Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package

ANDREW HEATHCOTE
University of Newcastle, Callaghan, New South Wales, Australia

Schwarz (2001, 2002) proposed the ex-Wald distribution, obtained from the convolution of Wald and exponential random variables, as a model of simple and go/no-go response time. This article provides functions for the S-PLUS package that produce maximum likelihood estimates of the parameters for the ex-Wald, as well as for the shifted Wald and ex-Gaussian, distributions. In a Monte Carlo study, the efficiency and bias of parameter estimates were examined. Results indicated that samples of at least 400 are necessary to obtain adequate estimates of the ex-Wald and that, for some parameter ranges, much larger samples may be required. For shifted Wald estimation, smaller samples of around 100 were adequate, at least when fits identified by the software as having ill-conditioned maximums were excluded. The use of all functions is illustrated using data from Schwarz (2001). The S-PLUS functions and Schwarz's data may be downloaded from the Psychonomic Society's Web archive, www. psychonomic.org/archive/.

Schwarz (2001) proposed the ex-Wald distribution, obtained from the convolution of Wald and exponential random variables, as a model of response time (RT) distribution, particularly in simple RT and go/no-go choice RT tasks. This article provides functions for the powerful commercially available S-PLUS package that produce maximum likelihood (ML) estimates (Edwards, 1972) for the parameters of the Wald and ex-Wald distributions and of another distribution based on the Wald, the shifted Wald. Sufficient mathematical detail is also given in this article to allow implementation in another suitable package. The efficiency and bias of the estimates obtained for the Wald distributions are examined in a Monte Carlo study, and the use of the functions is illustrated using data from Schwarz's (2001) Participant A.

Ex-Wald Convolution Components

Convolution provides the distribution of the sum of independent random variables and, so, is useful in modeling RTs that result from the sum of stages with independent and random completion times. Schwarz (2001) modeled decision time with a Wald random variable ($W > 0$) and nondecision time with an exponential random variable ($X \geq 0$), with RT being the sum of these two times: RT = W⊕X.

An exponential random variable X has a density of $f_E(x \mid t) = \frac{1}{t} e^{-x/t}$ and a cumulative density of $F_E(x \mid t) = 1 - e^{-x/t}$, where t is the mean (κ_1) of X.[1] The exponential distribution has a variance of $\kappa_2 = t^2$, and a third central moment [i.e., $\int (x - \kappa_1)^3 f(x)dx$, a measure of skew] of $\kappa_3 = 2t^3$. The exponential density (*dexp*) and cumulative density (*pexp*) functions are built-in features of S-PLUS, as is a function that generates exponentially distributed samples (*rexp*). For all of the distributions discussed here, density, cumulative density, and random functions are provided. Function definitions are given in Table 1, with function names given in italics (e.g., *dwald*) in the body of this article.

The Wald distribution can be derived from a sequential sampling evidence accrual model of the decision process in simple and go/no-go RT, since it is the distribution of first passage times (W) through a level $a > 0$ of a space and time homogenous Wiener diffusion process, with an initial value of zero, a drift of $m > 0$, and a variance of $\sigma^2 > 0$ (see Luce, 1986, for a discussion of sequential sampling models). In the RT context, a is a response threshold, and m is the mean rate of evidence accrual. Without loss of generality, one of the Wald parameters can be fixed, so we use the conventional normalization $\sigma = 1$ and drop this parameter in the following.

The Wald distribution has a density (*dwald*) of

$$f_W\left(w \mid m,a\right) = \frac{a}{\sqrt{2\pi w^3}} \exp\left[-\frac{\left(a - mw\right)^2}{2w}\right], \quad w > 0. \quad (1)$$

The Wald cumulative density function cannot be expressed in closed form but can be expressed in terms of the cumulative normal distribution function, Φ (built-in

Table 1
Function Definitions

	Comment
Density functions	
dwald(w,m,a,s=0)	shifted Wald, w is RT
dexw(r,m,a,t)	ex-Wald, r is RT
dexg(x,m,s,t)	ex-Gaussian, x is RT, m = μ, s = σ and t = τ
Cumulative density functions	
pwald(w,m,a,s=0)	p(W ≤ w) for shifted Wald
pexw(r,m,a,t)	p(R ≤ r) for ex-Wald
pexg(x,m,s,t)	p(X ≤ x) for ex-Gaussian
Random functions (n is number of samples)	
rwald(n,m,a,s = 0)	shifted Wald
rexw(n,m,a,t)	ex-Wald
rexg(n,m,s,t)	ex-Gaussian
Objective functions (minus loglikelihood, x is RT)	
negllswald(p,x)	shifted Wald, p = (m,a,s)
negllwald(p,x)	Wald with s = 0, p = (m,a)
negllexw(p,x)	ex-Wald, p = (m,a,t)
negllexg(p,x)	ex-Gaussian, p = (m,s,t)
Objective derivatives	
negllswald.gh(p,x)	shifted Wald gradient and Hessian
negllwald.gh(p,x)	gradient and Hessian for Wald with s = 0
negllexggrad(p,x)	gradient for ex-Gaussian
Heuristic starting points (p = heuristic parameter)	
waldstpt(x,shift=T,p=0.9)	Wald/shifted Wald
exwstpt(x,p = 0.5)	ex-Wald
exgstartpt(rt, p = 0.8)	ex-Gaussian
Fitting functions (rt is a response time vector)	
fitwald(rt,shift=T,p=0.9,	Wald (shift=F) and shifted Wald (shift=T)
start=waldstpt(rt,shift,p))	
fitexw(rt,p =.5,start=	ex-Wald, scaleit = T turns on scaling =
exwstpt(rt),scaleit=T)	1/start
fitexg(rt,p=0.8,start=	ex-Gaussian
exgstartpt(rt,p))	
Auxiliary functions	
uandv(x,y,firstblock=20,	Used by dexw, series approximation to the
block=0,tol=.Machine$double.	real (u) and imaginary (v) parts of
eps^(2/3),maxseries=20)	complex error function, erf(x + iy)
rew(x,y,...)	Used by dexw, real part of
	w(z) = exp(−z^2)[1−erf(−iz)]
ser.nlminb(object,tol =	Modified vcov.nlminb, object = fit object
1,scale=object$scale,eps=0.0	produced by nlminb, called within fitting
01,eps0=1)	functions.
chisq(rt,p,bin=25, minn = 5,	Chi square, bin = bin width, minn =
dist = "wald")	minimum expected observations in a bin,
	dist = "wald," "exw" (ex-Wald) and "exg"
	(ex-Gaussian), called in fitting functions.

function *pnorm* in S-PLUS), which has efficient numerical approximations:

$$F_W\left(w \mid m,a\right) = \Phi\left(\frac{mw - a}{\sqrt{w}}\right) + e^{2am}\Phi\left(-\frac{mw + a}{\sqrt{w}}\right). \quad (2)$$

Schwarz (2001) notes that the second summand in Equation 2 can cause numerical errors when the arguments to the exponential and Φ functions are large. The *pwald* function avoids numerical errors, in this case, by using Derenzo's (1977) approximation to Φ, followed by an algebraic simplification of the second summand, as was suggested by Schwarz.[2] An S-PLUS function that provides random samples from a different parameterization of the Wald distribution is available on the Web.[3]

The *rwald* function is based on this function but uses the parameterization in Equations 1 and 2.

The Wald distribution has a mean of a/m, a variance of $\kappa_2 = a/m^3$, and a third central moment of $\kappa_3 = 3a/m^5$. It is evident from the moments that a is a scale parameter and that m affects all moments. The leading edge of the Wald distribution is fixed at zero (this is also true of the ex-Wald), which makes it unrealistic as an exact model of RT, due to the minimum times required for postdecision processes. In simple RT, for example, Smith (1995) assumed a substantial shift (relative to overall RT) of around 66 msec. This defect can be corrected by adding a parameter, s, which shifts the leading edge and causes the mean to become $s + a/m$ but leaves the other central moments unaffected. The Wald functions in Table 1 include

the shift parameter, which is set to zero by default. Equations 1 and 2 can be parameterized to include the shift by substituting $(w - s)$ for w. For the random function, the shift parameter is added to the output.

The shift parameter provides an alternative way of modeling postdecision time, as a constant rather than as an exponentially distributed random variable in Schwarz's (2001) ex-Wald model. A constant for postdecision time is also unrealistic as an exact model of RT, since postdecision times are variable. Smith (1995) reviewed evidence that variability in postdecision times had a standard deviation of, at most, 14 msec, and likely substantially less. For Schwarz's (2001) Participant A data, analyzed below, this means that less than 5% of the variability in RT is due to postdecision variability. Hence, a constant nondecision time may provide a good approximation; it was the approximation adopted by Smith, and the same assumption is commonly made in fitting sequential sampling evidence accrual models of choice RT (e.g., Ratcliff & Rouder, 1998).

The Ex-Wald Distribution

Schwarz (2001) derives the density (function *dexw*) of an ex-Wald random variable (R) for the case in which $m^2 \geq 2/t$. Defining $k = \sqrt{(m^2 - 2/t)}$,

$$f_{EW}(r \mid m,a,t) = \frac{1}{t} e^{-[r/t - a(m-k)]} F_W(r \mid k,a). \quad (3)$$

Schwarz (2001) states that this case is of most interest in applications to RT. However, in practice, it is useful not to have to assume this bound, so the functions given here also work when $m^2 < 2/t$.

