Psychonomic Bulletin & Review
2005, 12 (2), 195-223

THEORETICAL AND REVIEW ARTICLES

A hierarchical model for estimating response time distributions

JEFFREY N. ROUDER, JUN LU, PAUL SPECKMAN, DONGCHU SUN, and YI JIANG
University of Missouri, Columbia, Missouri

We present a statistical model for inference with response time (RT) distributions. The model has the following features. First, it provides a means of estimating the shape, scale, and location (shift) of RT distributions. Second, it is hierarchical and models between-subjects and within-subjects variability simultaneously. Third, inference with the model is Bayesian and provides a principled and efficient means of pooling information across disparate data from different individuals. Because the model efficiently pools information across individuals, it is particularly well suited for those common cases in which the researcher collects a limited number of observations from several participants. Monte Carlo simulations reveal that the hierarchical Bayesian model provides more accurate estimates than several popular competitors do. We illustrate the model by providing an analysis of the symbolic distance effect in which participants can more quickly ascertain the relationship between nonadjacent digits than that between adjacent digits.

Response time (RT), the time taken to complete a task, is a common dependent variable that has been used to draw inferences about the nature of mental processing (e.g., Luce, 1986). Most researchers tend to analyze only mean RT, but a growing number are examining whole RT distributions as a means of providing extensive and insightful tests of cognitive and perceptual theories (e.g., Ashby, Tien, & Balakrishnan, 1993; Dzhafarov, 1992; Hockley, 1984; Logan, 1992; Ratcliff, 1978; Ratcliff & Rouder, 1998, 2000; Rouder, 2000; Rouder, Ratcliff, & McKoon, 2000; Spieler, Balota, & Faust, 1996; Theeuwes, 1992, 1994; Townsend & Nozawa, 1995; Van Zandt, Colonius, & Proctor, 2000; Vickers, 1980). Consider the following example, which demonstrates the appeal of distributional analysis.

Sternberg (1966) asked participants whether a probe item was presented in a study set. The resulting data were well described by a linear relationship between set size and mean RT, and this pattern is consistent with serial scanning in immediate memory. Consider the simplest exhaustive serial model, in which participants match a probe to each item in memory in succession. Let the time to match the probe to the ith item be S_i. The RT to answer for a study set of k items is $\mathrm{RT} = S_1 + \cdots + S_k + R$, where R is the time for residual processes, such as encoding the probe item and executing the response. In the simplest model, it is assumed that the times to match items follow a common distribution (S) and are independent. This model yields a number of predictions about how the data vary as a function of the number of items (k), including the following: mean(RT) = mean(R) + $k \times$ mean(S). This prediction about mean RT holds in experimental data (Atkinson, Holmgren, & Juola, 1969; Sternberg, 1966).

Townsend and colleagues, however, have repeatedly pointed out that this pattern among mean RTs can be consistent with other models, such as those based on parallel scanning with capacity limits (see Townsend & Ashby, 1983, for a review). Instead of relying on model predictions about mean RT alone, researchers can examine whole RT distributions. For the case of the simple model above, there are two other predictions: one about the variability and the other about the shape of RT distributions. The variability of an individual's RT distribution follows: Var(RT) = Var(R) + $k \times$ Var(S). The shape of an individual's RT distribution approaches that of a normal distribution as study set size (k) is increased (an application of the central limit theorem). Ashby et al. (1993) provided an exceedingly comprehensive analysis of RT distributions for the Sternberg memory-scanning task. They tested a number of distribution-level predictions for a wide class of serial and parallel models. They concluded that both serial models and unlimited-capacity parallel models are not consistent with the data. They favored a limited-capacity, self-terminating parallel model such as Ratcliff's (1978)

This research was supported by National Science Foundation Grant SES-0095919 to J.N.R., D.S., and P.S., by University of Missouri Research Board Grant 00-77 to J.N.R., and by National Science Foundation Grant DMS-9972598 to D.S. and P.S. We are grateful to Michael Stadler, who allowed us use of his data. We are indebted to Andrew Heathcote, who has provided constructive criticism and interesting ideas for development in his reviews. Correspondence concerning this article may be addressed to J. N. Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211 (e-mail: jeff@missouri.edu).

diffusion model. This example shows how distribution-level analysis may provide deeper insight into cognitive processes than just mean-level analyses do.

The main disadvantage of distribution analysis is that it requires a large number of observations to be effective. Moreover, these observations need to be independent replicates from a common source. To meet this requirement, researchers need to gather several hundred observations in each condition from each participant. For example, in testing distributional properties in memory scanning, Ashby et al. (1993) analyzed about 1,500 observations per participant per study set size. Each participant took part in 17 sessions and observed about 6,000 total trials. Likewise, in their analysis of RT distributions, Ratcliff and Rouder (1998, Experiment 1) collected about 10,000 total observations per participant.

The requirement of several replicates per participant and condition is often burdensome and prohibitive. In practice, many researchers have access to college participant pools. In these pools, it is convenient to gather data from a large number of participants who take part in a single session of a few hundred trials. These trials must be partitioned across several conditions. The upshot is that in many applications, the number of RT observations in conditions is in the tens, rather than in the hundreds or thousands.

This requirement of a large number of independent replicates would not raise problems if participants did not vary substantially from each other. If data from all participants were samples from the same underlying distributions, we could simply pool all the data and consider them as independent replicates from a common source. There is, however, often great variability in distributional properties across participants themselves. Figure 1 shows an example from a very simple experiment in which participants had to indicate the location of an asterisk on the screen. The distributional properties highlighted in the figure are shift (the time at which the distribution first attains mass), scale, and shape. Even though this task is simple, there is still substantial across-participant variability in these distributional properties. The first column shows 2 participants whose distributions vary in shift by a factor of two. The middle column shows 2 participants whose distributions vary in scale by a factor of two, and the last column shows 2 participants whose distributions vary from very skewed to nearly symmetrical. This variability in distributional properties across participants eliminates the possibility of considering all of the RTs as independent replicates from a common source.[1] One of the major challenges in using RT distributions is accounting for variability both within individuals and between them.

The goal of this article is to provide a statistical model for the estimation of RT distributions in cases in which researchers have data from several participants, but with only a small number of observations per participant. The model has three main properties. First, it is parametric; each participant's RTs are assumed to follow a three-

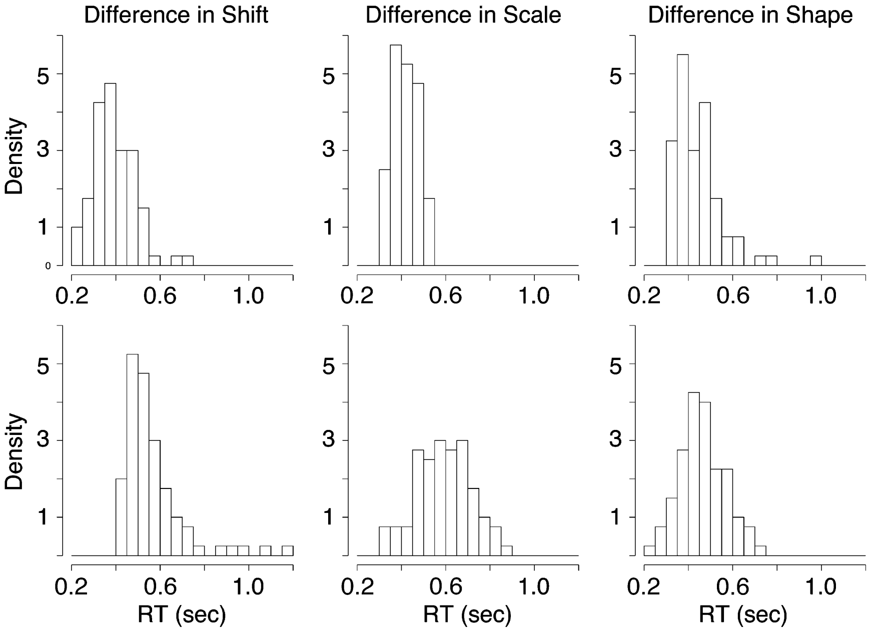

Figure 1. Histograms of response times (RTs) for selected pairs of participants. The left-hand column emphasizes differences in shift across participants. The middle and right-hand columns emphasize differences in scale and shape, respectively. Histograms are scaled so that the total area is one. The area of the rectangles, as opposed to their height, yields the proportion of observations within the corresponding interval. Figure reprinted with permission from Rouder, Sun, Speckman, Lu, and Zhou (2003).

