the task. Simulations have found that these faulty, early recordings can lead to inappropriate underestimation of the leading edge parameter, θ, which sensibly, will affect the estimation of the distribution shape parameters γ and α. One could also consider looking at RT values below three and above six (to preserve longer RT tail values) median absolute deviations (MADs, see Leys, Ley, Klein, Bernard, & Licata, 2013) from the median per cell, to infer whether these values may be contaminant RTs. Finally, also note that additional information on handling RT outliers is thoroughly described by Ratcliff and Tuerlinckx (2002).

Otherwise, the fitting method provided is naturally resistant to noisy or outlier data for the following four reasons—which any other method utilizing the same characteristics would also be: (1) A separate distribution (SW) is estimated per each design cell. Given that outliers tend to occupy a smaller percentage of design cells, and that there are typically many design cells, the mean parameter values for example will be much less affected by outliers in the raw data; (2) the fitting algorithm operates on the observed data quantiles, which naturally mitigate outlier effects; (3) the fitting algorithm uses the L1-norm distance (absolute distance), which is naturally more resiliant to outliers than the L2-norm distance (squared distances); (4) extreme quantiles are generally not fit by the algorithm, for example in real data cases the algorithm typically selects to fit the quantiles between the .01 to .99 range, but not further into the extent of .001 to .999.

Illustrations of the Shifted Wald on Data

In this section, the SW and the provided fitting method are first demonstrated on simulated data, and then on real data.

Application to Simulated Data

In this section the results are presented for a large simulated data study that consists of varying data set sizes. Specifically, observation length sizes between 1000 to as few as 15 observations are demonstrated; the simulation involves 1000 data sets (or e.g. experimental design cells) analyzed per observation size. The data-generating parameters were randomly drawn from uniform distributions. In the simulated analysis, the SW is fit (obtaining 3 fitted parameters) to each individual data set or design cell, and the recovery of parameters, as well as the fit of the observed data's quantiles, are calculated.

Table 1 provides the average parameter recovery trend from 1000 to 15 observations, across the 1000 data set simulations; and columns $E(RT)$ and $\mathrm{SD}(RT)$ provide the recovery of the expected value and standard deviation by equations (3) and (4). One can see that all parameters, as well as the mean and standard deviation, are strongly recovered

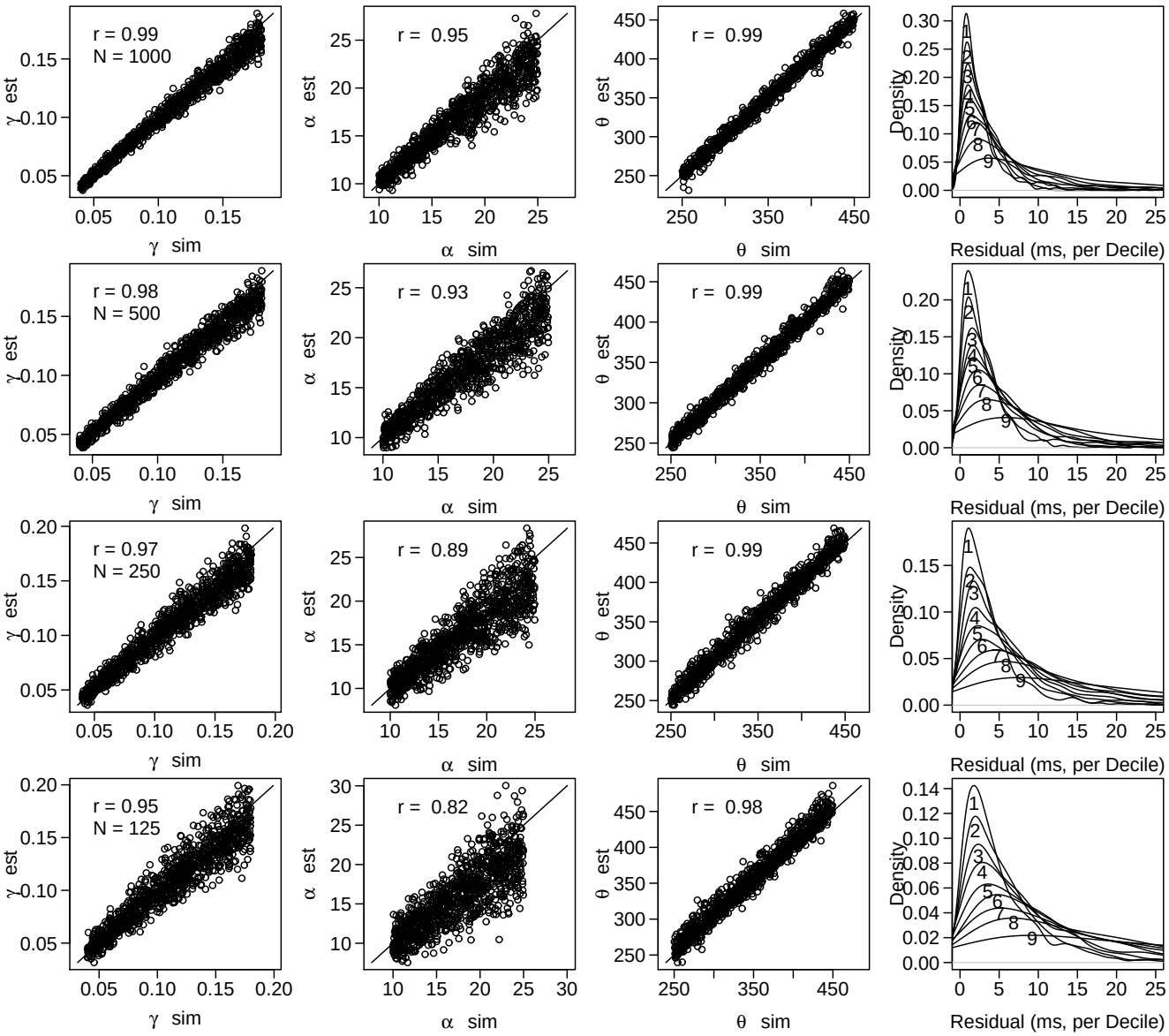

Figure 5. Application of the proposed method on four sizes of data: each row in the plot corresponds to 200 sets of data, that respectively have $N = 1000$, 500, 250, and 125 observations per data set.

with many observations, and even for low numbers such as near 125-50 observations. With very few observations, the recovery of parameter α is the most difficult. It is reasonable that parameter α is difficult to recover with fewer observations as it is primarily responsible for variance of the distribution around the mode. In the case of very few observations, this is a difficult measure to strongly recover in most any continuously-valued distribution.

Figure 5 then contains a visual plot of the parameter recovery results for the first four rows of the table (between 1000 to 125 observations), which can reflect if there are systematic trends that may not be reflected by the simple Pearson r correlation statistic. One can see that the model recovers the generating parameter values well and consistently, with almost no outliers. One may note that there is a small tendency for parameters γ and α to be very slightly underestimated together; however, since the parameter orderings are preserved in the recovery trends (e.g. the correlations are strong and the plotted recovery trend is packed and linear), parameter comparisons across fits of design cells remain ordered and meaningful. One may also note that fitting higher numbers of design cells may further increase fitting validity strength.