Schwarz (2002) examined the case in which $m^2 < 2/t$, which requires complex function theory. In this case, the ex-Wald density involves the real part of the complex function, $w(z) = e^{-z^2}[1 - erf(-iz)]$, where $z = x + iy$ is a complex number with real part x and imaginary part y and $erf(z)$ is the complex error function. The real part of $w(z)$ (function *rew*) can be derived from the real and imaginary parts of $erf(z) = u(x,y) + i\, v(x,y)$. Schwarz favors this method since series approximations exist for $u(x,y)$ and $v(x,y)$ that converge quickly for all x,y (Abramowitz & Stegun, 1965, 7.1.29; function *uandv*[4]). Defining $k' = \sqrt{(2/t - m^2)}$, see the equation at the bottom of the page.

For both cases, the cumulative density (function *pexw*) can be derived from a general result for convolutions involving exponential random variables (e.g., Ashby & Townsend, 1980), yielding an expression in terms of the Wald cumulative density and the ex-Wald density:

$$F_{EW}(r \mid m,a,t) = F_W(r \mid m,a) - t \times f_{EW}(r \mid m,a,t). \quad (4)$$

Samples from the ex-Wald distribution are obtained by summing samples from the Wald and exponential distri-

butions (function *rexw*). The ex-Wald has a mean of $t + a/m$, a variance of $\kappa_2 = t^2 + a/m^3$, and a third central moment of $\kappa_3 = 2t^3 + a/m^5$.

MAXIMUM LIKELIHOOD ESTIMATION

The log of the likelihood of a distribution f with parameter vector θ, given a data vector **RT** of length n, is

$$l(\boldsymbol{\theta} \mid \mathbf{RT}) = \sum_{i=1}^{n} \ln\left[f\left(RT_i \mid \boldsymbol{\theta}\right)\right]. \quad (5)$$

ML parameter estimates are obtained by finding the maximum of Equation 5 or, equivalently, by minimizing minus Equation 5 (functions *negllwald*, *negllswald*, and *negllexw* for the Wald, shifted Wald, and ex-Wald, respectively).

For shifted Wald and ex-Wald distributions, the minimum cannot be obtained analytically but, instead, must be found by a search function. The search function is not guaranteed to find the minimum of the "objective" function (i.e., *negllwald*, *negllswald*, or *negllexw*). A number of arguments to the search function must be appropriately set in order to gain good performance. These include bounds on the allowable parameter estimates, a starting point for the search, and ideally, functions to calculate the derivatives of the objective function and the relative scales of each parameter.

Fitting Functions

The functions *fitwald* and *fitexw* use the S-PLUS minimization *nlminb* function to obtain maximum likelihood estimates. For example, typing *fit ← fitwald(rt)* at the command prompt (note that ← is typed as a less than sign followed by a minus sign and should be read as "gets") in S-PLUS creates a list object called *fit* (any name may be used; *fit* will be used generically in the following to indicate the list object created by fitting). The *fit* object contains the estimated parameters for the Wald, given a data vector, *rt*, as well as a number of other components. The components can be referenced using the syntax *name$component*. For example, *fit$parameters* is a vector of length three containing the estimated parameters in the order (m,a,s) for the shifted Wald and (m,a,t) for the ex-Wald.

For both the Wald and the ex-Wald distribution, all parameters are bounded below by zero.[5] For the Wald distribution, the shift (s) parameter must be less than the minimum observed RT [min(**RT**)]. A starting point for search can be supplied by the user as the *start* argument to the fitting functions. For example, *fitwald(rt,start= c(0.2,70,100))* sets the starting point to $m = 0.2$, $a = 70$, and $s = 100$. If no *start* argument is supplied, the starting point is calculated automatically by the *waldstpt* or *exwstpt* functions. Both functions estimate one parameter, using

$$f_{EW}(r \mid m,a,t) = \frac{1}{t} e^{-\left[(a-m)^2 / (2r)\right]} \text{Re}\left[w\left(k'\sqrt{r/2} + ia/\sqrt{2r}\right)\right]$$

a heuristic, and then solve for the other two parameters, using the first two sample moments. Starting estimates based on the first three sample moments were found to perform poorly due to the high sampling variance of the third moment. For the Wald distribution, the shift parameter is estimated by the $s = p \times \min(\mathbf{RT})$ heuristic, with $p = .9$ by default. For the ex-Wald distribution, the exponential parameter is estimated by $t = p \times SD(\mathbf{RT})$, where SD is the sample standard deviation and $p = .5$ by default. The default values of p may be altered by specifying a single number as the start argument. For example, *fitwald(rt, start=waldstpt(rt=rt,p=.95))* sets p to .95.

The *nlminb* search function is faster and converges more reliably if the gradient (first derivatives) vector and Hessian (second derivatives) matrix of the likelihood function are specified. Approximate standard errors and correlations among the parameter estimates can be obtained by inverting the Hessian (Bates & Watts, 1988). The gradient vector and the Hessian matrix for minus the loglikelihood are

$$-\frac{\partial l\left(\mathbf{RT} \mid \boldsymbol{\theta}\right)}{\partial \theta_j} = -\sum_{i=1}^{n} \frac{\partial \ln\left[f\left(RT_i \mid \boldsymbol{\theta}\right)\right]}{\partial \theta_j} \qquad (6)$$

and

$$-\frac{\partial^2 l\left(\mathbf{RT} \mid \boldsymbol{\theta}\right)}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{n} \frac{\partial^2 \ln\left[f\left(RT_i \mid \boldsymbol{\theta}\right)\right]}{\partial \theta_j \partial \theta_k}. \qquad (7)$$

For the shifted Wald distribution, the first and second derivatives of $\ln(f_W)$ have straightforward analytic solutions, which are given in Table 2. The corresponding results for the Wald without a shift are obtained by setting $s = 0$. The *negllwald.gh* and *negllswald.gh* functions calculate the gradient and Hessian for the Wald and the shifted Wald distributions, respectively, using these

Table 2
First and Second Derivatives of $\ln(f_W)$

$$\partial \ln\left(f_W\right)/\partial m = a - m\left(w - s\right)$$

$$\partial \ln\left(f_W\right)/\partial a = m + 1/a - a/\left(w - s\right)$$

$$\partial \ln\left(f_W\right)/\partial s = \tfrac{3}{2}\left(w - s\right)^{-1} - \tfrac{a^2}{2}\left(w - s\right)^{-2} + m^2/2$$

$$\partial^2 \ln\left(f_w\right)/\partial m^2 = s - w$$

$$\partial^2 \ln\left(f_w\right)/\left(\partial m \partial a\right) = 1$$

$$\partial^2 \ln\left(f_w\right)/\left(\partial m \partial s\right) = m$$

$$\partial^2 \ln\left(f_w\right)/\partial a^2 = -1/a^2 - 1/\left(w - s\right)$$

$$\partial^2 \ln\left(f_w\right)/\left(\partial a \partial s\right) = -a/\left(w - s\right)^2$$

$$\partial^2 \ln\left(f_W\right)/\partial s^2 = \left[\tfrac{3}{2} - a^2\left(w - s\right)^{-1}\right]/\left(w - s\right)^2$$

equations. The *fitwald* function calls these functions to aid fitting, and they are also used by the *ser.nlminb* function, described in more detail below, to calculate approximate parameter standard errors and correlations. For the ex-Wald distribution, the gradient and Hessian are not so tractable. The *nlminb* function can still fit the ex-Wald distribution, using numerical finite differences, but it tends to be slower.

Several conditions are used by *nlminb* to terminate search, as reported in *fit$message*. Normal termination occurs when the objective function decreases by only a small proportion, by default $\max(10^{-10}, .\text{Machine\$double.eps}^{2/3})$, where .Machine$double.eps is the double precision floating point tolerance of the computer on which S-PLUS is running. Achieving this criterion is reported as "Relative Function Convergence" by S-PLUS. Convergence can also occur when parameter estimates change by only a small amount between iterations of the search (.Machine$double.eps$^{1/2}$ by default), called "X Convergence." Abnormal termination, indicating a failure of the minimization, occurs for a number of reasons, including exceeding a set number of objective function evaluations (200 by default) or search iterations (150 by default), or because the size of the finite differences used during search has become very small ($100 \times$ .Machine$double.eps by default), called "False Convergence."

In the Monte Carlo study, false convergence was a common problem with fits of the ex-Wald, even when search was started at the true parameter values. Greatly improved performance in this regard was obtained by setting the relative scales of the different parameters to the inverse of the starting point estimate, using the *scale* argument of *nlminb*. For example, for starting point (0.2, 70, 100), *scale=(1/0.2,1/70, 1/100)*. The *fitexw* function uses this setting of *scale* by default, but this can be turned off using the argument *scaleit=F*. Scale does not need to be set for *fitwald*, since it is calculated by default from the Hessian. More rarely, abnormal termination occurs because the function evaluation and search iteration limits are reached. This problem was avoided in the Monte Carlo study by doubling their default values in *fitexw*. All default values for *nlminb* can be changed using the S-PLUS *nlminb.control* function.

The parameter variance–covariance matrix for the ex-Wald and, hence, approximate parameter standard errors and correlations can be obtained by numerical finite difference methods, using the *vcov.nlminb* function of the MASS library [Venables & Ripley, 1999; the library is distributed with S-PLUS but must be loaded using the *library(MASS)* command]. The *vcov.nlminb* function can also make use of analytic derivatives if they are supplied to *nlminb*. For convenience, all fitting functions call an augmented version of *vcov.nlminb*, called *ser.nlminb*, which outputs parameter correlations and standard errors and protects against numerical errors. It creates *fit$se*, a vector of length three containing the standard error estimates, in the same order as the *fit$parameters*.