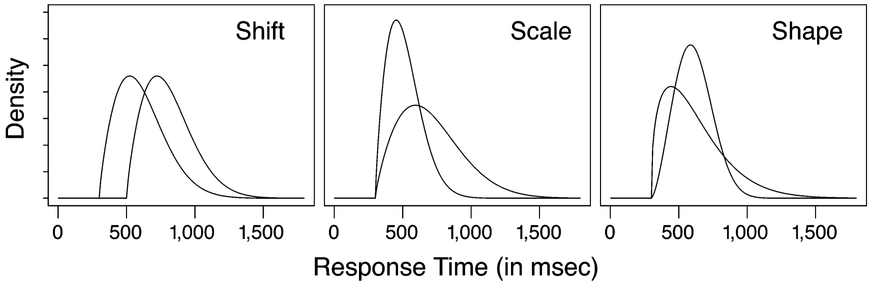

Figure 2. The Weibull parameters of shift, scale, and shape. Each plot shows the effect of changing one parameter while holding the other two constant.

parameter Weibull distribution[2] (Johnson, Kotz, & Balakrishnan, 1994). The Weibull is a flexible form whose parameters correspond to the heuristics of shift, scale, and shape.[3] The meaning of these parameters is depicted graphically in Figure 2. Second, it is hierarchical, with separate model components for variability within a participant and between participants. Each participant is accorded unique Weibull parameters, but a common underlying distribution describes the variability of these parameters. Third, the method of parameter estimation and hypothesis testing is Bayesian. Researchers typically have several choices of methods in performing inference in nonhierarchical models, including classical methods such as maximum likelihood (ML) and least squares. We choose Bayesian methods for the hierarchical models because of feasibility. Since the three-parameter Weibull is outside the family of generalized linear models, we know of no method to perform inference on the hierarchical version other than Bayesian methods. There has been a great deal of progress in the last decade in estimating hierarchical models with Bayesian methods (e.g., Gelman, Carlin, Stern, & Ruben, 1995), and this progress makes inference with the presented models feasible.

Before discussing the model, we provide a context for its intended role. It is important to differentiate between statistical modeling and substantive modeling. Statistical models, such as the analysis of variance (ANOVA), regression models, and structural equation models, are used in a different spirit than more substantive models, such as the diffusion model (Ratcliff, 1979) or the interactive activation model (McClelland & Rumelhart, 1981). The former are used for estimation and inference, whereas the latter are tested as truthful models of phenomena. We view our hierarchical Weibull model as best used in the spirit of the former; its usefulness is derived from the interpretability of its parameters, its ability to provide principled inference on these parameters with small samples per participant, its ability to provide reasonable inference even when model assumptions have been violated (this ability is termed *robustness to misspecification*), and its ability to provide a reasonable fit to data. We do not claim that the hierarchical Weibull models provide the best fit to all data. Other forms, such as the ex-Gaussian, may indeed do well. We do claim that hierarchical Weibull models are highly appropriate for inference on characteristics of RT distributions such as shift, scale, and shape in the common case in which researchers collect a limited number of observations from many participants.

This article is divided into three sections. First, we will present the model and method of parameter estimation. Then we will test the model's ability to provide accurate parameter estimates vis-à-vis other contemporary methods, such as ML estimation (MLE), Vincentizing, and quantile-based estimation (e.g., Heathcote, Brown, & Mewhort, 2002; Jiang, Rouder, & Speckman, 2004). The model outperforms all of these methods in estimating both individual- and group-level parameters. After demonstrating the benefits of the model, we will discuss other aspects, including the fit of the Weibull, the interpretation of the parameters, and the model's robustness to misspecification. Finally, we will fit the model to an experiment with a symbolic distance effect. Observers in the experiment had to decide whether a presented digit was less than or greater than 5. Digits far from 5 (e.g., 2 and 8) were more quickly classified than digits close to 5 (e.g., 4 and 6). The hierarchical Weibull model revealed a locus for the symbolic distance effect in terms of distributional properties. Symbolic distance affects the scale of RT distributions.

THE MODEL

The model that we will discuss was first presented by Rouder, Sun, Speckman, Lu, and Zhou (2003). They provided a detailed and formal account. We will provide a less formal and more accessible account of the advantages of the model for experimentalists. As has been mentioned, the model is parametric and hierarchical, and inference is done with Bayesian methods. Although the Bayesian approach has played a large role in theories of decision making (e.g., Edwards, 1965; Luce, 1959; Phillips & Edwards, 1966; Tversky & Kahneman, 1990), it has not been used extensively for data analysis in cognitive psychology (for notable exceptions, see Edwards, Lindman, & Savage, 1963; Myung & Pitt, 1997; Pitt, Myung, & Zhang, 2002; Sheu & O'Curry, 1998). Likewise, hierarchical modeling is better known in social and clinical psychology. Because of this lack of long-standing tradition in either Bayesian data analysis or hierarchical modeling, we will motivate our model with a true anecdote from a baseball game.

Baseball Example

In the late summer of 2000, the struggling Kansas City Royals were hosting the Boston Red Sox. Pitching for Boston was Pedro Martínez, who was having a truly phenomenal year. Many in the crowd came to see Martínez and his dominant pitching. Contrary to expectation, in the first inning, Kansas City scored five runs, and Boston none. At the end of the first inning, one of our colleagues, who is a loyal Royals fan and an APA editor, predicted a final score of 45–0. The reason this prediction is humorous is because it is both quite logical and wildly implausible. It is logical because 45–0 is obtained by multiplying the first inning scores by 9, the number of innings in a baseball game. It is wildly implausible on three accounts. First, there has never been a baseball game with such an extremely high score. Second, Boston was far superior to Kansas City. Third, Boston had the best pitcher in baseball on the mound. After the first inning, Martínez pitched well, allowing only one additional run. Kansas City lost the game by a score of 7–6; Martínez was the winning pitcher.

The reason the logical prediction of 45–0 was so bad is that it was based on a small sample, the result of a single

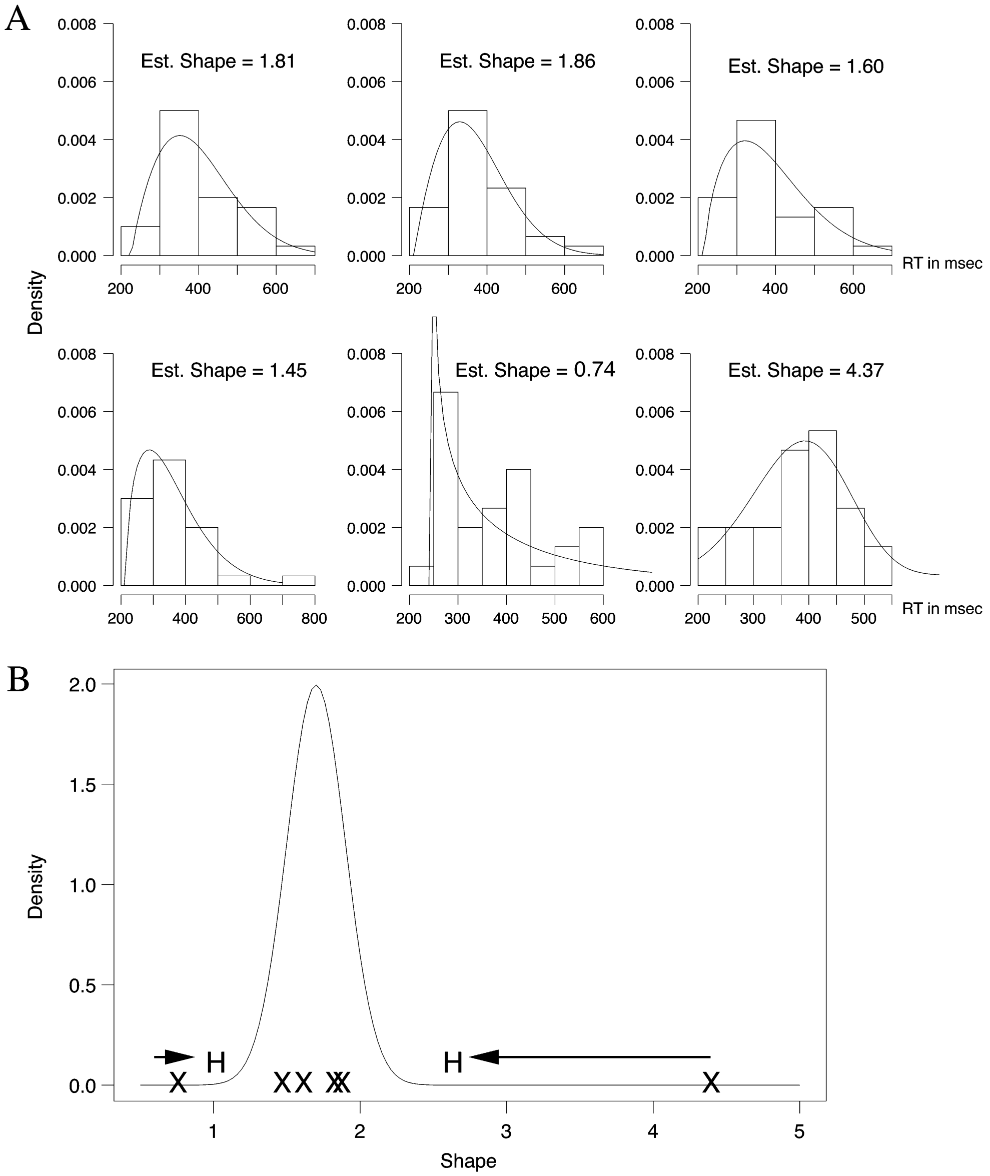

Figure 3. An illustration of shrinkage in a hierarchical model. Estimates of shape from histograms reflect sampling variability. When these estimates are constrained by a plausible parent distribution, they can be adjusted to more reasonable values.

inning of play. One way of improving the estimate is to combine it with our previous knowledge of baseball. Simply put, any person with any inkling about baseball could have produced a better estimate. This prior knowledge does not need to reflect anything more than a knowledge of the distribution of baseball scores. One need not know about pitchers and teams; all one needs for a vastly better estimate is the final score of a few hundred previous baseball games. Of course, the estimate would be even further improved by a detailed knowledge of baseball.