Finally, the right column provides the residual distribution

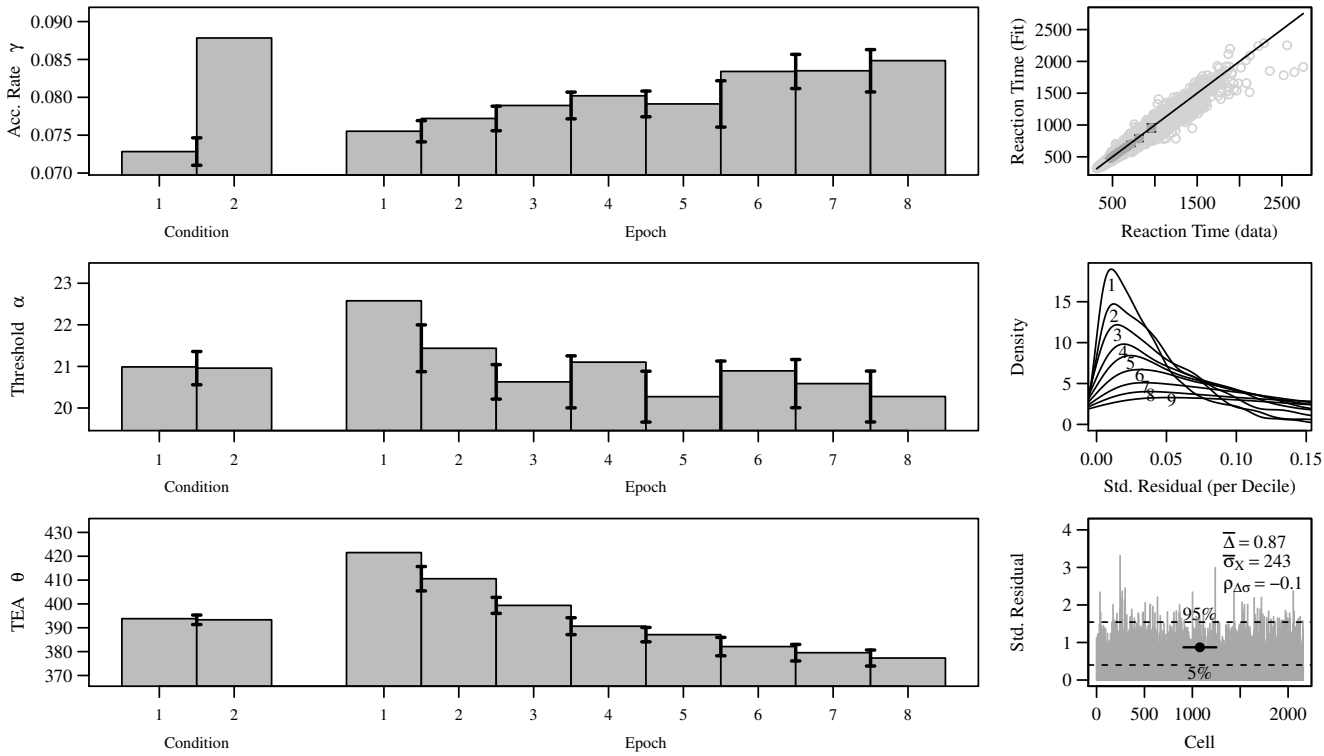

Figure 6. The SW fit to the manual response task: (left) main-effect mean parameter values with pairwise-difference error bars for each experimental factor; (right) model goodness-of-fit checks explained in detail within the model fit diagnostics section.

Observations	γ	α	θ	$E(RT)$	SD(RT)
$N = 1000$	0.99	0.95	0.99	1.00	1.00
$N = 500$	0.98	0.93	0.99	1.00	0.99
$N = 250$	0.97	0.89	0.99	1.00	0.99
$N = 125$	0.95	0.82	0.98	0.99	0.97
$N = 90$	0.93	0.78	0.98	0.98	0.96
$N = 50$	0.89	0.71	0.97	0.97	0.95
$N = 30$	0.82	0.61	0.95	0.94	0.87
$N = 15$	0.71	0.48	0.91	0.91	0.81

Table 1
Parameter Recovery, Average Pearson Correlations

diagnostic check with unstandardized residuals, to provide an opportunity to observe residual size on the natural scale of milliseconds. One can see that for this range of parameters (in simulated data), residual size generally occupies the range of 0 to 10 ms, and that residual size tends to improve (decrease) with increasing numbers of observations per cell.

Application to Real Data

In this section, the fitting approach is demonstrated on three published data sets (Casteau & Vitu, 2012; Goujon & Fagot, 2013; Hartendorp, Van der Stigchel, & Postma, 2013), that respectively represent data that arise from three canonical modes of responding: manual, vocal, and oculomotor modes. In each application, the results will be presented in the vocabulary of the SW as a simple cognitive process model for the data, and in tandem, with comments on the SW as a quantitative distribution measurement tool.

Manual Response Task. In this section, the fitting approach is demonstrated on a data set involving a manual-gesture response task. Collected by Goujon and Fagot (2013), baboons performed a visual search (VS) task, with contextual cues. The task consisted of visually searching for a target (the letter "T") that was embedded within configurations of distractors (letters "L"), which were either arranged predictively to locate the target (hence a contextual cue), or non-predictively (shuffled, without a cue), and the baboons responded by touching the target on the display screen. The experimenters explored an animal model (via baboons) of statistical learning mechanisms in humans, specifically the ability to implicitly extract and utilize statistical redundancies within the environment for goal-directed behavior. Twenty-seven baboons (species *Papio papio*) were trained to perform the task with contextual cueing.

As organized by the original researchers, there are three

meaningful partitions: the $C = 2$ predictive vs. non-predictive contextual cue conditions; the $E = 40$ time-points (epochs) to observe training effects, in which every unit step in E consists of 5 blocks (each block contains 12 trials, and thus each E contains 60 trials); and the $B = 27$ individual baboons. These three meaningful factors provide for $N = 2 \times 40 \times 27 = 2160$ separate distributions to each be individually fit by the SW; however 2158 were fit since one baboon did not have data for the 36th epoch. The average RT distribution length (number of observations) per design cell is $\bar{L} = 30$, with standard deviation, $\mathrm{SD}(L) = 1.10$.

Beginning with the model goodness-of-fit checks, the right column of plots in Figure 6 provides the information. The top plot contains the deciles of all $N = 2158$ distributions fit with the SW. As one can see, there is no systematic curvature in the plot and the SW performs systematically well on the data set. The plot also captures the range of the data, and that there are about 4-6 of the 2158 cells fit in which their 9th decile (upper right of the plot) are notably underestimated by the SW. Then the middle plot provides the distribution of standardized residuals for each of the nine deciles across the 2158 cells fit; here it is shown that the fit optimally satisfies an ordering of distribution modes and variances. Then finally, the bottom plot provides the sum standardized residual, Δ, by cell. Using the plot, one can observe which cells are more poorly fit. Overall, $\rho_{\Delta\sigma}$ is small at -0.1, which supports the Δ statistic as generally consistent. Furthermore, one can observe that given the fit is to real data with noise, the Δ's or $\bar{\Delta}$ are naturally larger here than fits to SW-simulated data (without noise or contaminant RTs) such as in Figures 4 and 11 (column one, third row).

The left column of Figure 6 provides the parameter main-effect results of the analysis for this manual-response VS task; in order to simplify the plot, the 40 epochs were averaged into eight training levels (each training level consists of five consecutive epochs). The left column with three plots contains the main-effect means, and pairwise-difference standard errors, of the model-fit measurements of the three SW parameters: γ, α, and θ, by experimental factor: the two conditions and eight training levels. The main-effect means are calculated by the mean of within-subject means for a given experimental level. The pairwise-difference standard errors are calculated for each pair of adjacent experimental levels, by computing the standard error of the within-subject differences between a pair of adjacent experimental levels; these standard error bars have been shown to be informative of the significance levels on the parameters that our ANOVA analyses return.

Beginning with the effect of the contextual cue condition on visual search latency in Figure 6, the latencies are considerably faster on average by an increase uniquely in the signal accumulation rate parameter, γ, when the cues are arranged in predictive patterns, and by no significant difference in the other parameters; this result is supported by ANOVAs across the three parameters ($F_\gamma(1,26) = 68.62, p < .001, \eta_p^2 = 0.73, \eta_G^2 = 0.113$; $F_\alpha(1,26) = 0.00, p = .95, \eta_p^2 = 0.00, \eta_G^2 = 0.000$; and $F_\theta(1,26) = 0.06, p = 0.80, \eta_p^2 = 0.00, \eta_G^2 = 0.000$); for an explanation of effect sizes η_p^2 and η_G^2, see work by Bakeman (2005). As a pure distribution measurement tool, the SW analysis replicates the faster latencies with predictive cues that the experimenters originally found using raw mean RT comparisons, however the SW locates exactly how the RT distribution is changed by the experimental manipulation: the tail of the RT distribution is much shorter when the cues are predictive, while the leading edge position and mode of the distribution are generally unchanged. This effect of the experimental cue factor is illustrated in Figure 7, where the RT distributions are plotted for the shuffled and predictive cue conditions (differing mainly by γ). The modes and leading edges of the RT distribution quantifications are about the same, and mainly the tail of the shuffled condition has a larger density.