Similarly it creates, *fit$cor*, a vector of length three that contains the parameter correlations, in the order $r(m,a)$, $r(m,s)$, $r(a,s)$ for the Wald and $r(m,a)$, $r(m,t)$, $r(a,t)$ for the ex-Wald. For some fits, *ser.nlminb* sets standard errors and correlations to missing [*fit$se* and *fit$cor* = (NA, NA, NA), where NA is the S-PLUS missing value symbol], either because the Hessian matrix cannot be inverted, since it is singular, or because variance estimates are negative.

MONTE CARLO STUDY

A Monte Carlo study was performed with two aims: to refine the implementation of the S-PLUS functions so that they are numerically robust for small samples and to determine the bias and efficiency of parameter estimates. A variety of parameter values were selected that spanned the range usually encountered in applications and, also, theoretically important cases, particularly for the ex-Wald.
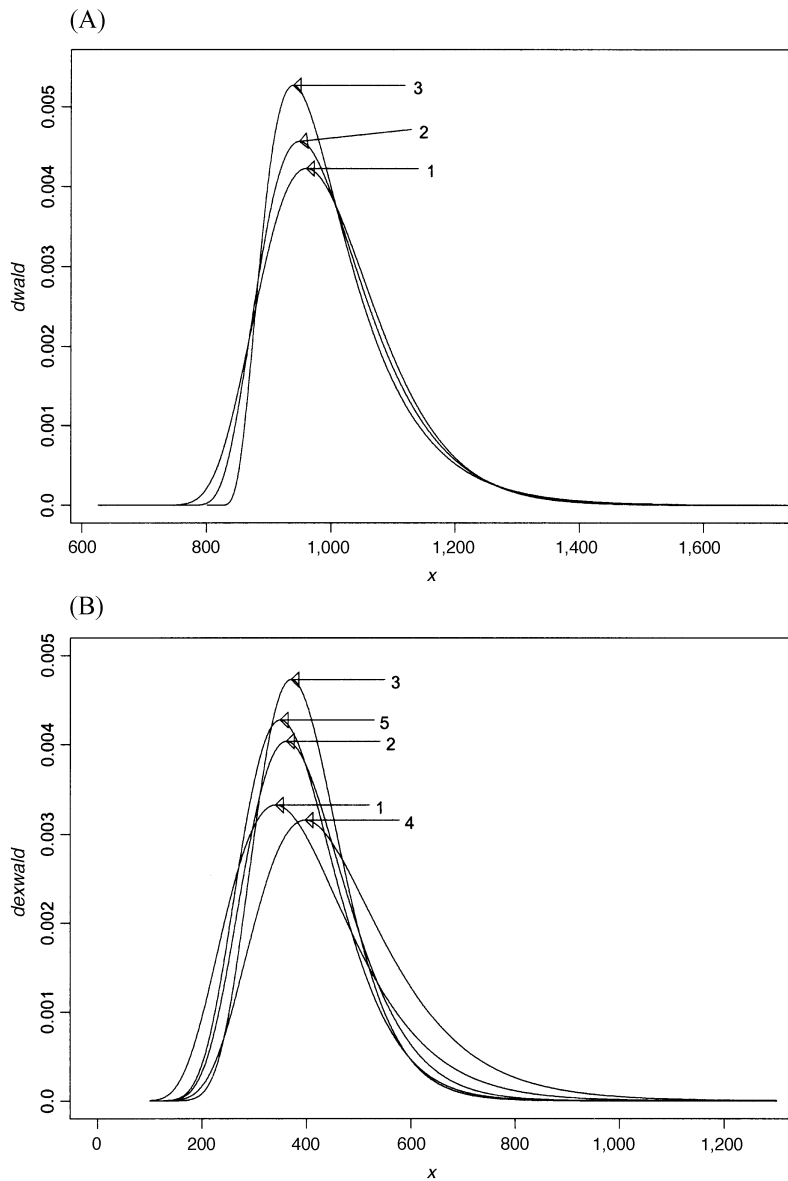
Figure 1. (A) Shifted Wald and (B) ex-Wald distributions used in the Monte Carlo study. In panel A, mean and standard deviation are constant, and skew increases with the number used to identify the distribution. In panel B, Distributions 3 and 4 are in the real case studied by Schwarz (2001) and Distributions 1 and 5 are in the complex case studied by Schwarz (2002), with Distribution 2 on the boundary.

Table 3
Shifted Wald Parameters and Percentages of Irregular Fits

	Distribution		
	1	2	3
m	0.18856	0.16262	0.14142
a	70.711	44.721	28.284
s	625	625	625
Mean	1,000	1,000	1,000
SD	102.7	102	100
$\mu_3^{1/3}$	66.6929	73.2623	79.3711
Fisher skew	0.2739	0.3708	0.5000
Irregular fits, $n = 40$	12.3	5.0	1.4
Irregular fits, $n = 60$	9.4	2.4	0.0
Irregular fits, $n = 80$	6.4	0.6	0.0
Irregular fits, $n = 100$	4.6	0.4	0.0
Irregular fits, $n = 120$	2.4	0.1	0.0

Results for the shifted Wald are described first, followed by results for the ex-Wald.

Figure 1 shows the three shifted Wald distributions investigated in the Monte Carlo study, and Table 3 supplies the corresponding parameters. Parameters were chosen so that the distributions were equated on location (mean = 1,000) and scale ($SD = 100$) but differed in skew. One thousand samples of sizes 40 to 120 were fit for each distribution. All shifted Wald fits in the Monte Carlo study used the analytic gradients and Hessian. Without them, some fits did not converge normally, and estimates were sensitive to starting points, whereas in the Monte Carlo study all fits converged normally and produced identical parameter estimates for starting points generated by *waldstpt*, using $p = .9, .95,$ and $.99$.

As is shown in Figure 1, for the Wald distributions examined here, skew was determined mainly by the left tail, with right tails having very similar shapes for all the distributions. The longer left tail of Distribution 1 caused the greatest problems with parameter estimates. Table 3 shows that Distribution 1 produced the greatest percentage of irregular fits (i.e., fits where the Hessian either was singular or produced negative parameter variance estimates). Irregular fits overestimated the m and a parameters and underestimated the s parameter. The underestimates would have been much worse if s had not been bounded below by zero.

Figure 2 shows the bias and efficiency of the parameter estimates as a function of sample size and distribution. Bias was measured by the mean deviation of the estimates from the true parameter values. Efficiency was measured by the standard deviation of the estimates. Bias decreases and efficiency increases with sample size and as the distributions became more skewed. Much better estimates were obtained by removing irregular fits, resulting in bias that was relatively minor even for the smallest sample sizes. However, parameter estimates were still quite variable for smaller sample sizes, particularly for the s and a parameters.

Figure 1 also shows the five ex-Wald distributions investigated in the Monte Carlo study. Parameters, which

are given in Table 4, were chosen to examine the effect of varying k^2: Schwarz's (2001) real solution applies for $k^2 \geq 0$, whereas solutions are complex for $k^2 < 0$. Distribution 2 was chosen to be a reference with $k^2 = 0$. Negative values of k^2 were produced by decreasing m (Distribution 1) or t (Distribution 5). Positive values of k^2 were produced by increasing m (Distribution 3) or t (Distribution 4). Where k^2 was manipulated by changing m (Distributions 1 and 3), a was also altered to maintain a constant mean. For these parameters, the second and third central moments decreased as k^2 increased when manipulated via m, whereas they increased when k^2 was manipulated via t. Skew, as indicated by Fisher's skewness measure, $\gamma = \mu_3/\mu_2^{3/2}$, always increased with k^2.

One thousand fits were performed at each sample size in the Monte Carlo study of the ex-Wald distribution. These fits produced bimodal and, in some cases, multimodal parameter distributions. Initially, parameter sampling distributions were obtained for $n = 120$ to 400. In order to check whether sampling distributions become unimodal for large sample sizes, sampling distributions were obtained for $n = 10,000$. For Distributions 3 and 4, where the true value of $k^2 > 0$, estimates of t were bimodal for $n = 400$, but unimodal and relatively symmetric for $n = 10,000$. Note that when the fitting algorithm was restricted to searching parameter combinations in which $k^2 > 0$, the sampling distributions for $n = 400$ were very similar to the unrestricted case, so bimodality was not due to allowing fits to sample the complex case.

For distributions in which the true value of $k^2 \leq 0$, sampling distributions for m and t were not unimodal even for $n = 10,000$. Restricting the fitting algorithm to search for parameter combinations in which $k^2 < 0$ did not help matters for Distributions 1 and 5, where the true value of $k^2 < 0$, producing sampling distributions that were shaped similarly to the unrestricted case. Sampling distributions for the three cases in which $k^2 \leq 0$ were obtained for $n = 50,000$. For Distribution 2, where the true value of $k^2 = 0$, unimodality was not obtained even for $n = 50,000$. For Distribution 1, sampling distributions were fairly unimodal for $n = 50,000$, although a long right tail remained for the m parameter. For Distribution 5, sampling distributions for m and t were trimodal for $n = 50,000$, with a sharp third peak appearing at the true parameter value.