This technique of using a small amount of prior knowledge to improve estimates can be used in the analysis of human performance in psychology experiments. Participants' data, especially when obtained from a small sample, are highly variable. The effects of this variability are acute for estimating higher order properties, such as scale and shape. By using the higher order properties of all the participants together as a baseline, we can better assess whether extreme estimates come about from sampling variability, and if they do, we can correct our estimates accordingly. The baseline need not be overly subjective; in our model, it reflects the contribution of the participants in the present experiment.

Hierarchical Prior Distributions

The key concept in the model is its hierarchical nature. In hierarchical models, parameters are assumed to come from underlying parent distributions.[4] Figure 3 provides an illustration of the advantage of using hierarchical models with parent distributions. The example is for the shape parameter, although any distributional parameter can be treated in a hierarchical manner. Panel A shows hypothetical histograms of 30 RTs from each of 6 participants. In fact, each of these histograms was sampled from a three-Weibull distribution with the same shift (200 msec), scale (200 msec), and shape (1.7). Due to sampling variability, the histograms are diverse. The best-fitting Weibull distributions are drawn over the histograms. The estimated shape parameters are shown, and these are also diverse, ranging from 0.74 (skewed right) to 4.37 (skewed left). Panel B shows a hypothetical parent distribution of the shape parameters. For the purpose of the example, let's assume that this distribution accurately reflects the distribution of shapes in a population. In practice, we only assume a parametric form for the parent distributions and estimate parent distribution parameters from the data. The six shape estimates from the histogram in panel A are indicated by Xs. On the basis of the parent distribution, it is obvious that the two extreme shape values are implausible and reflect a large degree of sampling variability. A reasonable correction is to adjust it toward a more probable value. This adjustment is shown by thick lines with arrows, and the resulting parameter estimates are denoted by Hs (for *hierarchical* estimates). As can be seen, these new values, which reflect the influence of the parent distribution, are closer to the true shape (1.7 in this case) than the original individual estimates are. The adjustment of an extreme estimate to a more moderate one is termed *shrinkage*.

The gains from hierarchical models have been well understood within the statistical literature (e.g., Dey, Ghosh, & Mallik, 2000; Kreft & de Leeuw, 1998). The main problem is that of tractability; although it is easy to postulate hierarchical models, it has traditionally been computationally difficult to analyze them. Over the last decade or so, there have been steady gains in Bayesian statistics that have made these hierarchical models more tractable (Gelman et al., 1995; Tanner, 1993). One recent example of success of Bayesian analysis in psychology is that in item response theory (Fox & Glas, 2001; Wang, Bradlow, & Wainer, 2002). Our model is based on these statistical advances and would not have been possible a decade ago.

In our model, the scale and shape parameters are treated hierarchically, but the shift parameter is not.[5] The parent distributions for scale and shape are *gamma* distributions. The gamma distribution is a two-parameter form; it is quite flexible and can be arbitrarily broad or narrow. The gamma was chosen on the basis of tractability and convenience. The parameters that determine the specific form of the parent distribution are themselves free parameters that are estimated from the data. A more technical specification of the model is given in Appendix A, and an extensive discussion of issues related to parameter estimation may be found in Rouder, Sun, et al. (2003). Software in Splus/R and in WinBUGS (Lunn, 2003; see the WinBUGS Development web-site) may be found at www.missouri.edu/~pcl.

THE ADVANTAGES OF THE HIERARCHICAL WEIBULL MODEL

In this article, we claim that the hierarchical Weibull model is useful in the analysis of RT distributions for four reasons: (1) It allows researchers to pool data across several participants, resulting in vastly improved inference with small samples per participant; (2) its parameters are interpretable in terms of psychological process; (3) it fits data reasonably well; and (4) it is reasonably robust to misspecification. In the following sections, we will examine all four of these reasons in turn.

Estimation With Small Samples

In this section, we will compare Bayesian estimation of the hierarchical Weibull model (HB) with that of other methods. The gold standard of estimation of distributions is ML estimation. ML is an accepted and recommended method in both the statistical literature (e.g., Hogg & Craig, 1978; Lehmann, 1991) and the psychological literature (e.g., Dolan, van der Maas, & Molenaar, 2002; Ulrich & Miller, 1994; Van Zandt, 2000; see Myung, 2003, for a tutorial review). Several good introductory texts cover ML, including Hogg and Craig (1978) and Lehmann (1991).

There is a problem encountered if ML is used to estimate Weibull parameters when the shape parameter is less than 1. When the shape is less than 1, the Weibull resembles an exponential but is even more skewed. In this case, ML estimates of the Weibull are not necessarily consistent (Cheng & Amin, 1983). In our opinion, shape

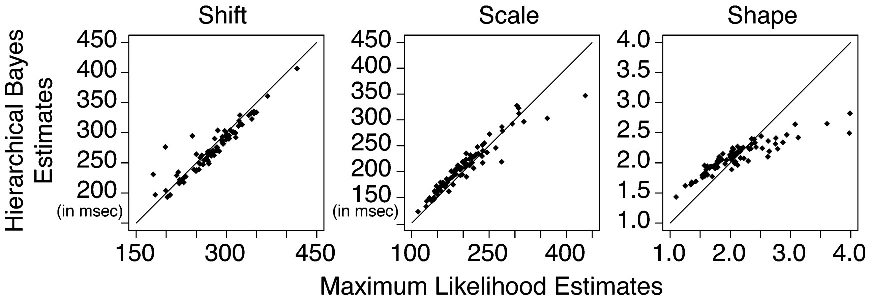

Figure 4. Maximum likelihood and hierarchical Bayesian estimates for Stadler's data. Adapted with permission from Rouder, Sun, Speckman, Lu, and Zhou (2003).

parameters are always greater than 1 in most perceptual and cognitive experiments. Therefore, ML has ample theoretical justification.

In the following subsection, we will estimate parameters from a data set, using HB and ML estimation. For certain participants, the estimates are different. After that, we will describe a large simulation in which 10 estimation methods are compared. The Bayesian method with the hierarchical Weibull model provided more accurate estimates than any other method did.

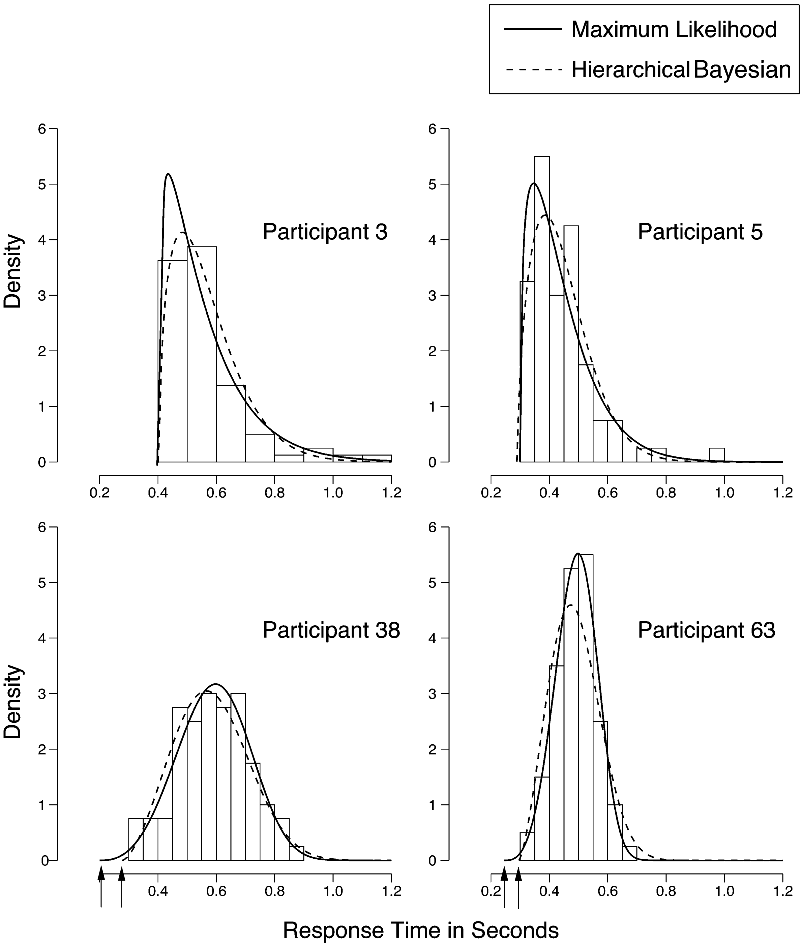

Figure 5. Histograms and fits for 4 selected participants. The solid and dotted lines show Weibull density with parameters from maximum likelihood and hierarchical Bayesian estimates, respectively. These particular participants were selected to maximize the difference between estimation methods. Reprinted with permission from Rouder, Sun, Speckman, Lu, and Zhou (2003).