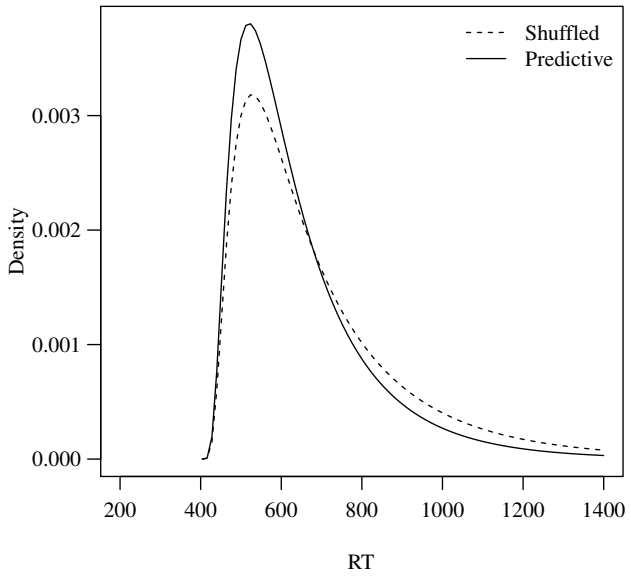

Figure 7. The shuffled and predictive RT distributions for the visual search, manual response task data set.

With regard to training effects on visual search latencies, all parameters were affected in ways that support faster RTs with more training, yet in different patterns; this result is also supported by ANOVAs across the three parameters ($F_\gamma(7,182) = 3.78, p < .001, \eta_p^2 = 0.13, \eta_G^2 = 0.021$; $F_\alpha(7,182) = 2.5, p = .02, \eta_p^2 = 0.09, \eta_G^2 = 0.024$; and $F_\theta(7,182) = 20.07, p < .001, \eta_p^2 = 0.44, \eta_G^2 = 0.045$). Over the training interval, both signal accumulation rate (increases) and external time to the accumulation (decreases)

adjust at a steady rate for improvement in RTs. In contrast, the signal criterion level, α, provides a sharp improvement (decreases) across training levels 1-3, and then appears to stabilize across the remaining training levels. As a distribution measurement tool, the SW analysis replicates the faster latencies with increased training and further specifies in which ways the distributions change: the leading edges become sooner with increased training, and the tail becomes shorter, the deviation around the mode becomes markedly smaller after the first three levels and then stabilizes. Next in the consideration of interaction effects, in aims for simplicity in these illustrations, we do not focus deeply on interaction details; however no significant interaction effects were found between training level and condition factors at the $p < 0.05$ level, but interaction is suggestive for each of the parameters if allowing the $p < 0.10$ level.

Vocal Response Task. In a picture-naming study by Hartendorp et al. (2013, Experiment 2), the authors explore the extent to which competing picture interpretation alternatives and distractor words influence vocal response latencies. The respondent views a picture and is instructed to name it vocally. In each trial, a morphed figure, consisting of a 2-dimensional blend (e.g. also see Burnett & Jellema, 2013) between two similarly-shaped objects (e.g. apple and heart) was presented; three different levels of *morphing* balance were assessed at $80/20\%$, $70/30\%$, and $60/40\%$. In addition, a distractor word was simultaneously presented as either identical to the object "apple-apple," semantically-related "orange-apple," or completely-unrelated "shirt-apple," termed as *priming*. The third factor is whether the distractor word was in relationship to the dominant morphing or the non-dominant, termed as *dominance*.

Twenty students from Utrecht University participated in the experiment. The $P = 3$ priming conditions, $M = 3$ morphing levels, and $D = 2$ dominance levels, with $N = 20$ participants provided for $N = 3 \times 3 \times 2 \times 20 = 360$ separate distributions to each be fit by the SW, with an average length of $\bar{L} = 11$, and $\text{SD}(L) = 1.2$; trials in which the non-dominant picture was named (a total of 4% from $4,015$ trials), were not included in the RTs analyzed.

Beginning with the model goodness-of-fit checks, the right column of plots in Figure 8 provides the information. The QQ plot shows no systematic misfitting of the deciles and fewer outliers than in the manual response data set fit. The decile residual distribution plot also shows a general ordering of deciles, albeit with more variance than the other experiment, which might be due to the smaller sample size of the average cell fit. This additional magnitude is also reflected in the by-cell summed residual plot, in which the Δ values are larger. Note also that $\rho_{\Delta\sigma}$ is at an appropriate value.

The left column of Figure 8 provides the parameter main-effect results of the analysis for this vocal response picture-naming task; in order to simplify the presentation, the main effects for cases only in which the distractor word is in relationship to the dominant morphing are presented. Beginning with the effect of picture morphing intensity on picture naming latency, pictures with a clear distinction (at least 70%) of the primary object, reduced picture naming latency by an increased signal accumulation rate, γ; and no distinct effect was observed in the other parameters. These results are supported by the ANOVA ($F_\gamma(2,38) = 7.79, p = .001, \eta_p^2 = 0.29, \eta_G^2 = 0.037$; $F_\alpha(2,38) = 0.30, p = .75, \eta_p^2 = 0.02, \eta_G^2 = 0.004$; and $F_\theta(2,38) = 0.85, p = 0.44, \eta_p^2 = 0.04, \eta_G^2 = 0.008$).

In regard to the effect of distractor word priming on picture naming latency in Figure 8, distinct significant effects are observed in each of the parameters ($F_\gamma(2,38) = 3.61, p = 0.03, \eta_p^2 = 0.16, \eta_G^2 = 0.037$; $F_\alpha(2,38) = 4.1, p = .02, \eta_p^2 = 0.18, \eta_G^2 = 0.051$; and $F_\theta(2,38) = 6.76, p = 0.003, \eta_p^2 = 0.26, \eta_G^2 = 0.055$). Firstly, faster RTs by an increased signal accumulation rate only occurs when the prime is identical to the target word. Secondly, slower RTs by a larger information accumulation criterion, α, occurs when the prime is semantically related to the word but not the true picture name. Thirdly, faster RTs occur by sooner leading edges of the distribution, θ, as the semantic prime becomes more similar to the picture name. Finally, in the case of γ, significant interaction effects were found between priming and morphing at the $p = .02$ level, but not for α and θ. Note that while this data set provides an interesting example for goodness of fit to cells with notably smaller numbers of observations in a vocal response task, the parameter results should be taken with caution due to the small numbers of observations per cell.

Oculomotor Response Task. In this section, the fitting approach is demonstrated on a data set involving an oculomotor response task. Collected by Casteau and Vitu (2012), adults performed saccadic eye movements in order to locate a target (an 'h' or 'k') at varying distances (aka *eccentricities*) from the central fixation point; either with also a distractor target (an 'o', at varying eccentricities), or without a distractor (control). In this paradigm, the RT of each trial is the saccade latency: the total fixation time prior to making a single saccade that arrives at the target stimulus. Eight students from Aix-Marseille Université, between ages 18 to 23 years, participated in the experiment; all reported having normal vision and were unaware of the purpose of the experiment.

The experimental design cells consist of $C = 2$ conditions (no-distractor, distractor), $T = 9$ levels of target eccentricities (distances from the fovea fixation point, in degree units from 1-6°), $D = 5$ levels of distractor eccentricities (from 0-3°), and $DT = 3$ levels of distractor-to-target distances (4°, 5°, and 6°). We fit the same sections of the balanced design cells as organized by the original experimenters: the control condition over subjects and the levels of target eccentricities,