Fitting was also less tractable for the ex-Wald than for the shifted Wald. Table 4 shows that, in many cases, fits were irregular. In contrast to the shifted Wald, excluding these fits did not reduce bias or increase efficiency. For Distributions 1 and 2, most fits were well behaved for $n = 400$. For the other distributions, a substantial proportion of fits had problems for $n = 400$. These problems largely disappeared for Distributions 3 and 4, where the true value of $k^2 > 0$, when $n = 10,000$. However, problems reappeared for Distribution 2 when $n = 10,000$, and both Distributions 2 and 5 still had a substantial proportion of irregular fits even for $n = 50,000$. Note that all ex-Wald fits reported here used the true parameter
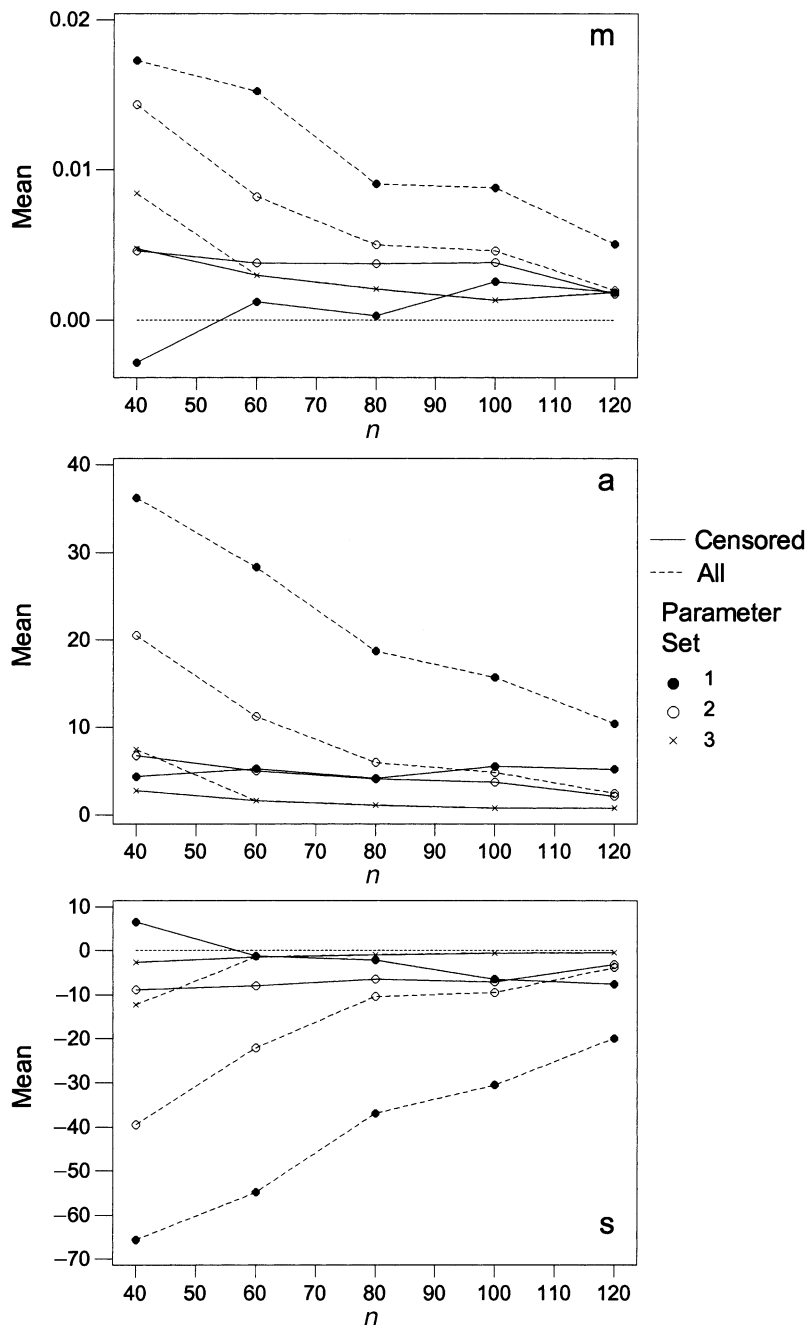
Figure 2. Shifted Wald parameter estimates of mean bias (estimated − true) and efficiency (parameter standard deviations, *SD*) for all fits and for fits with ill-conditioned Hessian estimates censored. The abscissa is sample size (*n*), and the panels represent results for the *m* (mean rate of evidence accrual), *a* (response threshold), and *s* (shift or minimum) parameters.

values as the starting point for search. When starting values were generated using the default heuristic in *exwstpt*, similar results were obtained, but with a slightly increased percentage of irregular fits.

Despite irregularity in fits and multimodality in parameter estimate distributions, Figure 3 shows that bias

decreased and efficiency increased with increases in sample size for all the distributions. However, overall performance was poor relative to results for the shifted Wald, and for the *t* parameter, in particular, estimates remained variable even for the very large samples. Distribution 2 showed the greatest bias, and for smaller sam-
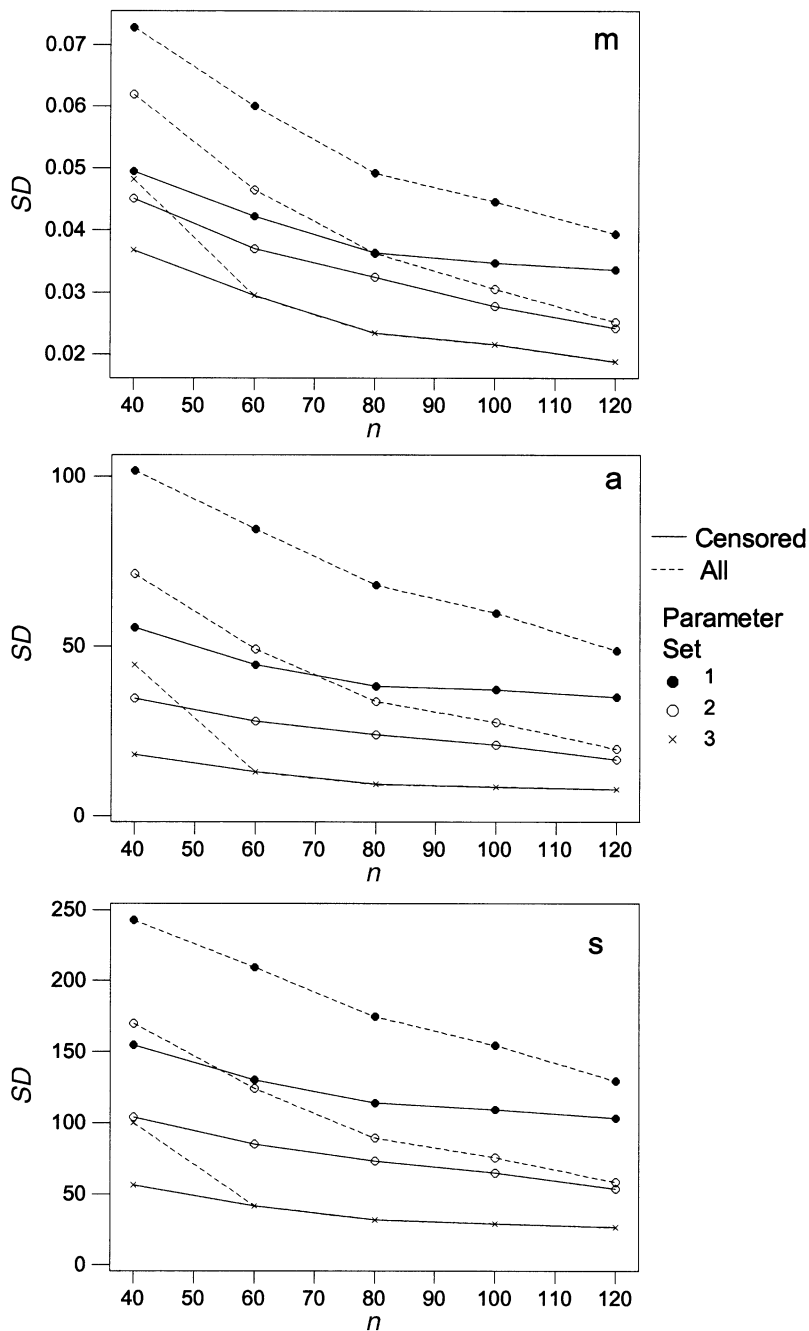
Figure 2 (Continued).

ples the bias for the *t* parameter was in the opposite direction to the other distributions. Distributions 1 and 4 produced particularly variable estimates of *t* for smaller samples. Although variability decreased quickly with increases in sample size for Distribution 4, the decrease was much slower for Distribution 1.