Application to a Data Set

Rouder, Sun, et al. (2003) have provided an example in which both ML and HB estimates are produced from the same data set. The set, collected by Michael Stadler, consists of 80 observations for each of 80 individuals.[6] Figure 4 shows the relationship between ML and HB estimates as scatterplots. The points represent estimates from individuals. The *x*-axis value of a point denotes the ML estimate, and the *y*-axis value denotes the HB estimate. Overall, many of the points cluster on the diagonal, indicating concordance between the HB and the ML estimates. The big difference is in the shape parameter. Here, the points deviate substantially from the diagonal. The slope is less than 45º, indicating greater variability for the ML estimates than for the HB estimates. To better understand the nature of these differences, it is helpful to show the fits of the Weibull density to individuals' data (Figure 5). The top two panels show data that are fairly skewed, and the ML predictions track this skew well. HB predictions are a little less skewed. According to the hierarchical interpretation, the degree of skew in the data is atypical, given the rest of the participants, and may reflect sampling variability. The Bayesian prediction, which takes this into account, is more moderate. The same dynamics are evident in the bottom panels. Here, the data are atypically symmetric, and the Bayesian predictions are more skewed than the ML predictions. Overall, the extreme Bayesian estimates are less extreme than their ML counterparts and are more like those from typical participants.

Simulation Study

The above analysis shows where HB estimation diverges from more conventional ML estimation. To assess which estimation method is most accurate, we performed a Monte Carlo simulation study. In addition to HB and ML, we included several other methods used or proposed in experimental psychology. We split the methods into two types: those that estimate individuals' shift, scale, and shape parameters and those that estimate group-level shift, scale, and shape.

Individual-Level Methods

Bayesian estimation with the hierarchical Weibull. As has been discussed, Weibull parameters are assumed to be randomly sampled from parent distributions (see Appendix A for details). Estimation is done through Monte Carlo Markov chain techniques, with details provided in Rouder, Sun, et al. (2003). The model yields individual estimates of shift, scale, and shape parameters.

Maximum likelihood. Parameters are obtained by maximizing the likelihood function with the simplex routine (Nelder & Mead, 1965). Each individual's RT distribution is fit separately, yielding individual estimates of shift, scale, and shape.

Quantile maximum likelihood. Quantile maximum likelihood (QML) is a new method from Heathcote et al. (2002). Unlike conventional ML, estimates are based on sample quantiles.[7] The basic idea is that the likelihood of the parameters can be expressed as a function of the sample quantiles. The estimates are those values that maximize the likelihood of the parameters, given the sample quantiles.[8] QML is applied to an individual's sample quantiles to obtain individual parameter estimates. We experimented with several choices of sample quantiles and found that estimators were most accurate when all of the data points served as sample quantiles.[9] Choices with fewer quantiles, such as 5 or 10, led to dramatically worse estimation.

Quantile least-squares. In the quantile least-squares (QLS) method, the parameter estimate is that which minimizes the summed squared error between the sample quantiles and the predicted quantiles. Jiang, Rouder, and Speckman (2004) have shown that QLS is the most efficient means of estimating parameters from sample quan-

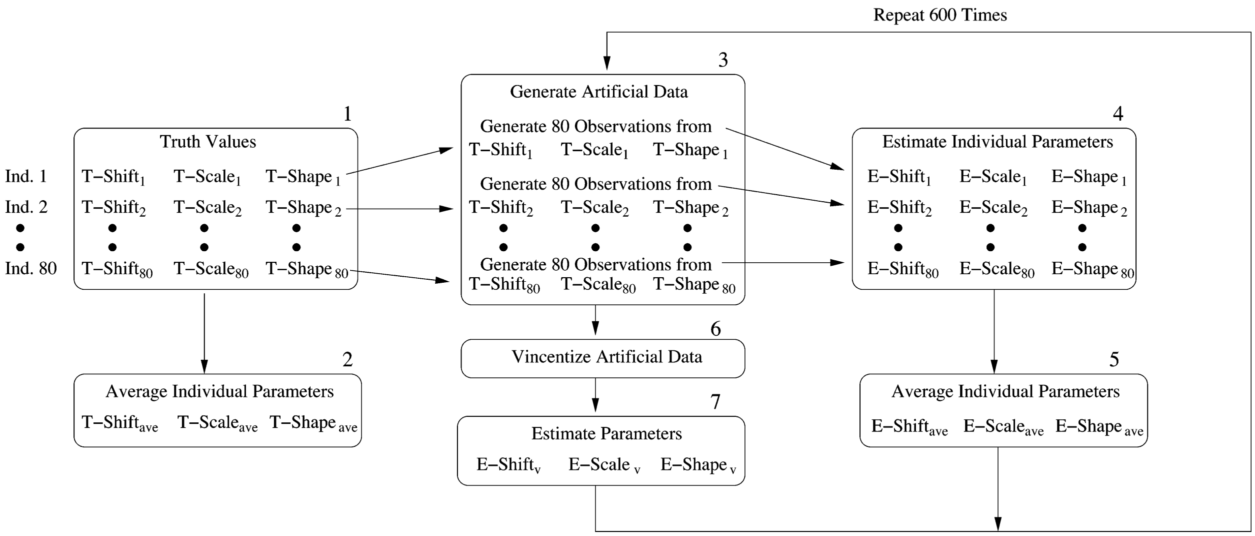
Repeat 600 Times
1
Truth Values
Ind. 1
Ind. 2
Ind. 80
T−Shift$_1$ T−Scale$_1$ T−Shape$_1$
T−Shift$_2$ T−Scale$_2$ T−Shape$_2$
T−Shift$_{80}$ T−Scale$_{80}$ T−Shape$_{80}$
2
Average Individual Parameters
T−Shift$_{ave}$ T−Scale$_{ave}$ T−Shape$_{ave}$
3
Generate Artificial Data
Generate 80 Observations from
T−Shift$_1$ T−Scale$_1$ T−Shape$_1$
Generate 80 Observations from
T−Shift$_2$ T−Scale$_2$ T−Shape$_2$
Generate 80 Observations from
T−Shift$_{80}$ T−Scale$_{80}$ T−Shape$_{80}$
6
Vincentize Artificial Data
7
Estimate Parameters
E−Shift$_v$ E−Scale$_v$ E−Shape$_v$
4
Estimate Individual Parameters
E−Shift$_1$ E−Scale$_1$ E−Shape$_1$
E−Shift$_2$ E−Scale$_2$ E−Shape$_2$
E−Shift$_{80}$ E−Scale$_{80}$ E−Shape$_{80}$
5
Average Individual Parameters
E−Shift$_{ave}$ E−Scale$_{ave}$ E−Shape$_{ave}$

Figure 6. A schematic of the simulation tests.

tiles for a restricted class of distributions called *location–scale models*. Although the three-parameter Weibull is not a location–scale model, it is reasonable to suspect that this method will do well. Rouder and Speckman (2004) used this method to estimate parameters from the Weibull, ex-Gaussian, and shifted-Wald with varying degrees of success. To use the method, it is necessary to choose a set of sample quantiles, and we chose to use all of the data points as sample quantiles.[10] Choices with few quantiles, such as 5 or 10, led to dramatically worse estimation.

Group-Level Methods

Vincentizing + quantile maximum likelihood. Vincentizing is a popular nonparametric method for constructing group-level RT distributions (Heathcote, Popiel, & Mewhort, 1991; Ratcliff, 1979; Vincent, 1912). The basic idea is that individuals' sample quantiles are averaged to produce a group distribution.[11] For example, the 10th percentile of the group distribution is the average of individuals' 10th percentiles. The method of estimating parameters from a Vincentized form is a well-used method of obtaining stable estimates with small sample sizes (e.g., Andrews & Heathcote, 2001; Logan, 1992; Ratcliff & Rouder, 2000; Spieler et al., 1996). Hence, a comparison of Vincentizing with the HB estimation is of particular interest. In the Vincentizing + QML (V+QML) method, estimates are obtained from averaged quantiles by the QML method described above.

Vincentizing + quantile least squares. This method is identical to V+QML, with the exception that QLS is used to estimate parameters from averaged quantiles.

Parameter averaging. An obvious method for constructing group-level estimates is to average individual parameter estimates. This method can be implemented with all the individual parameter estimation methods listed above. When using this method, we add the suffix PA (for parameter averaging) to the method label. For example, HB+PA refers to the method of averaging individual parameters obtained with the HB method. There are four PA methods: HB+PA, ML+PA, QML+PA, and QLS+PA.