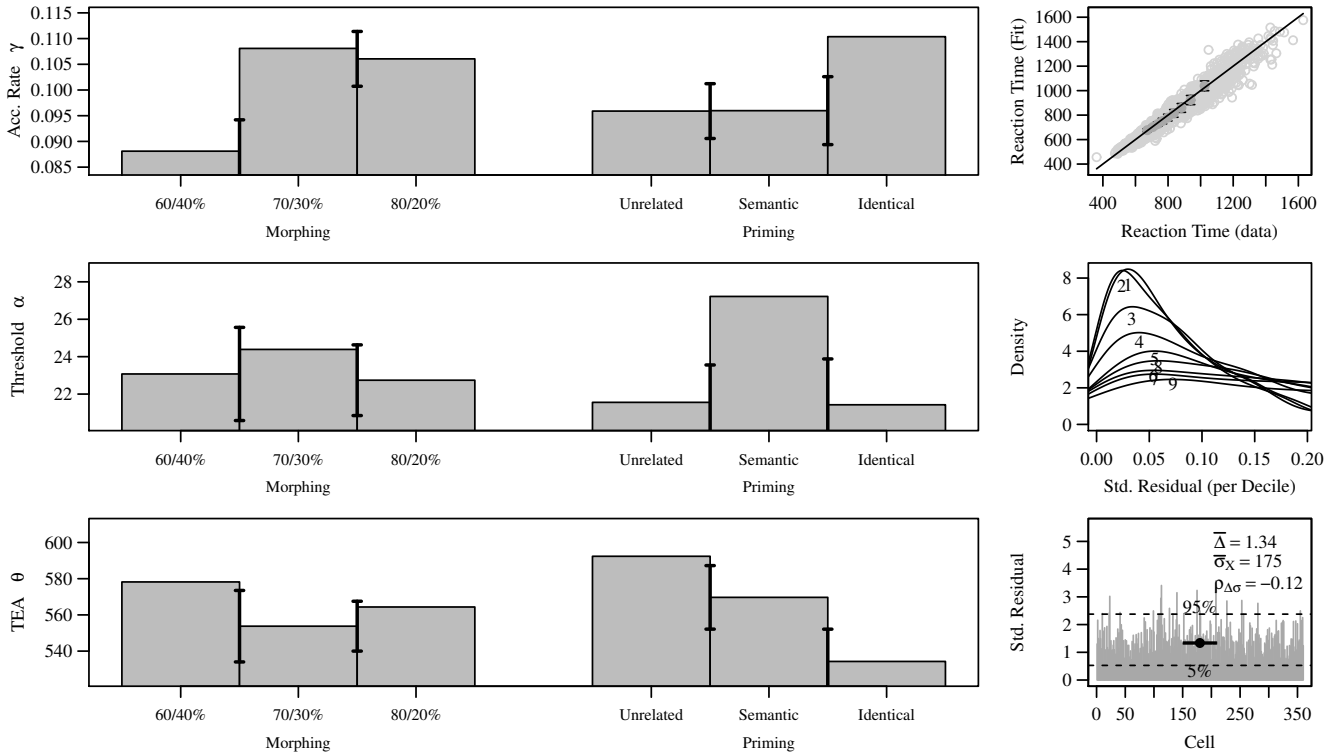

Figure 8. The SW fit to the vocal response task, picture-naming with word and visual distractors: (left) main-effect mean parameter values with pairwise-difference error bars for each experimental factor; (right) model goodness-of-fit checks explained in detail within the model fit diagnostics section.

and the experimental condition over subjects, the distractor eccentricities, and distractor-target distances. Each of these unique combinations lead to $N = (8 \times 9) + (8 \times 5 \times 3) = 192$ separate distributions total to each be fit by the SW; the average distribution length for each cell fit is $\bar{L} = 25$ with $\mathrm{SD}(L) = 5.3$; trials with blinks, or more than one saccade to arrive at the target are not included in the RTs analyzed.

Beginning with the model goodness-of-fit checks, the right column of plots in Figure 9 provides the information. One can see that the data of this paradigm is also fit well by the SW. In the QQ plot there is no systematic curvature, and the SW performs systematically well on the data set. The plot also captures the range of the data, and that there are 5 of the 192 design cell cases in which the 9th decile is notably underestimated by the SW. One can also observe that the observed RTs here occupy a faster interval than in the other experiments. Secondly, the decile residual distribution plot also shows an appropriate ordering of the distributions. Thirdly, the by-cell residual sum plot shows the Δ values to be larger than the manual response task data, but smaller than the vocal response task data, and $\rho_{\Delta\sigma}$ at 0.07 is the smallest of the three experiments.

The left column Figure 9 provides the parameter main-effect results of the analysis for this oculomotor response task; in order to simplify the presentation, the main effects only for presence of distractor and distractor eccentricities are illustrated; target eccentricity effects are then presented in a second plot. Beginning with the effect of presence of distractor on saccade latency, the lack of distractor decreases saccade latency by an increased signal accumulation rate, γ, and a flat overall decrease in external time, θ ($F_\gamma(1,7) = 3.85, p = 0.09, \eta_p^2 = 0.35, \eta_G^2 = 0.122$; $F_\alpha(1,7) = 0.40, p = .55, \eta_p^2 = 0.05, \eta_G^2 = 0.028$; and $F_\theta(1,7) = 28.8, p = .001, \eta_p^2 = 0.80, \eta_G^2 = 0.122$). In this case, the leading edge positions of RT distributions from trials with the presence of a distractor are systematically larger than in trials with no presence of distractor, and the accumulation rate of the signal therein is reduced as well; therefore the tail of the RT distribution is also notably shorter in distractor cases.

Next, the effect of distractor eccentricity on saccade latency provides for larger latencies as the distractor is closer to the fovea fixation point (eccentricities near 0). This is observed with larger signal thresholds needed, α, as the dis-

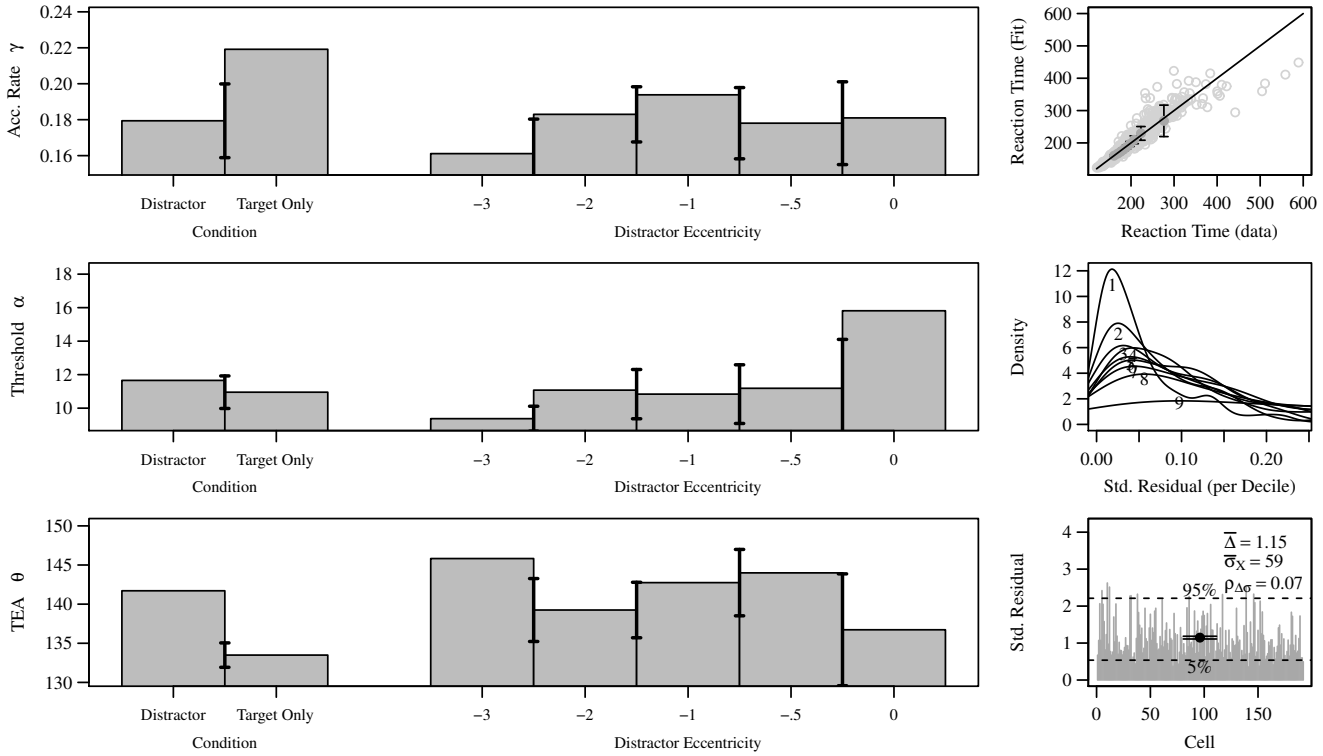

Figure 9. The SW fit to the oculomotor response task, saccadic eye movements for targets with distractors: (left) main-effect mean parameter values with pairwise-difference error bars for each experimental factor; (right) model goodness-of-fit checks explained in detail within the model fit diagnostics section.

tractor is closer, and no other distinct effects are provided by the other parameters (γ ($F_\gamma(4,28) = 0.60, p = 0.67, \eta_p^2 = 0.08, \eta_G^2 = 0.035$; $F_\alpha(4,28) = 2.58, p = .06, \eta_p^2 = 0.27, \eta_G^2 = 0.209$; and $F_\theta(4,28) = 0.98, p = 0.4, \eta_p^2 = 0.12, \eta_G^2 = 0.046$). Incidentally, this kind of trend is also observed in the mean latencies of the original paper, and here we see that the longer duration to move the eye is explained by increased signal threshold parameter α values when the distractor is placed more eccentrically at the fovea fixation point.