The results of the Monte Carlo study indicate that maximum likelihood estimation in finite samples from the shifted Wald and ex-Wald distributions does not always produce the unimodal normally distributed parameter estimate distributions observed in regular cases, particularly in smaller samples and particularly for the ex-Wald. It is important to note that multimodal parameter estimate distributions were not due to failures of search because of convergence to a local minimum. The same parameter estimates were obtained for a range of starting points, and profiles of the likelihood function (e.g., Figure 6) for particular samples had only one minimum. It

Table 4
Ex-Wald Parameters and Percentages of Irregular Fits

	Distribution				
	1	2	3	4	5
m	0.15	0.2	0.25	0.2	0.2
a	52.5	70.0	87.5	70.0	70.0
t	50.0	50.0	50.0	115.0	35.0
k^2	-0.0175	0	0.0225	0.0226	-0.0171
Mean	400	400	400	465	385
SD	134.371	106.066	90	148.24	99.875
$\mu_3^{1/3}$	98.006	77.681	69.768	148.284	67.276
Fisher skew	0.06292	0.07111	0.08171	0.08216	0.06740
Irregular fits, $n = 120$	10.4	74.5	26.2	41.1	20.4
Irregular fits, $n = 160$	42.0	20.5	18.6	17.9	11.7
Irregular fits, $n = 200$	40.3	1.5	17.5	4.9	10.6
Irregular fits, $n = 300$	4.9	0.3	17.7	13.0	15.4
Irregular fits, $n = 400$	0.4	0.2	18.2	11.1	16.2
Irregular fits, $n = 10{,}000$	0.1	46.7	0.5	0.0	6.1
Irregular fits, $n = 50{,}000$	4.3	10.1	–	–	15.2

was also not due to failures in the numerical approximations used here. Benchmarking against numerical integration showed these approximations to be accurate and slightly more accurate in some cases than Schwarz's (2001) approximation (see note 3).

Multimodality in likelihood estimates is not unexpected theoretically; it occurs when distribution functions do not fulfill regularity conditions (Edwards, 1972). Evidently, likelihood estimation of the shifted Wald and, particularly, the ex-Wald is irregular. However, Figures 3 and 4 display consistency, in that both the bias and the standard deviation of parameter estimates decreased with increased sample size. Hence, estimation based on likelihood remains useful in practice, as long as sample sizes are sufficiently large.

EXAMPLE APPLICATION

The application of the fitting functions is illustrated in this section, using the data from Participant A shown in Figure 3 of Schwarz (2001) and reproduced in Figure 4 here. Schwarz (2001) also examined the fit of the ex-Gaussian distribution to these data. As is shown in Table 1, S-PLUS functions are supplied to fit the ex-Gaussian, using similar notational conventions to the Wald and ex-Wald functions. Analytic gradients are supplied by *negllexggrad*. Maximum likelihood estimation is regular for the ex-Gaussian, and gradients were sufficient for good performance. Heathcote's (1996) ex-Gaussian estimation program, RTSYS, can use analytic Hessians in fitting and performed similarly for these data. The *exgstpt* function estimates starting points, using heuristics given in Heathcote, and usually performs well with the default values.

Schwarz's (2001) data come from a go/no-go paradigm, so the decision process might reasonably be modeled by a one-barrier sequential sampling (Wiener) process and, hence, a Wald distribution. Table 5 contains the parameter estimates for the ex-Wald obtained from *fitexw*, and the estimated distribution functions are plotted

as solid lines in Figure 4. As is predicted by the sequential sampling model, manipulation of the probability of a go trial ($p = .5$ or $.75$) mainly effects the a (decision criterion) parameter, whereas manipulation of numerical distance ($d = 1$ or 4) mainly effects the m (strength of evidence) parameter.

The exponential component of the ex-Wald models nondecision times, and corresponding parameter estimates (t) do not vary much across conditions, as was expected. Estimates of the t parameter vary slightly from those reported by Schwarz (2001), especially in the $p = .75$, $d = 1$ condition, where the fit at the minimum was so ill conditioned that the Hessian could not be inverted. Some small differences are to be expected due to variation in the details of the numerical approximations and search algorithms used. Such numerical differences are magnified when the minimum of the objective function is ill defined, as appears to be the case for the $p = .75$, $d = 1$ condition. Note also that the parameters found by *fitexw* are in the complex case (i.e., $m^2 < 2/t$) examined by Schwarz (2002).

In order to explore the $p = .75$, $d = 1$ case further, Figure 5 plots profiles of the objective function (*negllexw*) in the region of the solution by holding two parameters fixed at their estimated values and varying the third. The plots can be interpreted like a plot of sum-squared error in least squares regression, with smaller values of minus loglikelihood, which is sometimes called the *deviance*, indicating a better fit. Note that the deviance functions are smooth as changes in m and t cause a change from the real to the complex solutions. The minimum is sharply defined as a function of m and is fairly well defined for a, and in both cases the minimum is approximately quadratic in shape, as is assumed by the standard error and correlation estimates based on the Hessian. However, the minimum as a function of t is very poorly defined and is clearly nonquadratic, since deviance rises dramatically as t approaches zero. Similar, although less marked, behavior for t was found in the other conditions, and in these cases the estimates of t are within two standard er-
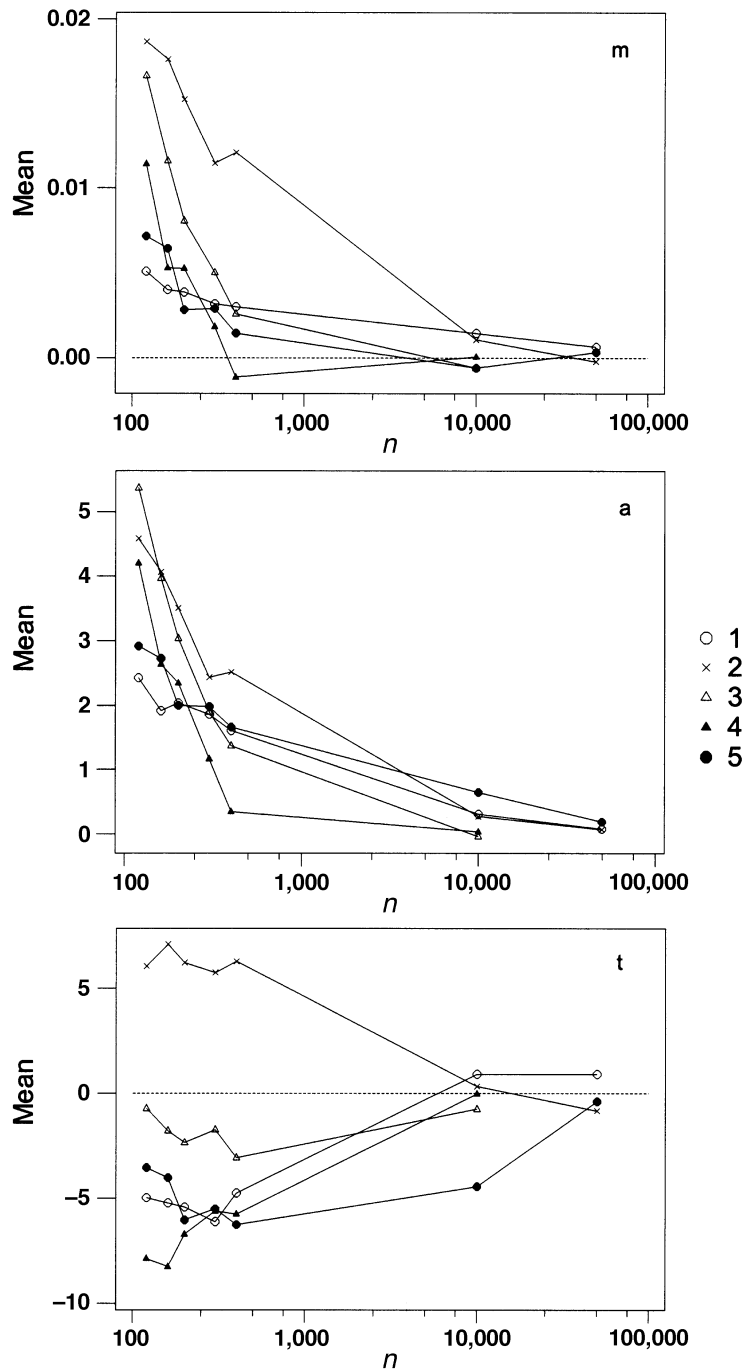
Figure 3. Ex-Wald parameter estimates of mean bias (estimated − true) and efficiency (parameter standard deviations, *SD*). The abscissa is sample size (*n*), on a $\log_{10}$ scale. The panels represent results for *m* (mean rate of evidence accrual), *a* (response threshold), and *t* (exponential mean) parameters.

rors of zero, indicating that there is no strong evidence for an exponential component.