Simulation Method

To assess the relative performance of the individual- and group-level estimation methods, we performed a set of simulation studies. Figure 6 provides a schematic of the simulations for the individual-level methods. The first step is to pick "true" values for the simulations. To ensure that we started with a realistic degree of participant variability, we used the ML estimates of the Weibull parameters from Stadler's data set as true values (see Figure 4). There are three estimates per individual and 80 individuals in the set. The resulting 240 true values are depicted in Box 1 in Figure 6 as "T–Shift," "T–Scale," and "T–Shape." We defined the true group parameter as the arithmetic average across the true individual parameters—for example, the true group shift parameter is the average of the individual true shift parameters (see Box 2 in the figure). Artificial data were simulated using individual true values (see Box 3). Individual-level parameters were estimated (Box 4) and then averaged to produce group-level estimates (Box 5). The artificial data were Vincentized (Box 6), and the resulting averaged quantiles were used to produce group-level estimates (Box 7). The process of generating data from true values and estimating parameters was repeated 600 times to obtain the sampling properties of all of the methods.

Two sample sizes were used in the simulations. In the first test, an artificial data set consisted of 80 observations for each of the 80 individuals (shown in Figure 6). These sample sizes are fairly typical for large experiments in cognitive and perceptual psychology. In the second test, the number of observations per individual was reduced to 20. Although 20 seems like a small number, it is typical of the numbers of trials per cell in multifactor experiments. According to conventional wisdom, 20 observations per individual is not sufficient for distributional analysis without the aid of Vincentizing (e.g., Andrews & Heathcote, 2001). Therefore, this small-sample simulation provides a stringent test for HB estimation.

Results

The appropriate statistic to consider is estimation error: the difference between the true value of a parameter and its estimate. By considering errors across all of the data sets, it is possible to construct the sample error distributions. The root mean square error (RMSE) serves as a summary statistic. Table 1 shows RMSE for the first test (80 observations per individual) under the columns "RMSE." The columns labeled "HB Gain" provide a convenient comparison of each method to HB. It is the RMSE of estimates from an alternative method divided by the RMSE of the HB estimates. If the HB estimates are more accurate, the gain is greater than 1. For example, if the gain of the HB estimate over an alternative is 2.0, HB estimates are, on average, twice as accurate as the estimates from the alternative.

The HB method is best for estimating individuals' parameters. The results, however, are more equivocal for estimating group averages. All of the PA methods as a group fared well and outperformed the Vincentizing-based methods. As has been pointed out by Rouder and Speckman (2004) and Thomas and Ross (1980), Vincentizing is not a theoretically justified method for three-parameter distributions, such as the Weibull or the ex-Gaussian. The reason is that in these cases, Vincentized estimators are not consistent; that is, they do not become arbitrarily accurate with sufficiently large sample sizes.[12] The PA methods discussed above are consistent, and given a sufficiently large sample, they can be made arbitrarily accurate. This lack of consistency in estimates from Vincentizing explains its relatively weaker performance with larger sample sizes. Although all of the PA methods performed well, averaged HB estimates held a slight edge for estimating group shift and group scale, whereas averaged ML was best for estimating group shape. Overall, HB+PA and QML+PA were the most accurate group-level methods.

Table 1
Estimation Errors: 80 Observations per Participant

	Error in Shift (sec)		Error in Scale (sec)		Error in Shape	
Method	RMSE	HB Gain	RMSE	HB Gain	RMSE	HB Gain
Individual-Level Methods						
HB	.0222	–	.0271	–	.388	–
ML	.0302	1.36	.0347	1.28	.489	1.26
QML	.0330	1.49	.0371	1.37	.531	1.37
QLS	.0521	2.37	.0557	2.06	.756	1.94
Group-Level Methods						
HB+PA	.00335	–	.00354	–	.0766	–
ML+PA	.00600	1.79	.00772	2.18	.0628	.82
QML+PA	.00340	1.01	.00383	1.08	.0663	.86
QLS+PA	.00738	2.20	.00787	2.22	.0843	1.10
V+QML	.00994	2.97	.00995	2.81	.1390	1.82
V+QLS	.01240	3.70	.01201	3.39	.2369	3.09

Note—RMSE, root mean square error; HB, hierarchical Bayesian; ML, maximum likelihood; QML, quantile ML; QLS, quantile least squares; PA, parameter averaging; V, Vincentizing.

Table 2
Estimation Errors: 20 Observations per Participant

Method	Error in Shift (sec) (RMSE)	Error in Scale (sec) (RMSE)	Error in Shape (RMSE)
Individual-Level Methods			
HB	0.0381	0.0448	0.548
ML	58,173	58,173	1,120,786
QML	89,420	89,420	1,491,868
QLS	93,887	93,877	1,034,443
Group-Level Methods			
HB+PA	0.0076	0.0084	0.147
ML+PA	7,502	7,502	142,739
QML+PA	11,434	11,434	189,261
QLS+PA	10,636	10,636	117,951
V+QML	0.0177	0.0186	0.191
V+QLS	0.0150	0.0137	0.329

Note—RMSE, root mean square error; HB, hierarchical Bayesian; ML, maximum likelihood; QML, quantile ML; QLS, quantile least squares; PA, parameter averaging; V, Vincentizing.

Table 2 shows the results from the second simulation, in which there were only 20 observations per individual. The results are dramatic. Reasonable individual estimates could be obtained only by HB. The ML, QML, and QLS methods failed completely. Likewise, for group-level estimates, HB+PA provided reasonable estimates, whereas ML+PA, QML+PA, and QLS+PA failed completely. The Vincentizing methods did not fail dramatically but were about half as accurate as the HB+PA method.

The reason for the dramatic difference in RMSE stems from a pathology of the Weibull for distributions that are skewed left—for example, those with a high shape parameter value. In this case, changes in the shape parameter have little effect on the distribution. Figure 7 shows this phenomenon. Here, three Weibull densities are drawn. Each density has the same mean and standard deviation. The shapes are varied from 7 to 7,000. Even though the parameters vary by several orders of magnitude, the three densities are very similar. The fact that large differences in parameters do not produce large differences in the densities means that parameter estimation is highly unstable. Fortunately, this parameter instability is present only for Weibull distributions with high shape values, such as those over 4. Typical RT distributions are characterized by Weibulls with shapes between 1.5 and 2.5.

The tradeoff demonstrated in Figure 7 raises problems when Weibull parameters are estimated from small samples. Due to random variability, the sample distribution for small samples may appear to be roughly symmetric or even skewed left. In this case, the Weibull estimates may change by several orders of magnitude. Indeed, this happened in our simulations, and the resulting extreme values dominated the RMSE measure.[13] The influence of the parent distribution in the hierarchical model is greatest in the relatively rare cases in which the samples are symmetric or skewed left. In these cases, extreme parameter estimates are inconsistent with the parent distribution and are not obtained.

Extreme estimates in nonhierarchical approaches (ML, QML, and QLS) are not outliers in the conventional sense. Often, researchers exclude extreme or outlying points from RT analyses. The rationale is that these points may reflect extraneous psychological processes not under consideration. These extreme estimates in the nonhierarchical methods are part of the sampling distribution of the estimators and do not arise from some extraneous process. Hence, there is no logical argument for excluding them. Even if one chooses to exclude extreme estimates, the sampling distribution has smooth tails, indicating that there is no natural method with which to classify whether an estimate is extreme.

The ill-behavior of nonhierarchical estimates comes from the previously discussed pathology of the Weibull. This type of pathology is not evident in the ex-Gaussian and is evident to a far less extent in the shifted Wald (Rouder & Speckman, 2004). Without HB estimation, the Weibull is a poor choice as a descriptive model with small sample sizes, because of these statistical considerations. The HB approach can, in theory, be adopted with other descriptive distributions. Gains in estimation would be expected in these cases too, although these gains would not be as dramatic as with the Weibull.

Researchers may question whether conventional methods can be used with small sample sizes if outlying observations are truncated or censored. Unfortunately, these poor estimates do not result because of extreme observations. In fact, in all cases, no simulated RT was below the smallest true shift parameter of 180 msec. The poor estimates occur when the overall shape of the distribution is skewed left—that is, when typically long observations do not occur in sufficient numbers. Truncation of extreme observations will not lead to good estimation for conventional methods in these cases.