Finally for purposes of simplicity, Figure 9 does not include the distractor-target distance parameter values; and no significant effects were found in the model parameter results for this factor. This was also the case in the results of the original paper that analyzed the raw mean saccade latencies. In addition, no significant interactions were found in the ANOVA analysis between distractor eccentricity and distractor-target distance levels in the parameters.

The effect of target eccentricity on saccade latency for non-distractor trials is instead displayed in the left plot of Figure 10, in which a decreasing trend on saccade latency over the first few eccentricities, that then then levels out, is observed. This effect is located in the external time parameter θ; and no distinct trends are observed in the other parameters ($F_\gamma(8,56) = 0.39, p = 0.92, \eta_p^2 = 0.05, \eta_G^2 = 0.029$; $F_\alpha(8,56) = 0.35, p = 0.94, \eta_p^2 = 0.47, \eta_G^2 = 0.046$; and $F_\theta(8,56) = 2.72, p = .01, \eta_p^2 = 0.28, \eta_G^2 = 0.199$). The trend in θ replicates the trend in the median latencies observed as in the right plot of Figure 10, and has a similarity to also the raw mean latencies, as shown in the original paper (see left

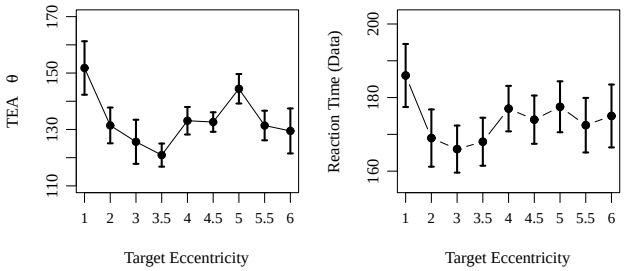

Figure 10. The effect of target eccentricity in the control condition is observed to be significant on parameter θ (left). The trend replicates the effect on the median RT values (right).

plot, Figure 3 of Casteau & Vitu, 2012). Therefore, the RT distributions follow a similar stepwise pattern in their leading edge positions over varying target eccentricities, without significant differences between the RT distribution shapes.

When to Apply the SW

We have demonstrated the SW on both simulated and real data applications. The SW may be generalized to a number of additional paradigms, since the SW is characterized by a probability density function that can be applied to any positively-valued data with a degree of right skew. Then a goodness of fit may be analyzed by model fit checks, such as the ones provided, to determine whether the model fit is appropriate, and if the fit is strong or poor. In Figure 4 for example, we show the behavior of the SW when it is applied to various distributions that can take on incompatible shapes (e.g. lacking skew, unimodality). Then by establishing the criterion that the model should first reproduce sufficiently the observed data RTs, before one interprets the results, the model fit diagnostics were assessed before interpreting the parameter results in our real data application sections. Then in these real data applications, example possible data sets were demonstrated in which a SW fit provided a useful distributional measurement and/or process-model analysis of the data.

When the SW may exhibit poorer performance

Generally, it is less-likely that the SW will appropriately fit the RT distribution when it is applied to data with many error responses. This is because it is well-known that generally, the error responses have a different distribution than the correct responses (e.g. Ratcliff & Rouder, 1998); thus a single distribution fit to two different underlying distributions may indeed cause a misfit or poorer fit. In this case the SW is more likely to satisfy the model fit checks when it is applied separately to the sets of errors and correct responses than both at the same time. However, note that in this case the SW will not be explanatory in a predictive fashion for rates of correct/error responses, but rather explanatory in a descriptive fashion for cases of correct or error responses; it would hence serve as a more elementary model than a further complex one, which predicts either response within a single accumulation process, such as the DDM.

The second possible case in which the SW may not fit well to the RT data is during very long response time tasks, where a participant may be switching between a number of response strategies (each resulting in a different type of RT distribution), that are hence not easily parsible/predictable by the experimental conditions (for additional information, see Van Maanen, de Jong, & Van Rijn, 2014). Hence without being able to parse the data for when a participant is changing strategies, the non-parsed RT distribution may not be of consistent form, or of easily predictable form for the SW in this case. Thus in summary, any RT distribution that is not behaving in a shape conformable to what a SW distribution can reproduce, may fail the model fit checks. In these cases, there may be more complex models of accumulation worth considering for fitting the data.

Considering Other RT Process Models

More complicated process models, have the potential to provide more detailed information of the process underlying a response task when there is enough and adequate data to fit them. With the right parameterization, they may also be capable of predicting more complexly-shaped RT distributions.

Model Complexity. However, it is worth noting that as a cost of greater model complexity (number of parameters), fewer experimental condition levels may be analyzed in the data; and it is not always the case that a more complex model is more useful for an RT data set. For example, a key benefit of the SW is it is a very simple process model with only three parameters, that can be easily applied to fit experimental data at a high analytical resolution: e.g. being specified to model each level of each experimental condition. In contrast, accumulation models with additional numbers of parameters may have to aggregate data over several condition levels in order to have enough observations to appropriately estimate the additional parameters; and these levels may not be linearly ordered, such that their aggregation could be problematic, or much less informative.

Thus while having more parameters may be more informative, data sets that lack sufficient numbers of trials along each type of observation and experimental condition combination, might not allow an appropriate fit of a model with such complexity, at such an equivalently high resolution. For example the three data sets analyzed in our applications had at times, as few as 20 trials in a design cell for just one characteristic response, and any more complex accumulation model with additional parameters or absorbing boundaries would be *overparameterized* or inappropriate for analyzing the data at this resolution.

Extended SW Model. In order to account for more complicated signal accumulation processes, the SW process in (7) may also be extended or augmented. However, the resultant distribution is no longer mathematically, directly equivalent to a SW distribution, but it still may be estimated by a SW distribution. In some cases, extended or more complicated processes may be adequately summarized by a single, standard SW process, and this can be observed in the model fit checks; for example, in our real data applications, Figures 6, 8, and 9. However, when the extended processes are too strongly different, the SW will instead provide a cruder summary, and poorer fit than the native model.

Specifically, this is illustrated in columns two and three of Figure 11, where the response processes of (7) are strongly altered, and the SW provides a poorer fit: column two per-

tains to a sum of two SW processes (the first has a slow drift $\gamma_1 \in [.04, .08]$, the second has a fast drift $\gamma_2 \in [.12, .18]$); column three pertains to a single SW process, in which the first 100 ms of the process has a slow negative drift ($\gamma_1 = -.03$), and after 100 ms the drift is faster and positive, $\gamma_2 \in [.04, .18]$. In both cases, a SW may fit and summarize the process with a single aggregate positive drift, γ, though the fit is markedly poorer than data simulated from a regular SW distribution, as in the first column of Figure 11. These simulations were intentionally constructed to illustrate a clearly poorer fit when the basic process of (7) is strongly altered. However, a large number of other parameter combinations were found in which these two extensions may be fit almost as well with a regular SW model; sensibly, when these extended processes become easily summarized by a single, standard SW process with single positive drift, γ. It is also worth mentioning that when doing the same simulations, but using much fewer observations (e.g. less than 50), we found that these extended processes become more difficult to differentiate from one another.

In addition to sums of SW processes or internal drift shifts, in which there are multiple drift rates, e.g. γ_1 and γ_2, the process in (7) may also be extended in other notable ways: such as by introducing between-trial error in the basic accumulation rate γ, or a collapsing threshold over time α (though recent work suggests that this might often not be necessary, see Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015). Furthermore, if one has enough observations for a number of characteristic responses, one can consider fitting a multi-accumulator SW model (see Usher, Olami, & McClelland, 2002; Vickers, 1970, 1979), in which multiple SW accumulators race against each other, and this is known as the race model. Alternatively, each characteristic response can be partitioned into a separate design cell, and thus a separate SW is estimated per characteristic response for each condition. Finally, one may also consider extending the modeling of the TEA parameter, θ, as being exponential in value by using a Wald distribution that is shifted exponentially, termed the ex-Wald by Schwarz (2001).