Table 6 contains the results of fits of the shifted Wald, using *fitwald*. The deviance values indicate that fit is very slightly better than that for the ex-Wald and the Hes-

sian is well behaved for all the conditions. However, standard errors are quite large, especially for the estimated shift (*s*) parameters, which fall within one standard error of zero. Hence, there is no evidence that the shift parameter is nonzero. Figure 4 plots the best-fitting
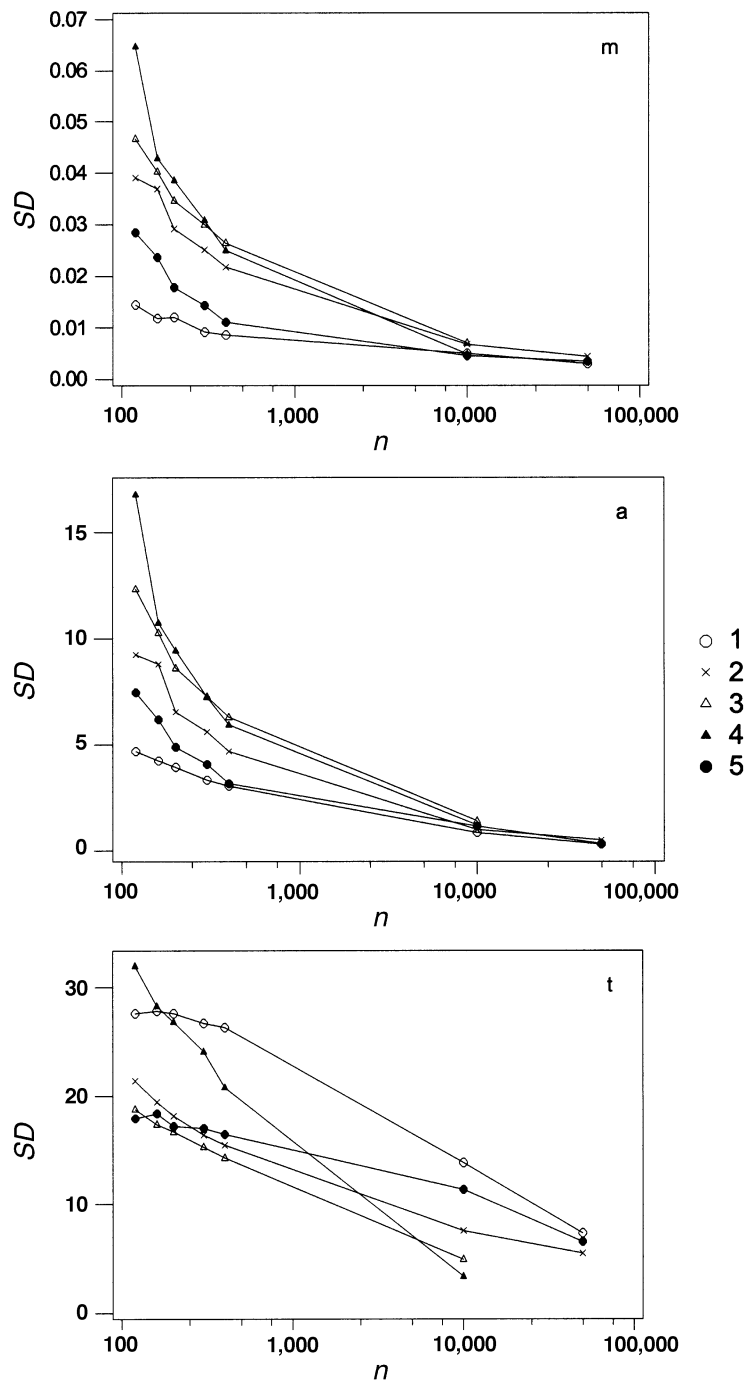
Figure 3 (Continued).

shifted Wald densities as dashed lines; in most cases, they are indistinguishable from the ex-Wald densities, except for a very slight negative shift on the left of the peak.

Since standard errors for both the shifted Wald and the ex-Wald models provide little evidence for a reliable non-decision (s or t) component, the simpler two-parameter Wald distribution, with shift fixed at zero, was fit. The resulting deviance values (Table 7) are only slightly larger than those for the three-parameter distributions. Since the Wald distribution is a special case of (i.e., is nested within) the shifted Wald and ex-Wald distributions, the reliability of the reduction in deviance can be determined using a likelihood ratio test. Annotating the deviance for the nested (Wald) model as d_n and the deviance of the full (shifted Wald or ex-Wald) model as d_f, the likelihood
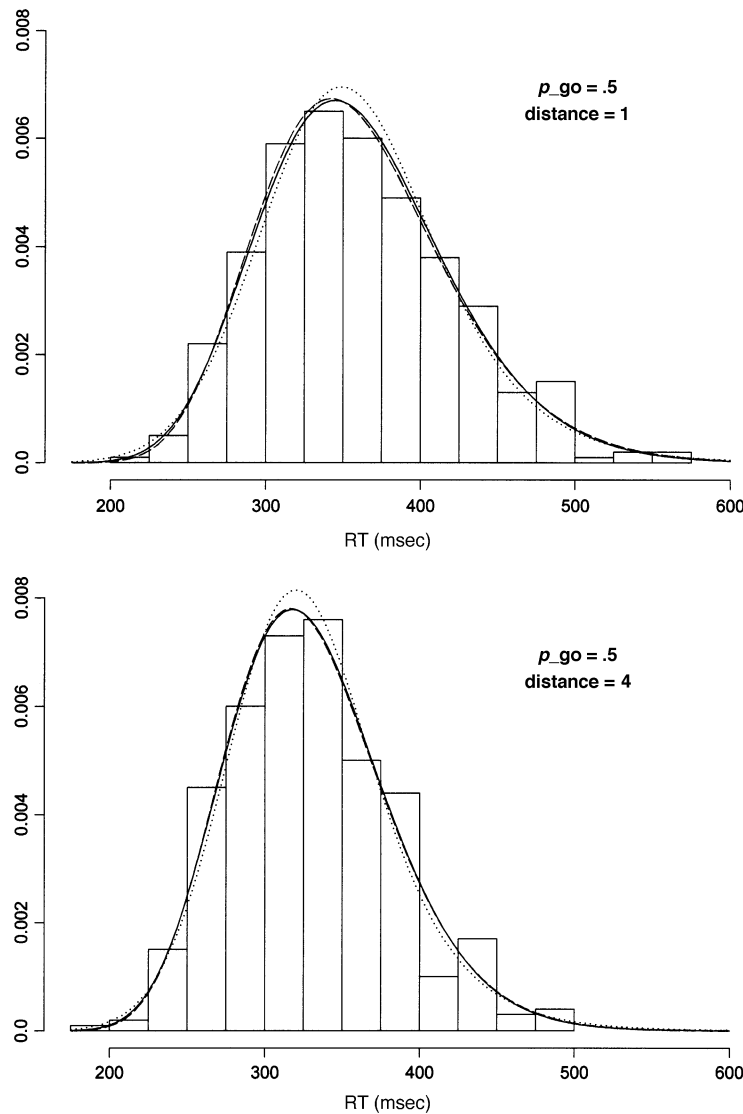
Figure 4. Probability histograms of Schwarz's (2001) data for Participant A, with fitted probability density functions for the ex-Wald and Wald (solid lines), the shifted Wald (dashed lines), and the ex-Gaussian (dotted lines). Each panel provides results for one of four experimental conditions, formed by the factorial combination of the probability of a go response (*p*_go = .5 or .75) and numerical distance (*d* = 1 or 4).

ratio statistic is $\lambda = 2(d_n - d_f)$. The λ statistic is approximately distributed as χ^2, with degrees of freedom equal to the difference in the number of parameters for the full and nested models ($df = 3 - 2 = 1$).

Table 7 shows that no reliable differences in deviance were found in any condition.[6] Hence, the two-parameter Wald distribution provides the most parsimonious model of these data. Furthermore, its parameter estimates follow the pattern predicted by the sequential sampling model and have much smaller standard errors than those in either of the three-parameter models. The estimated Wald densities are plotted in Figure 4 as solid lines. They

are so close to the estimated ex-Wald densities, which are also plotted as solid lines, that no difference is discernable. It appears that the small estimated values of *t* have a negligible effect on the shape of the ex-Wald.

The results of χ^2 tests based on the histograms in Figure 4 also indicate little difference in fit between the three distributions. Summing across conditions, for the ex-Wald, $\chi^2(29) = 45.5$, $p = .026$; for the shifted Wald, $\chi^2(30) = 46.6$, $p = .03$; and for the Wald, $\chi^2(29) = 45.5$, $p = .03$. The χ^2 statistic for individual fits is calculated by all fitting functions, using an algorithm (*chisq*) similar to that in Schwarz (2001) and stored in *fit$chisq*. The
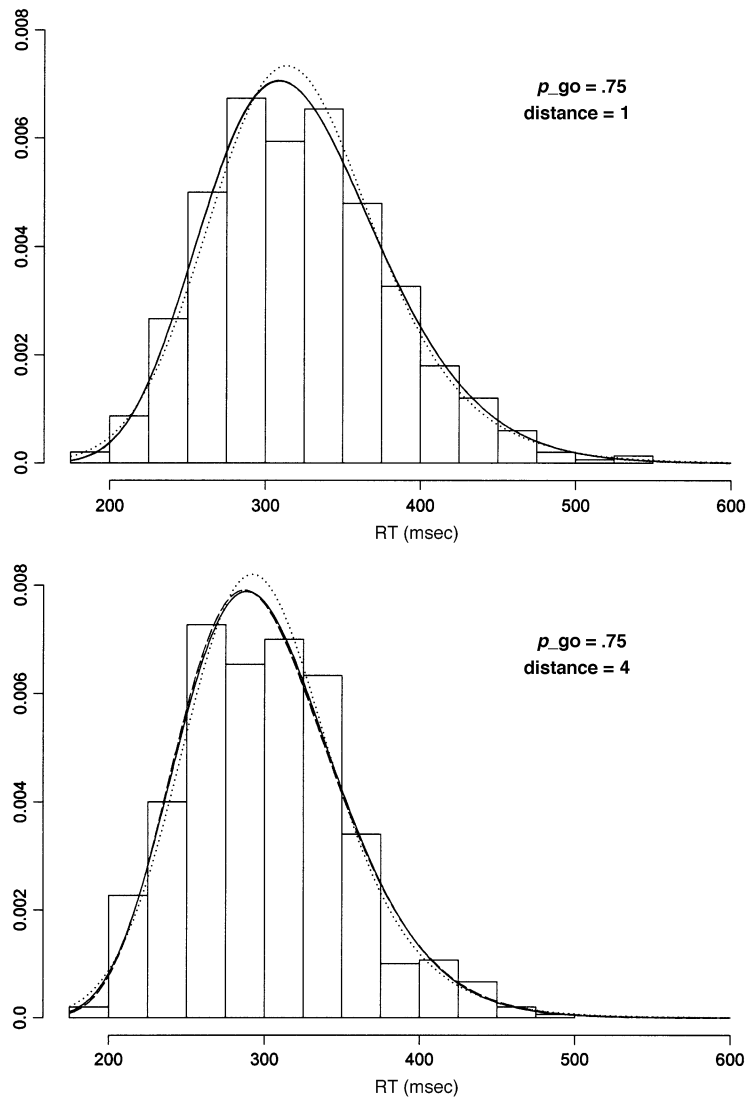
Figure 4 (Continued).

ex-Wald results are larger than those reported by Schwarz (presumably due to numerical estimation differences and, possibly, differences in the aggregation of bins with an expected frequency of less than 5) and achieve significance at the .05 level. Note that interpretation of these χ^2 statistics as measures of absolute fit is not recommended, both because their sampling distribution is often very poorly approximated by the χ^2 distribution and because they can vary depending on the choice of bin location and width (25 msec in this case, following Schwarz).