Overall, the results are clear and consistent. The HB method is superior especially for those researchers who gather a few observations per condition from several par-

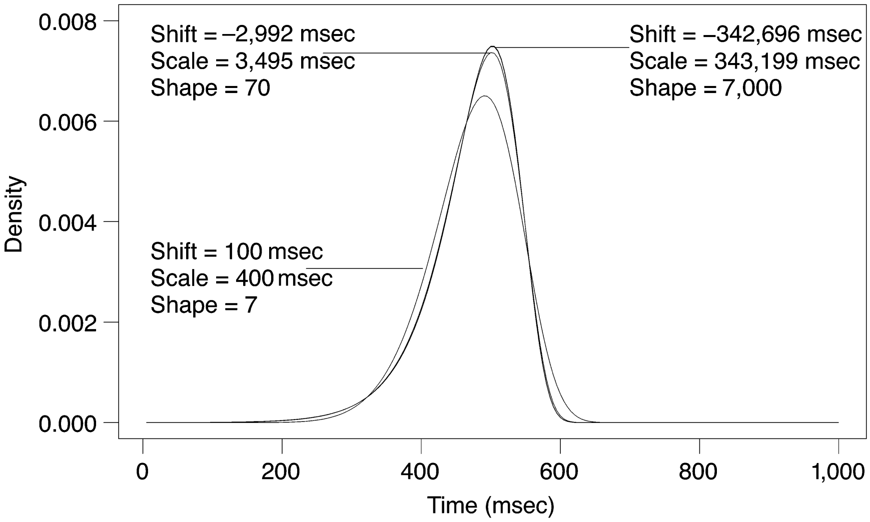

Figure 7. The Weibull distribution is "ill behaved" when skewed left. Distributions with vastly different parameters mimic each other. This behavior raises problems for all of the individual-based methods except the hierarchical Bayesian.

ticipants. We have run these simulations with other true values and for different sample sizes (Jiang, 2002; Rouder, Sun, et al., 2003). The HB method always provides more accurate estimates (has smaller RMSEs) for individual parameters than does any other method.

Interpretation of Weibull Parameters

One of the reasons we find the hierarchical Weibull model useful is that its parameters are interpretable, often in terms of psychological processes. In this section, we will describe three different approaches to interpreting Weibull parameters.

The Weibull as a Descriptive Model

At the least-committed level, the Weibull can be regarded as a convenient descriptive form. The goal then is to provide robust measures of shift, scale, and shape in different experimental conditions with few observations per participant. Shift, scale, and shape are fairly meaningful characteristics of distributions. Accurately measuring the effects of manipulations and group membership on these characteristics can provide motivation for new theories and test beds for existing ones. In this sense, the Weibull model is used analogously to the ANOVA, except that the distributional assumptions are far more realistic and the dependent variables are shift, scale, and shape, rather than the mean.

The Weibull can be used in a descriptive fashion as an intermediary in the fitting of more complicated, theoretically invested models. For example, consider the following possible strategy for fitting Ratcliff's diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998). The diffusion model is usually fit when there are hundreds of observations per participant per condition, since there is a loss of parameter stability with smaller numbers of observations per participant. The hierarchical Weibull model advocated here can be used to potentially increase stability in small-sample applications. In the first stage, the Weibull parameters are obtained by the method presented here. The advantage is that the hierarchical formulation provides a sophisticated means of pooling data across several participants. Then the diffusion model is fit to these Weibull parameters, instead of directly to the data. This process of using one distribution as an intermediary in fitting the diffusion model is not novel. Ratcliff (1978) used the ex-Gaussian in this capacity. This approach will not yield consistent estimates, since the Weibull only approximates the diffusion model densities and should be used with care, rather than programmatically.

The Weibull as a Stage Model

We offer a stage-based, process-oriented interpretation of the Weibull. As Balota and Spieler (1999) have noted, experimental psychologists make a broad, longstanding distinction between two types of processes: central and peripheral[14] (often, the terms *decision* and *residual* are used to describe central and peripheral processes; see, e.g., Dzhafarov, 1992; Luce, 1986). Peripheral processes are quick sensory and motor processes that occur automatically, whereas central processes are processes that require conscious control and attention (e.g., Hasher & Zacks, 1979; Jacoby, 1991; Luce, 1986; Schneider & Shiffrin, 1977). For example, eye movements to a location of a bright flash rely mainly on peripheral processes, whereas maintenance of a 10-digit phone number in memory relies mainly on central processes. It is common to assume that the latency of peripheral processes is small and of low variance, whereas the latency of central processes is large, variable, and skewed right (e.g., Hohle, 1965).

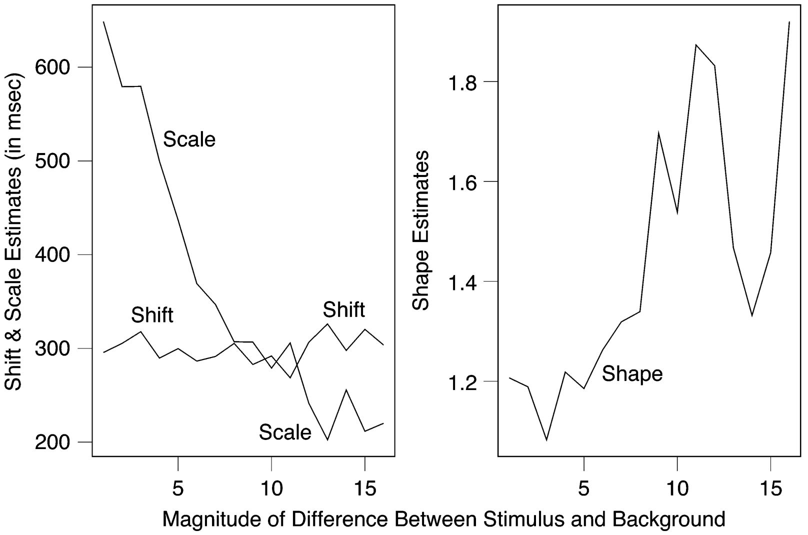

Figure 8. Shift, scale, and shape parameter estimates as a function of the magnitude of the difference between the background and the target stimulus. Data are from Ratcliff and Rouder (1998).

This distinction between central and peripheral processing can be implemented within the Weibull distribution. The Weibull scale and shape parameters index the central processes, whereas the shift parameter indexes peripheral processes. Differences in the structure of central processes across groups or conditions would be manifested as a difference in the shape parameter. Difference in the structure of central processes would include the insertion of stages (e.g., Ashby & Townsend, 1980; Balota & Chumbley, 1984) or changes in processing strategy (e.g., Treisman & Gelade, 1980). However if the central processes follow the same structure across different groups or conditions but the speed of execution is different, there would be differences in the scale parameter, but not in the shape parameter. Finally, differences in the speed of peripheral processes are manifested largely in changes in the shift parameter (e.g., Balota & Spieler, 1999; Hockley, 1984; Ratcliff, 1978). A shift parameter is included in several decision-making RT models (e.g., Busemeyer & Townsend, 1993; Link, 1975; Rouder, 2001) and is a measure of the irreducible minimum (Dzhafarov, 1992; Hsu, 1999)—that is, the minimum possible latency for encoding and responding to a stimulus.

The stage model interpretation offered above is plausible but untested. Perhaps the best way to assess this interpretation is to test whether benchmark manipulations selectively influence model parameters (e.g., Rouder, 2004). The model passes the selective influence test if benchmark manipulations affect only the intended parameter. We hope that the community of researchers will identify selective influence tests of distributional quantities in various domains.

One easy-to-identify selective influence test of the stage model parameter interpretation is that peripheral processes (shift parameter) should not be much affected by decision-critical stimulus variables. These variables presumably affect only central components. This selective influence test can be performed with Ratcliff and Rouder's (1998) Experiment 1. In that experiment, 3 participants observed squares that varied in brightness and indicated whether the brightness was greater than or less than a gray background. Accuracy in this task ranged from ceiling for dark and light stimuli to chance for stimuli that were similar in brightness to the background. According to the stage model interpretation, the peripheral processes (shift parameter) should not vary with brightness. We analyzed the correct response distributions as a function of luminance.[15] Weibull parameters were estimated separately for each participant and luminance level by maximizing likelihood.[16] Then these estimates were averaged across participants. In this experiment, there was a fair amount of symmetry, in that correct RTs to the darkest stimuli were similar to those to the lightest ones. Likewise, correct RTs (and probabilities) to stimuli slightly brighter than the background were similar to those to stimuli slightly darker than the background. Because this symmetry was evident across several levels of luminance, Ratcliff and Rouder (1998) collapsed the data. For ease of presentation, we averaged estimates in the same manner. Estimates for shift and scale are displayed in the left panel of Figure 8 as a function of the distance in luminance between the stimulus and the background. Shift estimates appear invariant across these different conditions, especially when compared with scale estimates.