More Complex Accumulation Models. Some mentionable more complex accumulation models than the regular SW (Luce, 1986; Ricciardi, 1977; Wald, 1947) are the DDM (Ratcliff, 1978), E-LATER/LBA (Brown & Heathcote, 2008; Nakahara et al., 2006), and multi-accumulator SW (race, LaBerge 1962; Usher et al. 2002; Vickers 1970, 1979) models, previously mentioned in the introduction. Generally, all of these models aim to handle more than one characteristic response within a single accumulation process: either in which there is a single accumulator with two boundaries (DDM), or multiple accumulators each with one boundary that race against each other (E-LATER/LBA, race). They hence describe a more complex accumulation process. Thus a single characteristic response observed over varying latencies may be well-described by a SW, but two or more may be more interestingly-modeled by the DDM, E-LATER/LBA, and race models.

The SW can of course still be applied to such data, but it will provide a more aggregate description, as a cruder model. For example in the fourth plot of Figure 11, the SW is fit to classical DDM simulated data (as in Ratcliff & Rouder, 1998; Ratcliff et al., 1999, thus not the extended DDM), and in which the upper and lower threshold responses are grouped in the same design cells. While the fit is acceptable, it is a worse fit than to data simulated by the SW in the first plot. It is also worth noting that there are other randomizations of DDM parameters, which may provide better or worse (e.g., by the extended DDM) recovery of the observed quantiles when fit by a SW.

Accumulation Model Similarities. Many accumulation models, such as the SW, DDM, race, LATER, and E-LATER/LBA are highly similar in that they all share the same continuous time-stochastic accumulation process (a type of Brownian motion), where a single latent quantity, X, continuously accumulates until it reaches a threshold; and elementary changes in the accumulation process rules easily define one model or another. In this section, we explain the relationships.

Drift Diffusion Model. The DDM contains the same exact accumulator as the SW. However, a lower absorbing boundary is created, and negative values for γ are allowed (achieved by adding parameter z, as indicative of a starting value for X, which is between 0 and α; and letting the lower boundary be at value 0) . Note that a negative γ results in a preference for choosing the secondary characteristic response option, which is a preference for X_t to be absorbed in the lower boundary.

Race Model. The race model contains the same exact accumulator as the SW. Then any $k > 1$ type of response options can be more exhaustively modeled by installing k replica instances of the SW accumulator that each race against each other. The first accumulator that reaches its threshold is the response observed.

LATER, E-LATER, and LBA. The LATER, E-LATER and LBA have the same type of accumulator as the SW. However, the inter-time error in the accumulation of X_t is set to 0, and rather the error is placed on modulating the slope, or accumulation rate γ. Thus in the these models, X_t accumulates at a linear constant rate. Then also as in the SW, one can make many accumulators of this kind race against each other until the first one wins; in the case of one accumulator it is called the LATER model, in the case of two or more, it is the E-LATER/LBA (they are equivalent). Note also that when one uses multiple accumulators, one can insert an additional parameter, a different starting point (of X_t) for each of the accumulators, if all accumulators for example, share the same threshold. The LATER model is principally different from the SW because it necessarily predicts RTs in

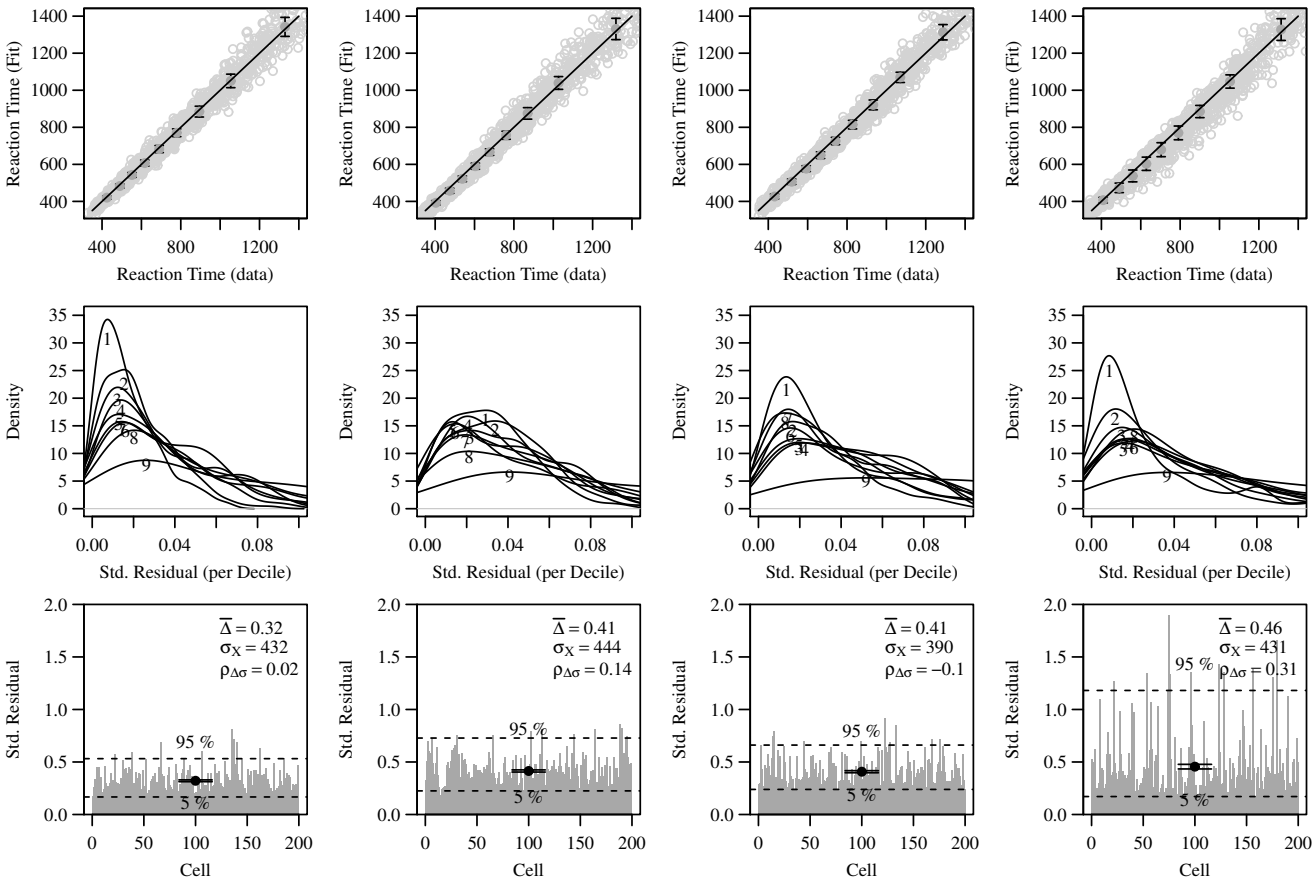

Figure 11. The SW fit to data simulated by more complex random-walk simulations; each with 300 observations, 200 design cells each. From left to right: the SW; sum of two SWs, the first process has a slower drift than the second, and $\theta = 0$ for the second; a single SW process but the γ for the first 100 ms is set to -0.03; classic DDM. These model fit checks are explained in detail within the model fit diagnostics section.

which their inverse is distributed normal (this is akin to the lognormal model, which necessarily predicts the log of the RTs are distributed normal), while this is not necessarily the case for the SW; they are also different because the LATER allows negative drift rates, and in these cases the process will usually not terminate.

Supra-model. Where elementary changes in the accumulation process easily define one of these models from another, it is apparent that all of these approaches are very closely related, and it could be said that they constitute the very same supra-model. It is suggested that one should choose the model that the data can appropriately support, based on whether the data provide enough numbers of observations for the level of model complexity, and if the model appropriately satisfies the model fit diagnostics. Preferably, the more complex model will provide more information along each experimental condition if the data have enough observations.

Considering Other Fitting Methods

In the present paper we aimed to present a fitting method that is effective, practical, and easy to implement. Its effectiveness was demonstrated on both simulated data criteria and real data applications. Given the effort toward practicality, efficiency, and ease of use, the fitting method provided serves as a baseline in which additional developments may be made. Future developments may focus on exploring other renditions of maximum likelihood estimation or deviance criterion minimization (e.g. see Basak & Balakrishnan, 2012; R. S. Chhikara & Folks, 1974; Heathcote, 2004; Koutrouvelis et al., 2005; Nagatsuka & Balakrishnan, 2013; Padgett & Wei, 1979; Vladimirescu & Tunaru, 2003), in order to fit the distribution. In addition, expert users may consider some of the more data-intensive fitting approaches, such as the hierarchical Bayesian approach. In this section, we first discuss ways in which the proposed fitting method may be modified; and secondly, we discuss as-

pects of the hierarchical Bayesian framework, which may be preferable for expert users, given that it has additional levels of complexity, and is much more time-intensive for large RT data sets.