Table 8 contains the results of fitting the data with the ex-Gaussian distribution. In agreement with Schwarz's (2001) results, the ex-Gaussian clearly fits worse than

Table 5
Ex-Wald Parameter Estimates, Standard Errors (*SE*s), Correlations (*r*), and Minus Loglikelihood ($-l$)
for Schwarz's (2001) Participant A

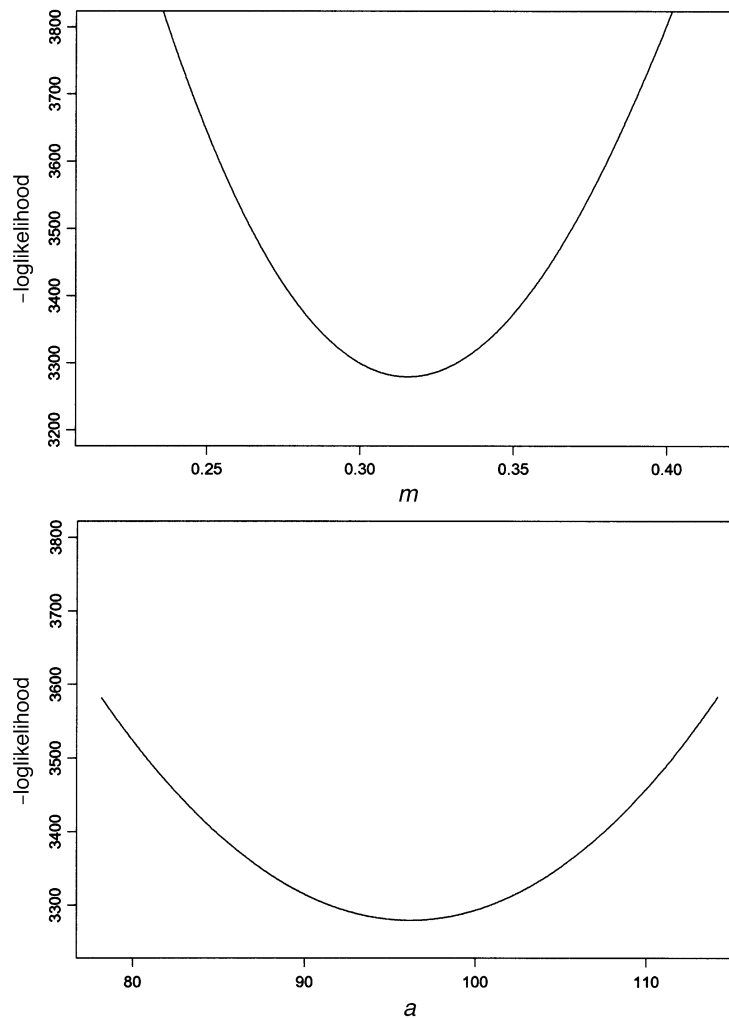| Condition	*m*	*a*	*t*	*SE*(*m*)	*SE*(*a*)	*SE*(*t*)	*r*(*m,a*)	*r*(*m,t*)	*r*(*a,t*)	$-l$
$p = .5, d = 1$	.321	108	22.5	.032	6.39	16.3	.94	.91	.73	2,206.76
$p = .5, d = 4$	.359	112	19.3	.040	7.57	15.5	.96	.93	.80	2,146.22
$p = .75, d = 1$	.316	96	19.3	NA	NA	NA	NA	NA	NA	3,279.50
$p = .75, d = 4$	.349	98	20.8	.029	5.08	10.8	.95	.91	.76	3,213.06

Figure 5. Objective function (minus loglikelihood) profiles around the fitted ex-Wald parameter values for Schwarz's (2001) Participant A in the *p* = .75, *d* = 1 condition. For each panel, the abscissa label indicates the parameter varied (*m*, the mean rate of evidence accrual, *a*, the response threshold, and *t*, the exponential mean) while the other parameters are held at their fitted values.

the Wald distributions, in terms of both deviance and the total histogram-based χ^2 over conditions [$\chi^2(31) = 57.9$, $p = .002$]. As is shown in Figure 4, which plots the estimated ex-Gaussian densities as dotted lines, they have heavier left tails and sharper peaks than the Wald distributions, but similar right tails. The ex-Gaussian parameter estimates were, however, quite sharply defined, as was indicated by small standard errors.

The ex-Gaussian parameter estimates were less correlated than was the case for any of the Wald distributions. A substantial correlation among parameter estimates can raise problems for fitting but does not necessarily indicate that the model is inadequate. High correlations may arise because of the parameterization adopted (*parameter effects curvature*; see Bates & Watts, 1988), and alternative parameterizations may reduce correlations.

However, limited experimentation with different parameterizations of the Wald distribution, including the mean and dispersion parameterization adopted by the *invgauss* distribution functions described in note 4, did not result in any improvement, so the sequential sampling parameterization was retained because of interpretability.

DISCUSSION

This article provides S-PLUS functions to simulate and fit the Wald, ex-Wald, shifted Wald, and ex-Gaussian distributions. The ex-Wald and shifted Wald functions were extensively tested in a Monte Carlo simulation, which allowed automatic starting point estimates and details of the fitting algorithm to be fine-tuned in order to achieve robust numerical performance even with relatively small
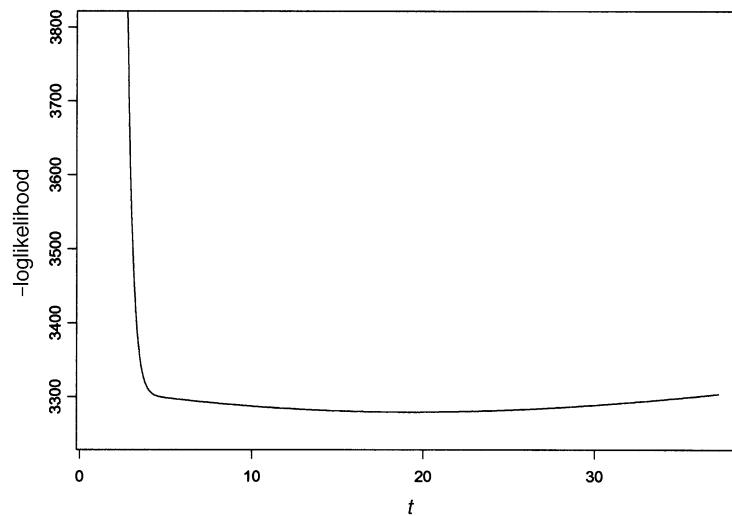
Figure 5 (Continued).

samples. The ex-Wald functions cover both the restricted parameter case considered by Schwarz (2001) and the complementary case considered by Schwarz (2002) and so provide a complete estimation solution.

The Monte Carlo study revealed that estimation of the shifted Wald and ex-Wald distributions can be irregular, in the sense that parameter estimate distributions can be multimodal, particularly for smaller sample sizes and particularly for the ex-Wald distribution. However, for both distributions, bias decreased and efficiency increased with sample size, indicating that estimation was consistent, at least for the parameter values examined. Hence, likelihood estimation for these distributions remains useful despite the irregularity. However, caution should be exercised with smaller samples and even with larger samples when the complex case of the ex-Wald applies. The S-PLUS software can then be used to investigate potentially problematic cases with the same Monte Carlo methods as those reported here.

The ex-Wald distribution required larger samples to obtain the same level of bias and efficiency as the shifted Wald. It is difficult to provide global recommendations for the minimum sample size required in applications, since estimation performance varies depending on the true parameter values. However, overall, the Monte Carlo results suggest a minimum sample size of at least 100 for

the shifted Wald and 400 or more for the ex-Wald. For samples of the minimum or greater size, useful Hessian-based parameter standard error and correlation estimates are usually obtained for the shifted Wald and for the ex-Wald, at least in the real case. However, caution should be exercised in interpreting Hessian-based estimates, since they assume that estimation is regular. Likelihood ratio testing, as was described in the Example Application section, can be used to provide a more reliable inference.