It may prove more difficult to specify a selective-influence test to differentiate scale and shape. Scale indexes the speed of processing, whereas shape indexes the architecture. It may be more contentious to specify a particular manipulation as one that affects speed or architecture exclusively. On a practical level, the example

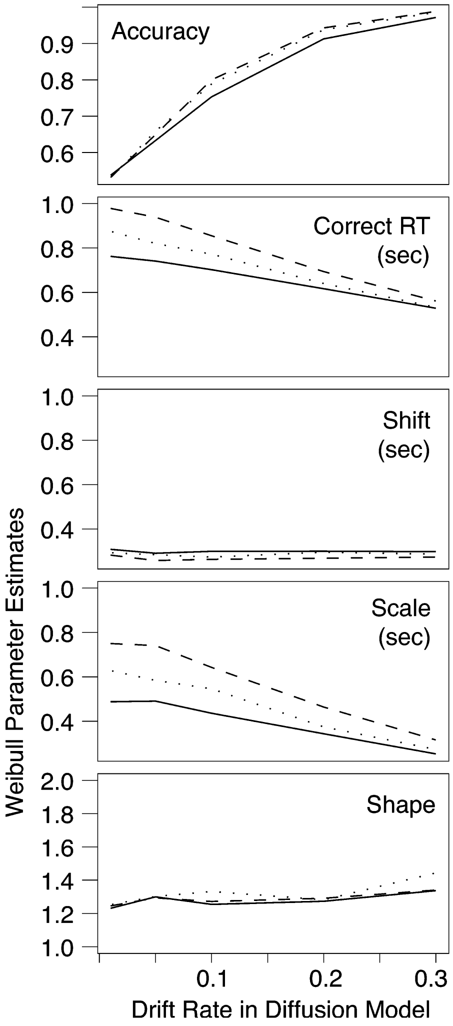

Figure 9. Effect of drift rate on performance. The top two panels show how drift rate affects accuracy and mean response time (RT) for correct responses, respectively. The bottom three panels show how drift rate affects Weibull parameters of shift, scale, and shape, respectively. The effects were obtained by simulating the diffusion model with parameters from Ratcliff and Rouder (1998, Table 1). The solid, dashed, and dotted lines are for simulations in which estimates from Participants N.H., J.F., and K.R. served as true values, respectively.

from Ratcliff and Rouder's (1998) experiment is informative. If one believes a priori that stimulus luminance does not change processing architecture, invariance of shape will be expected. This does not occur in Figure 8 (right panel). If one maintains the stage-based interpretation, it must be that luminance affects architecture. Of course, this is not unreasonable; it is plausible that participants may engage additional processing steps for obviously difficult stimuli, especially when instructed to be accurate. If a recheck mechanism is added for difficult decisions, it may have a very long latency, and the increase in skew comes about from a mixture of slow trials in which the mixture is performed with quicker, unchecked trials. From the figure, it is unclear whether this decrease in shape is sudden or is more gradual. The challenge before theorists is to explain why the shape of the distribution becomes more skewed with increasing difficulty.

One model that appears to be in conflict with the stage-based interpretation of the Weibull parameters is the diffusion model. The diffusion model does not give rise to location–scale–shape RT distribution. In particular, the drift rate affects both scale and shape. In the luminance identification paradigm discussed previously, the greater the difference between the stimulus and the background, the more extreme the drift rate. As the drift rate increases in absolute value, the variance of the RT distributions decreases, and the shape changes. The direction of the shape change, whether more skewed or symmetric, depends on the particular parameters of the process. This shape change poses a challenge to the stage interpretation presented here. Clearly, a change in drift rate is not a change in architecture, yet a change in drift rate does result in a change in shape.

This argument, although true, poses more of a theoretical than a practical challenge. The change in shape predicted by the drift rate changes is quite small, often within the realm of sampling noise. Figure 9 provides an informative example. Data were simulated[17] from a diffusion model with the parameter values reported in Ratcliff and Rouder's (1998) Table 1 (accuracy–stressed condition). The five panels show how various properties change as a function of drift rate. The empirical statistics, accuracy and mean correct RT, vary through the full range of performance for this task. The effect on the Weibull parameters is predominantly on scale; importantly, there is only a small effect on shape. In sum, although the diffusion model does predict some shape changes within a single architecture, the degree of these changes is small.

The stage-based interpretation still needs to be benchmarked, rather than assumed. It is likely that there are some domains in which the stage-based interpretation will be quite reasonable and others in which it will not. It is hoped that, over time, researchers in various domains will perform selective influence tests similar to those discussed here. If there is sufficient consensus that parameters behave reasonably within a domain, the stage interpretation can be especially useful in investigating which manipulations affect the architecture of processing versus its rate.

The stage-based interpretation has important consequences for data analysis. Shape serves as the primary characteristic of interest, rather than mean or variance. Shape indexes cognitive architecture, and it should be analyzed first because it is difficult to interpret changes in processing speed (scale) across conditions if there are accompanying changes in the processing of architecture (shape). Measures of speed are particular to given architectures and are not comparable across architectures. For example, the value of the speed of scanning in a serial process is not comparable to the value of the speed of information gain in parallel counters. If there is a significant and consistent change in shape, the main theoretical enterprise should focus on how to explain this shape change. Only if the shapes are relatively constant can questions about scale (processing speed) be asked and answered.

The Weibull as a Race Model

It is possible to commit to the Weibull distribution as a theoretically oriented model of RT. One interpretation is provided by Logan (1988, 1992), who capitalizes on a limit property of the Weibull. The Weibull distribution describes the distribution of the winning times of a race process. Logan accounts for the process of automatizing a response, a skill, or a task by assuming that identically distributed memory traces race each other to be recalled. The RT is the time of the fastest trace to be recalled. Under reasonable conditions, RT is distributed approximately as a Weibull.[18] Hence, the Weibull is a principled choice when researchers are willing to believe that RT is the result of a race among latent processes.

Fit of the Hierarchical Weibull Model

In this section, we will assess the fit of the hierarchical Weibull model to Stadler's data. In many endeavors, researchers search for models that explain the largest degree of variability in their dependent measures. We make no claims that the hierarchical model is the best model in this regard. It may be that for several data sets, other models, such as the ex-Gaussian or the diffusion model, may do a better job of describing the precise details of data. We treat our model as a statistical model, rather than as a substantive one; its benefit lies in the ability to do estimation and inference in typical applications. Given the intended statistical use of the model, we show that the hierarchical Weibull model fits reasonably well. In the first subsection, we will assess how well the Weibull accounts for individual RT distributions; in the second, we will assess how well the gamma distribution accounts for variation across individuals.

Fit of the Weibull Distribution

A conventional approach to assessing the fit is to compute chi-square goodness-of-fit statistics. We do so here with the caveat that because the Weibull is irregular, the chi-square statistic may not converge to the chi-square distribution. The distribution of the statistics will be, in fact, larger than that of the corresponding theoretical distribution, although the correction is not easily obtained. To compute a chi-square statistic, we first divided the range of variability of each participant's distribution into eight bins. The inner six bins had the same range; the two outer bins were twice the range of the inner bins. This small deviation from uniformly sized bins helps avoid small bin counts that tend to occur in the outer bins. We used the simplex routine to minimize the chi-square fit statistic for each participant separately. This resulted in 80 chi-square statistics. If the model fit well, each chi-square statistic should be an independent sample from a chi-square distribution with four degrees of freedom.[19] To assess fit, we compared the empirical cumulative distribution function of the obtained chi-square statistics with the theoretical cumulative distribution function (Figure 10, left panel). The empirical cumulative distribution function of the chi-square statistics is the line with discontinuities. Chi-square value is plotted on the *x*-axis; the proportion of obtained values below a specific value is plotted on the *y*-axis. The center dotted line is the theoretical cumulative distribution function for the chi-square distribution with four degrees of freedom. The surrounding two dashed lines denote pointwise 95% *estimation error bounds*.[20] Theoretically, for each value on the *x*-axis, there is a 95% probability that the empirical cumulative distribution function would fall within the upper and lower bounds if the obtained chi-square statistics do follow the appropriate chi-square distribution and, by extension, if the Weibull assumption is correct. As can be seen, the obtained empirical distribution function falls within acceptable ranges, indicating a good fit of the

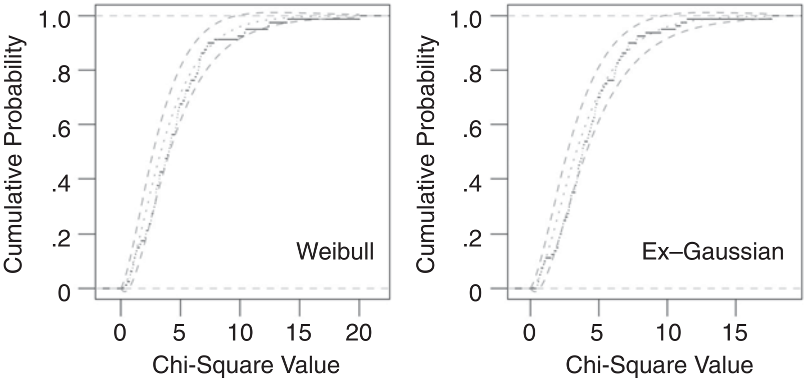

Figure 10. Chi-square fit statistics. The solid line with discontinuities denotes the empirical cumulative distribution function of individuals' chi-square fit statistics. The dotted line is the theoretical cumulative distribution function (CDF) that the chi-square fit statistics should follow if response time is distributed as a Weibull. The dashed lines are the 95% estimation error bounds on the CDF.