Extending the Current Method. The proposed fitting method combined techniques of deviance criterion minimization, and maximum likelihood estimation to fit the model parameters. There is room for customization as to which $(i.)$ quantiles to minimize and $(ii.)$ which kind of distance (e.g. absolute, or squared differences) to use in the minimization search. Our algorithm checks a number of options and selects the better fit; however these options may be expanded to include additional candidates, or instead reduced, to optimize run-time.

Our work in exploring $(i.)$ and $(ii.)$ has found that large ranges for $(i.)$ e.g. from 0.01-0.99 with near 100 quantiles provide the best recovery of parameters. It has been found particularly important to fit the exterior quantiles. For example not fitting between quantiles 0.01-0.05 results in an overestimated leading edge parameter, θ, which sensibly, also negatively affects the recovery of the distribution shape parameters, γ and α. However, while it is ideal to fit the exterior quantiles, real data may have uninformative outliers (contaminant RTs) that one may not want to provide weight to, thus small adjustments, such as the quantile range 0.02-0.98, may provide for better fits in some cases. The code provided for example, calculates both (and also others), and selects the better fit. In regard to $(ii.)$ the squared distance has been found to provide nearly similar recovery of parameters, however a worser recovery for quantiles 0.7-0.9 (this can be observed in the decile residual distribution plot), and it is also more sensitive to be affected by outlier values.

Bayesian Framework. A noteworthy distinction between the direct likelihood estimation approaches and the Bayesian framework, is the ability to impose a hierarchical model with the Bayesian approach. The primary effect of the hierarchical model is the ability to constrain the within- or between-subject error (depending on how one chooses the hierarchy) in the estimation of the parameters. In contrast in the case of our approach, there is no constraint *a priori* on the within-subject error, and so significant differences may be a little more difficult to find in an ANOVA analysis on the parameters; but at the same time, the results achieved with the method applied herein are not dependent on any Bayesian 'prior beliefs.'

A number of arguments in favor of hierarchical Bayesian models (e.g. Anders & Batchelder, 2013; Averell & Heathcote, 2011; Kemp, Perfors, & Tenenbaum, 2007; Oravecz, Anders, & Batchelder, 2013; Rouder, Morey, & Pratte, 2013; Zeigenfuse & Lee, 2010) are provided in previous works (Lee, 2008; Lee & Wagenmakers, 2014; Rouder et al., 2013). The location and variance in parameter estimations can also be constrained in the non-hierarchical Bayesian framework, simply by specifying informative prior beliefs for the non-hierarchical distributions. However in the context of analyzing very large RT data sets, it may be worth considering that the run time for the Bayesian framework analysis is exponentially longer, in comparison to the presented method, and one will also have to verify that there was valid mixing in the Bayesian model fit by the assessment of within-chain auto-correlation, convergence, between-chain similarity, and so forth (see Gelman, Carlin, Stern, & Rubin, 2004, for mixing terminology); and also that the fit resembles the data. Thus obtaining an appropriate Bayesian fit that resembles well the RT data may take a number of days if not longer, and thus the approach may be more appropriate for expert users; despite if the approach may possess a number of advantages.

Therefore the Bayesian approach is a probabilistic framework for assessing parameter values, and provides additional opportunities to constrain the locations and variances in which the parameters are estimated that the direct likelihood estimation or deviance criterion approaches do not have in this way. Depending on the case of usage, each of these approaches have unique benefits, drawbacks, or implicit assumptions, and the practitioner should utilize the model-fitting method that best satisfies their analytical needs.

General Discussion

A methodology and comprehensive illustration of the SW distribution for RT data analysis was developed. An effective fitting method was established, its mathematical properties are provided in the Appendix, and R code to apply it is included as a supplementary file. Simulated data applications show the effectiveness of the fitting method, and real data applications show that the SW model appropriately fits the real data that arise from three canonical modes of responding: manual, vocal, and oculomotor modes; and the model results are quite sensible along the experimental factors. It was shown that the SW distribution can be used as a cognitive process model, or alternatively, as a quantitative distributional measurement tool without theoretical implications.

The SW was discussed in the context of other accumulation models, and was shown that in cases of mainly only one characteristic response observed at varying latencies, and over various experimental conditions, more complex accumulation models may be too overparameterized to be appropriately fit to such data. Thus in these cases, the SW may provide a useful, and simple process model across a number of experimental factor levels; whereas in cases of many observations across all response options per experimental condition, more complex accumulation models, such as the DDM, race, or E-LATER/LBA models, would be more informative and should be used.

Most accumulation models however share the same understructure, as in (7), and elementary adjustments will define

one popular model from the other. For example in the case of the SW and DDM, Gerstein and Mandelbrot (1964) made no grand proprietary distinction between the two models in applications of neuronal spike modeling. Other works have also noted the close relationships (see respectively Chapter 3, and pages 8–24, R. Chhikara, 1988; Jones & Dzhafarov, 2014, for more information). But it is also important to note that when the basic accumulation rules are adjusted, it is indeed expected that the model parameters have a different significance. For example, the single drift rate of the DDM corresponds to the shapes of two underlying distributions (e.g. corrects and errors) while the single drift rate of the SW corresponds to the shape of the full RT distribution. Matzke and Wagenmakers (2009) perform a direct exercise to show this, that data simulated with a two-boundary accumulation process that also allows negative drifts (the DDM), fitted with a single boundary accumulation process that only allows positive drifts (the SW), provide parameter results that do not correspond exactly to the other. They also demonstrate the same result when one tries to fit the SW to data with many observations on both of two different characteristic responses (e.g. corrects and errors), which are well-known to have separate underlying RT distributions; that rather the DDM should be fitting. As mentioned, this is a typical example case in which the SW is not likely to perform as well on model fit diagnostics; unless the SW is separately fit for each characteristic response, or if the distribution of the two characteristic responses are adequately similar to one another in the task.

We have made suggestions for when to consider applying the SW and when one may consider applying other accumulation models of higher complexity. The model fit diagnostics may also provide an indication if the SW is appropriately accounting for the observed data. While there are certainly appropriate situations and data that could considerably benefit from a SW analysis approach, currently there are very few publications in the psychological literature that utilize the distribution. For example, in the three published data sets we analyzed, they are all tasks in which the original authors found errors to be so sparse and uninformative; in these cases, a more complex model such as the DDM would be overparameterized with not enough errors to fit per design cell, and in contrast the SW is not overparameterized, it is able to provide an interesting process or measurement model analysis of the data, that is a more sophisticated statistical analysis than simple raw mean and standard deviation comparisons. Therefore, through our detailed accounts, demonstrations, and discussions of the SW, both as a possible distribution measurement tool, and cognitive process model, we hope to have advocated the distribution's use, as well as to have facilitated a deeper understanding of the SW, and its position in the context of accumulation modeling.

Appendix

Fitting Method Details

This section details the fitting method, which is a combination of maximum likelihood estimation (MLE) and deviance criterion minimization. An overview of the method is as follows:

[1] Select a candidate β value
[2] Calculate $\hat{\theta}$ and $\hat{\alpha}$ using MLEs (10) and (11)
[3] Calculate $\hat{\gamma}$ using (13)
[4] Calculate the deviance criterion using (15)
[5] Repeat 1–4 across β's near-entire parameter space
[6] Select the distribution with the smallest deviance value

The next paragraphs explain these steps in further detail.

To obtain candidates for β in [1], a simple search algorithm selects candidates in the near-entire plausible range, such as from $(0.001, 1000)$. Then for each β, there are closed-form MLEs to calculate $\hat{\theta}$ and $\hat{\alpha}$ for [2], which as in Nagatsuka and Balakrishnan (2013), are:

$$\hat{\theta} = X_0 - \hat{\alpha}_0^2 \int_0^\infty (1 - F[z; \hat{\beta}, 1, 0])^N \, \mathrm{dz} \tag{10}$$

$$\hat{\alpha} = \left(\frac{1}{M} \sum_{k=1}^{M} [X_k - \hat{\theta}]^{-1} - [\bar{X} - \hat{\theta}]^{-1} \right)^{-1/2}, \tag{11}$$

where X_0 is the data minimum, X_k is a data point, M is the number of data points, and $F(\cdot)$ is the cdf of (1). The data submitted for X are the observed data quantiles ($\mathbf{Q}^O$, defined in the next paragraphs), as they have been found to provide more stable data measures that are less disturbed by outliers, than by using the raw data values. By MLEs (10) and (11), it is a four-step process to calculate parameters $\hat{\theta}$ and $\hat{\alpha}$. First an initial estimate of $\hat{\theta}^*$ is calculated via (10) by using the following initial estimate for $\hat{\alpha}_0$,

$$\hat{\alpha}_0^* = \sqrt{\frac{(\bar{X} - X_1)^3}{\frac{1}{M}\sum_k^M [X_k - \bar{X}]^2}}. \tag{12}$$

Then $\hat{\alpha}_0$ is calculated by (11) with $\hat{\theta} = \hat{\theta}^*$. Next with $\hat{\alpha}_0$, the final $\hat{\theta}$ is simply calculated by (10), and likewise $\hat{\alpha}$ by (11).