The example application confirmed Schwarz's (2001) finding that Wald distributions provide a clearly superior model of go/no-go RT data, as compared with the ex-Gaussian. In particular, the Wald distributions both fit better and provided theoretically meaningful parameter estimates. The shifted Wald fit these data at least as well as the ex-Wald, indicating that it can also provide a viable approximation to simple and go/no-go RT distribution. The ex-Wald and shifted Wald represent two extreme approaches to dealing with nondecision time; the shifted Wald assumes that nondecision time is always greater than zero but is not variable, whereas the ex-Wald assumes that nondecision time is usually close to zero but is variable.

Both approaches are approximations; in reality, nondecision time must both have a lower bound greater than zero and be variable. Modeling both aspects of nondeci-

Table 6
Shifted Wald Parameter Estimates, Standard Errors (*SE*s), Correlations (*r*), and Minus Loglikelihood
(−*l*) for Schwarz's (2001) Participant A

Condition	*m*	*a*	*s*	*SE(m)*	*SE(a)*	*SE(s)*	*r(m,a)*	*r(m,s)*	*r(a,s)*	−*l*
p = .5, *d* = 1	.284	88	49.3	.036	30.2	67.6	.981	−.957	−.99	2,206.55
p = .5, *d* = 4	.330	100	25.5	.043	35.7	69.6	.982	−.960	−.99	2,146.15
p = .75, *d* = 1	.307	99	0.2	.032	28.5	59.6	.981	−.959	−.99	3,279.51
p = .75, *d* = 4	.313	85	30.2	.032	23.5	48.0	.981	−.957	−.99	3,212.88

Table 7
Wald Parameter Estimates, Standard Errors (*SE*s), Correlations (*r*), and Minus Loglikelihood (−*l*) for
Schwarz's (2001) Participant A and Likelihood Ratio (λ) Tests (*p* = probability under the null
hypothesis of no difference) Comparing the Wald and Ex-Wald [λ(ew)] and the Wald and Shifted Wald
[λ(sw)]

Condition	*m*	*a*	*SE(m)*	*SE(a)*	*r(m,a)*	−*l*	λ(ew)	*p*	λ(sw)	*p*
p = .5, *d* = 1	.308	111	.011	3.9	.972	2,206.74	0	1	0.38	.54
p = .5, *d* = 4	.344	114	.012	4.0	.975	2,146.21	0	1	0.11	.74
p = .75, *d* = 1	.307	100	.009	2.9	.969	3,279.51	.02	.89	0	1
p = .75, *d* = 4	.332	100	.010	2.9	.971	3,213.03	0	1	0.31	.58

Table 8
Ex-Gaussian Parameter Estimates, Standard Errors (*SE*s), Correlations (*r*), and Minus Loglikelihood
(−*l*) for Schwarz's (2001) Participant A

Condition	*m*	*s*	*t*	*SE(m)*	*SE(s)*	*SE(t)*	*r(m,s)*	*r(m,s)*	*r(s,t)*	−*l*
p = .5, *d* = 1	318	46.5	41.6	6.62	4.17	6.54	0.76	−0.89	−0.77	2,577.66
p = .5, *d* = 4	294	39.3	36.4	5.31	3.39	5.26	0.73	−0.87	−0.74	2,515.87
p = .75, *d* = 1	285	44.1	39.2	4.73	2.95	4.66	0.71	−0.87	−0.72	3,833.99
p = .75, *d* = 4	266	39.4	35.4	4.12	2.56	4.05	0.70	−0.86	−0.71	3,768.00

sion time would, at the very least, require one extra parameter, such as the shifted rectangular distribution used by Ratcliff and Tuerlinckx (2002). The greater complexity of this approach increases estimation difficulties, so both the shifted Wald and the ex-Wald remain useful as approximations that, at least in some cases, are likely sufficiently accurate.

Schwarz (2001) improved the efficiency of estimation by making theoretically motivated assumptions about constancy of parameters across conditions. This represents a superior approach, as compared with separate estimation of distributions for each condition, and underlines the advantages of using RT distribution models that are theoretically motivated. Likely, the greater power of this approach would have led to nondecision time estimates significantly greater than zero, in contrast to the results obtained in the example application here. This application was given mainly to illustrate use of the S-PLUS functions, and in practice Schwarz's (2001) approach should be used where possible.

In order to fit models that fix parameters across conditions, the functions provided here must be extended. The powerful S-PLUS environment makes this extension relatively straightforward but requires some knowledge of the S-PLUS language. In particular, a new objective function must be created with the appropriate parameterization and ability to identify data from different conditions. The results of estimation on individual conditions can be used to obtain starting points for the estimation of such models, facilitating search, which must be performed in a higher dimensional parameter space.

REFERENCES

ABRAMOWITZ, M., & STEGUN, I. (1965). *Handbook of mathematical functions*. New York: Dover.

ASHBY, F. G., & TOWNSEND, J. T. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology*, **21**, 93-123.

BATES, D. M., & WATTS, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.

DERENZO, S. E. (1977). Approximations for hand calculators using small integer coefficients. *Mathematics of Computation*, **31**, 214-225.

EDWARDS, A.W. F. (1972). *Likelihood*. London: Cambridge University Press.

HEATHCOTE, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments, & Computers*, **28**, 427-445.

LUCE, R. D. (1986). *Response times*. New York: Oxford University Press.

RATCLIFF, R., & ROUDER, J. F. (1998). Modeling response times for two-choice decisions. *Psychological Science*, **9**, 347-356.

RATCLIFF, R., & TUERLINCKX, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, **9**, 438-481.

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

SCHWARZ, W. (2002). On the convolution of Inverse Gaussian and exponential random variables. *Communications in Statistics: Theory & Methods*, **31**, 2113-2121.

SMITH, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review*, **102**, 567-593.

VENABLES, W. N., & RIPLEY, B. D. (1999). *Modern applied statistics with S-PLUS* (3rd ed.). New York: Springer-Verlag.

NOTES

1. All the parameters are specified by exactly the same symbols in equations given in the text and the S functions. Consequently, the usual convention of using Greek letters for parameters is not adopted.

2. Schwarz (2001) suggests that the approximation be applied when $(mw+a)/\sqrt{(w)} \geq 5.5$. However, the *pnorm* function implemented in S is much more accurate than the one used by Schwarz, and benchmarking against numerical integration results suggests that the second summand, as expressed in Equation 2 using *pnorm*, is more accurate than the approximation unless *pnorm* = 0, and/or e^{2am} = Inf, where Inf is the value assigned by S when a number is larger than can be represented in double-precision floating point notation. Hence, *pwald* uses the approximation only when one or both of these conditions is true, in which case it yields results accurate to approximately six significant figures.

3. http://www.statsci.org/s/invgauss.s, with help at http://www.statsci.org/s/invgauss.html. These functions are parameterized in terms of the mean of the distribution $\mu = a/m$ and its "dispersion," $\lambda = a^2$. For the Wald without a shift, these parameters have closed form maximum likelihood estimates, the arithmetic sample mean for μ, and

$$1 \big/ \hat{\lambda} = \frac{\sum\limits_{i=1}^{n}\left(x_i^{-1} - \hat{\kappa}_1^{-1}\right)}{n}.$$

When this parameterization is used, the Wald distribution is usually called the *Inverse Gaussian*. Note that these functions are not built into S. To access these functions, they must be saved to a text file (e.g., invgauss.s) and compiled [e.g., source("invgauss.s")].

4. Because of the computational cost of memory allocation in S, calculating each term in the series sequentially, with a check for convergence of the approximation at each step, is much slower than calculating a larger number of terms as a vector. The *uandv* function allows vectors of terms to be calculated and can check for convergence after each set is calculated. The default setting, which calculates a vector of the first 20 terms and does not check for convergence, was accurate and fast in all cases examined in the numerical studies reported here. Using more than 20 terms can cause numerical problems in some cases.

5. In the Monte Carlo study, it was necessary to set the minimum values of (m,a,t) to $(1e{-}8,1e{-}8,1)$ for the ex-Wald in order to avoid occasional floating point errors.

6. In theory, the full model should never have a greater deviance than the nested model. However, where the full model is not sufficiently constrained by the data, numerical problems can make it difficult for the search algorithm to find the minimum, and so slightly larger deviance values may be found for the full model. This was the case for all but one of the ex-Wald fits. In these cases, λ was set to zero.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, http://www.psychonomic.org/archive/.

To access these files or links, search the archive for this article using the journal (*Behavior Research Methods, Instruments, & Computers*), the first author's name (Heathcote), and the publication year (2004). The downloadable archive, Heathcote-BRMIC-2004.zip, contains two files:

FILE: rtfit.scc.

DESCRIPTION: This script file contains the S-PLUS functions described in Heathcote (2004), as a 17-k ASCII file. Once the script file is saved, its functions must be compiled [e.g., source("rtfit.ssc")]. These functions work with both S-PLUS 2000 and S-PLUS 6, with the former having been tested more extensively, using both an NT4 and a Windows 2000 system.

FILE: schwarz-a.dat.

DESCRIPTION: This file contains the data from Schwarz (2001) Participant A as a 24-K, three column, tab-delimited, ASCII file. Column names are on the first line, and each following line represents the results for one trial. The first column classifies the go probability for each trial (.5 or .75), the second column classifies numeric distance (1 or 4), and the final column contains the trial RT (msec).

Note that these files and possible future extensions (e.g., extra distribution functions, updates for new versions of S-PLUS, etc.) are also available on the author's Web site: http://www.newcastle.edu.au/school/behav-sci/ncl/.