Then for [3], since β and $\hat{\alpha}$ are already known, it is trivial to calculate $\hat{\gamma}$ since

$$\hat{\gamma} = \frac{1}{\beta \hat{\alpha}}. \tag{13}$$

In [4], the suggested deviance criterion, Δ, is the sum of distances between M specified quantiles of the observed and real data, between probabilities (p_1, p_2). We denote $\mathbf{Q}^O = (Q_k^O)_{1\times M}$ for the observed data quantiles and $\mathbf{Q}^F = (Q_k^F)_{1\times M}$ for the candidate fit, where Q_k^* is the quantile at probability

$$p_k = p_0 + (k-1)\frac{(p_0 - p_M)}{M}. \tag{14}$$

Then we suggest the L1-norm distance measure for Δ, as

$$\Delta = \sum_{k=1}^{M} \left| Q_k^O - Q_k^F \right| . \quad (15)$$

As a result of [5], there is a Δ calculated for every candidate parameter set. The parameter set with the best (smallest) Δ is the chosen fit, which is step [6].

Customizing the Method

One may notice that there is room to adjust the algorithm on the range (p_1, p_2), M, and the distance measure. Many simulations were run, and have found that generally the larger ranges (e.g. 0.01, 0.99), and a large number of points $M = 100$, provide superior recovery of parameters, and fit well the quantiles on real data applications. In addition, the L-1 norm distance is more resilient to outliers, and has been found in simulation studies, to recover better the 7th-9th deciles than the L-2 norm distance (squared differences), in the context of this method.

While the largest ranges (0.001, 0.999) are ideal in clean (simulated) data settings, smaller-adjusted ranges (0.01, 0.99) may help in real data cases that may have noisy or contaminant RTs, which are not indicative of the general distribution produced along the experimental condition, and/or may also worsen the fit of the overall RT distribution. The algorithm indeed contains an additional step [7], which performs [1]-[6] along a number of slightly adjusted ranges around (0.01, 0.99), and selects the better fit according to a criterion of 100 residual points within the bulk of the distribution, quantiles $(.05, .95)$. Thus the provided fitting approach may be easily customized in a number of ways: to explore additional combinations of (p_1, p_2), M, distance types, and criterions, according to the aims of the researcher.

Finally, it is worthwhile to note that the variety of additional ranges searched in [7] are close to (0.01, 0.99), e.g. (0.02, 0.98), because simulation studies have found that smaller ranges which do not seek to fit the lower quantiles, such as ranges between (0.10, 0.90), do not recover as well the parameters. This is principally because the leading edge parameter, θ, is not appropriately recovered, which strongly affects the recovery of the distribution shape parameters; and it is a similar case being that the right-tail is not fit either. Therefore, using only slightly smaller ranges may be advantageous when working against uncharacteristic outliers, while in contrast, much smaller ranges will negatively affect parameter recovery.

References

Anders, R., & Batchelder, W. H. (2013). Cultural consensus theory for the ordinal data case. *Psychometrika*. doi: 10.1007/s11336-013-9382-9

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 514.

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Current Directions in Psychological Science*, *20*(3), 160–166.

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*, *59*(4), 495–523.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). Wiley New York.

Basak, P., & Balakrishnan, N. (2012). Estimation for the three-parameter inverse Gaussian distribution under progressive type-ii censoring. *Journal of Statistical Computation and Simulation*, *82*(7), 1055–1072.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*(2), 117–133.

Burnett, H. G., & Jellema, T. (2013). (Re-)conceptualisation in Asperger's syndrome and typical individuals with varying degrees of autistic-like traits. *Journal of Autism and Developmental Disorders*, *43*(1), 211–223.

Carpenter, R. (1981). Oculomotor procrastination. *Eye Movements: Cognition and Visual Perception*, 237–246.

Casteau, S., & Vitu, F. (2012). On the effect of remote and proximal distractors on saccadic behavior: A challenge to neural-field models. *Journal of Vision*, *12*(12), 14.

Cheng, R., & Amin, N. (1981). Maximum likelihood estimation of parameters in the inverse Gaussian distribution, with unknown origin. *Technometrics*, *23*(3), 257–263.

Chhikara, R. (1988). *The inverse Gaussian distribution: Theory, methodology, and applications* (Vol. 95). CRC Press.

Chhikara, R. S., & Folks, J. L. (1974). Estimation of the inverse Gaussian distribution function. *Journal of the American Statistical Association*, *69*(345), 250–254.

Crow, E. L., & Shimizu, K. (1988). *Lognormal distributions: Theory and applications* (Vol. 88). M. Dekker New York.

Folks, J., & Chhikara, R. (1978). The inverse Gaussian distribution and its statistical application–a review. *Journal of the Royal Statistical Society. Series B (Methodological)*, 263–289.

Fréchet, M. (1927). Sur la loi de probabilité de lâĂŹécart maximum. In *Annales de la societe polonaise de mathematique* (Vol. 6, pp. 93–116).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (second ed.)*. Boca Raton, FL.: Chapman & Hall/CRC.

Gerstein, G. L., & Mandelbrot, B. (1964). Random walk models for the spike activity of a single neuron. *Biophysical Journal*, *4*(1), 41–68.

Goujon, A., & Fagot, J. (2013). Learning of spatial statistics in nonhuman primates: Contextual cueing in baboons (papio papio). *Behavioural Brain Research*, *247*, 101–109.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Gumbel, E. J., & Lieblein, J. (1954). *Statistical theory of extreme values and some practical applications: a series of lectures* (Vol. 33). US Government Printing Office Washington.

Hartendorp, M. O., Van der Stigchel, S., & Postma, A. (2013). To what extent do we process the nondominant object in a morphed figure? evidence from a picture–word interference task. *Journal of Cognitive Psychology*, *25*(7), 843–860.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience*, *35*(6), 2476–2484.

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 678–694.

Heathcote, A., Popiel, S. J., & Mewhort, D. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin*, *109*(2), 340.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*(4), 382.

Jeansonne, M. S., & Foley, J. P. (1991). Review of the exponentially modified Gaussian (emg) function since 1983. *Journal of Chromatographic Science*, *29*(6), 258–266.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*(1), 1.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307–321.

Koutrouvelis, I. A., Canavos, G. C., & Meintanis, S. G. (2005). Estimation in the three-parameter inverse Gaussian distribution. *Computational Statistics & Data Analysis*, *49*(4), 1132–1147.

LaBerge, D. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27*(4), 375-396. Retrieved from `http://dx.doi.org/10.1007/BF02289645` doi: 10.1007/BF02289645

Laming, D. R. J. (1968). Information theory of choice-reaction times.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1-15.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Link, S. W. (1992). *The wave theory of difference and similarity*. Psychology Press.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford University Press.

Lukacs, E. (1955). A characterization of the gamma distribution. *The Annals of Mathematical Statistics*, 319–324.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817.

Mulder, M., Van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences–a model-based review. *Neuroscience*, *277*, 872–884.

Nagatsuka, H., & Balakrishnan, N. (2013). A consistent method of estimation for the parameters of the three-parameter inverse Gaussian distribution. *Journal of Statistical Computation and Simulation*, *83*(10), 1915–1931.

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*(8), 1027–1046.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735.

Oravecz, Z., Anders, R., & Batchelder, W. H. (2013). Hierarchical Bayesian modeling for test theory without an answer key. *Psychometrika*. doi: 10.1007/s11336-013-9379-4

Padgett, W., & Wei, L. (1979). Estimation for the three-parameter inverse Gaussian distribution. *Communications in Statistics-Theory and Methods*, *8*(2), 129–137.

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*(1), 53.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, *111*(2), 333.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contami-