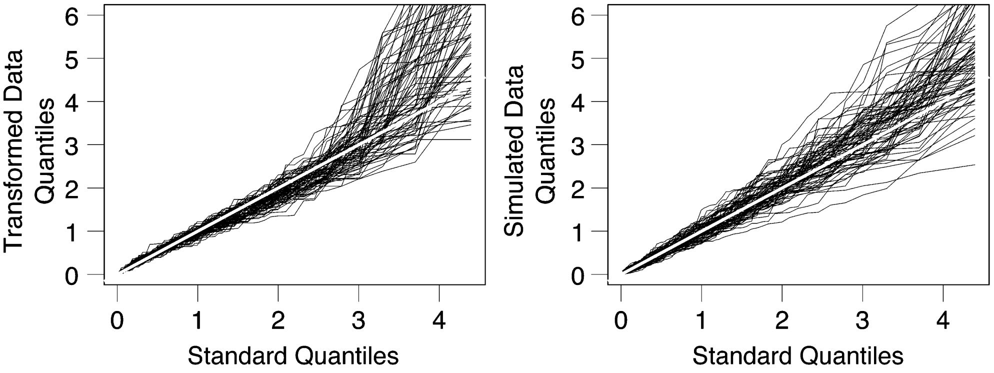

Figure 11. Quantile–quantile (QQ) plots for transformed data (left) and simulated data (right). The diagonal, which would describe a perfect relationship, is indicated with a white line. The figure shows that although the QQ plots for data look similar to those for simulated data of the same sample size, there is a misfit in the tail. The data have slightly heavier tails than those predicted by the Weibull. Data are transformed by hierarchical Bayesian estimates.

Weibull to empirical RT distributions. The right panel of the figure shows the same plot of chi-square statistics for the fit of the ex-Gaussian distribution. The fits are comparable.

We also used quantile–quantile (QQ) plots to graphically explore the fits of distributions. Our plots make use of the fact that Weibull random variables can be transformed to exponential ones. If $\mathbf{Y}$ is a Weibull random variable, $\mathbf{X} = [(\mathbf{Y} - \psi)/\theta]^{\beta}$ is a standard exponential random variable with density $f(x) = \exp(-x)$, where ψ, θ, and β are the shift, scale, and shape of $\mathbf{Y}$. We performed this transformation for each individual's data, using each individual's HB parameter estimates. Quantiles from the transformed data are plotted against those for a standard exponential (Figure 11, left panel). If the data were distributed as a Weibull, the relationship in the QQ plot should be a straight line with a slope of 1 (white line). There are some evident variations from the expected line. The curves are below the line for values between 2 and 3 and above it for values between 3 and 4. This discrepancy indicates that the data have more mass in the tail than does the Weibull distribution. The right

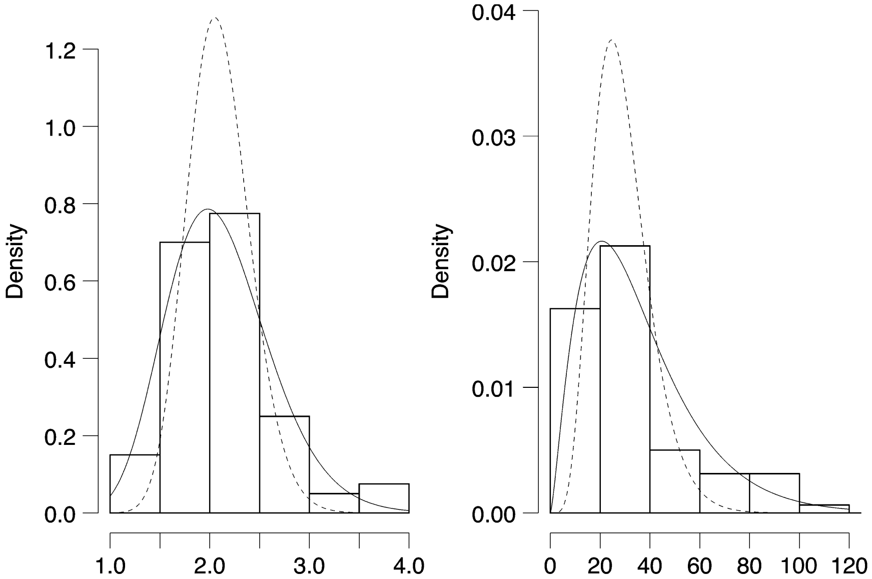

Figure 12. The left panel shows the histogram of maximum likelihood (ML) shape parameter estimates. The solid line shows the best-fitting gamma density. The dotted line is the parent distribution from the hierarchical Bayesian analysis, which is, as was expected, narrower than the ML estimates. The right panel shows the same for the function of scale and shape that was treated hierarchically [$g(\theta, \beta) = \theta^{-\beta}$; see Appendix A]. In both cases, the assumption of a gamma-distributed parent seems quite reasonable.

panel shows a random sample from the exponential distribution (80 observations for 80 participants). The comparison of the panels confirms that there is a slight misfit with the Weibull: It tends to underestimate the tail in some of the participants.

The result of the preceding analyses is that the Weibull distribution fits reasonably well, although not perfectly. In particular, the Weibull's tail is not as heavy as that observed in the data. The reason for this misfit is unclear, but it may very well be that there are a few outliers in the data that come about from atypical processing. For example, if a few participants lost attention on a few trials, the resulting large RTs would cause the observed misfit in the QQ plot. As was mentioned earlier, our goal is to provide a reasonably well fitting statistical model for pragmatic measurement. The fit of the Weibull is more than adequate for this purpose.

Fit of the Parent Distribution

Our hierarchical Weibull model assumes that there is a parent distribution from which individual parameters are drawn and that this distribution has a gamma form. It is reasonable to ask whether such a choice is judicious, especially since it was made without recourse to psychological considerations. The gamma, for example, is unimodal. In certain cases, it may be that participants' parameters are not distributed unimodally. It may be that participants cluster into modes—for example, those that have a symmetric shape and those that have a skewed shape, without many in between. To test this possibility, we plotted the obtained individual ML parameters. Figure 12 shows that ML estimates were unimodal and well fit by a gamma distribution (solid line). The actual gamma distribution that was estimated as the parent in the hierarchical analysis is shown as a dotted line. The estimated parent distribution is narrower because the extreme ML estimates typically reflect a large degree of sampling noise. It is the narrowness of the parent distribution (which is not assumed but estimated) that gives rise to shrinkage in the HB estimation.

Researchers heavily concerned about bimodality can take solace in three facts. First, the effects of misspecification of the prior will be overcome by collecting larger amounts of data. As the amount of data increases, the impact of the prior becomes smaller and smaller. In the asymptotic limit, the prior plays no role. Second, as we will discuss in the next section, the effects of misspecification of the parent distribution are rather marginal. Third, it may be possible to specify reasonable priors that allow for two or more modes.

Robustness to Misspecification

As was mentioned previously, we treat the hierarchical Weibull model as a statistical model, rather than as a more substantive one. One consequence of this treatment is that the inference should be relatively robust to misspecification of the model. In this section, we will explore the consequences when the underlying data are not distributed as a Weibull. We show through simulation that reasonable inferences about location, scale, and shape are possible even when the data are distributed as an ex-Gaussian.

The ex-Gaussian is usually parameterized as the addition of a normal random variable with an exponential random variable. The distribution has three parameters—μ, σ, and τ—corresponding to the location and scale of the normal and the scale of the exponential. There is a little-known alternative parameterization in the location–scale form, as is discussed in note 2. The alternative is based on parameter η, where $\eta = \tau/\sigma$. In this parameterization, μ serves as the location parameter, σ serves as the scale parameter, and η serves as the shape parameter.[21] Figure 13 shows the dependence of the distribution on the parameters. In this parameterization, σ scales both the normal and the exponential components proportionately. Changes in σ alone do not affect higher order shape-related properties, such as skew and kurtosis. Only the value of η determines these higher order properties.

The question at hand is whether we can reliably recover changes in ex-Gaussian location, scale, and shape parameters with the hierarchical Weibull model. To answer this question, we performed a small simulation experiment in which artificial data were distributed as an ex-Gaussian. In the simulation, we constructed two groups of 50 hypothetical participants contributing 50 hypothetical observations each. Each hypothetical participant had his or her own parameters (μ, σ, η), and these parameters were sampled from parent distributions. There were two sets of parent distributions, one for each group.

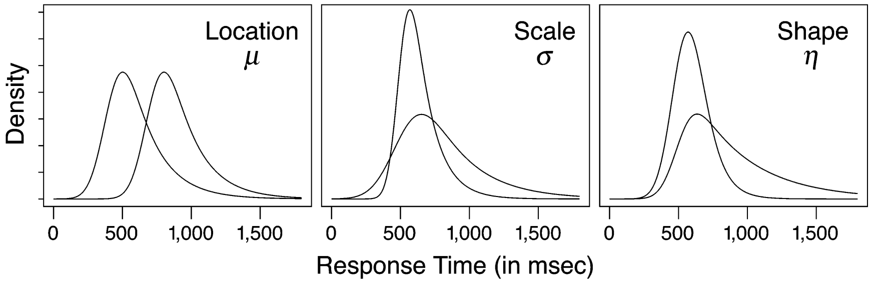

Figure 13. The ex-Gaussian parameters of location, scale, and shape. Each plot shows the effect of changing one parameter while holding the other two constant.

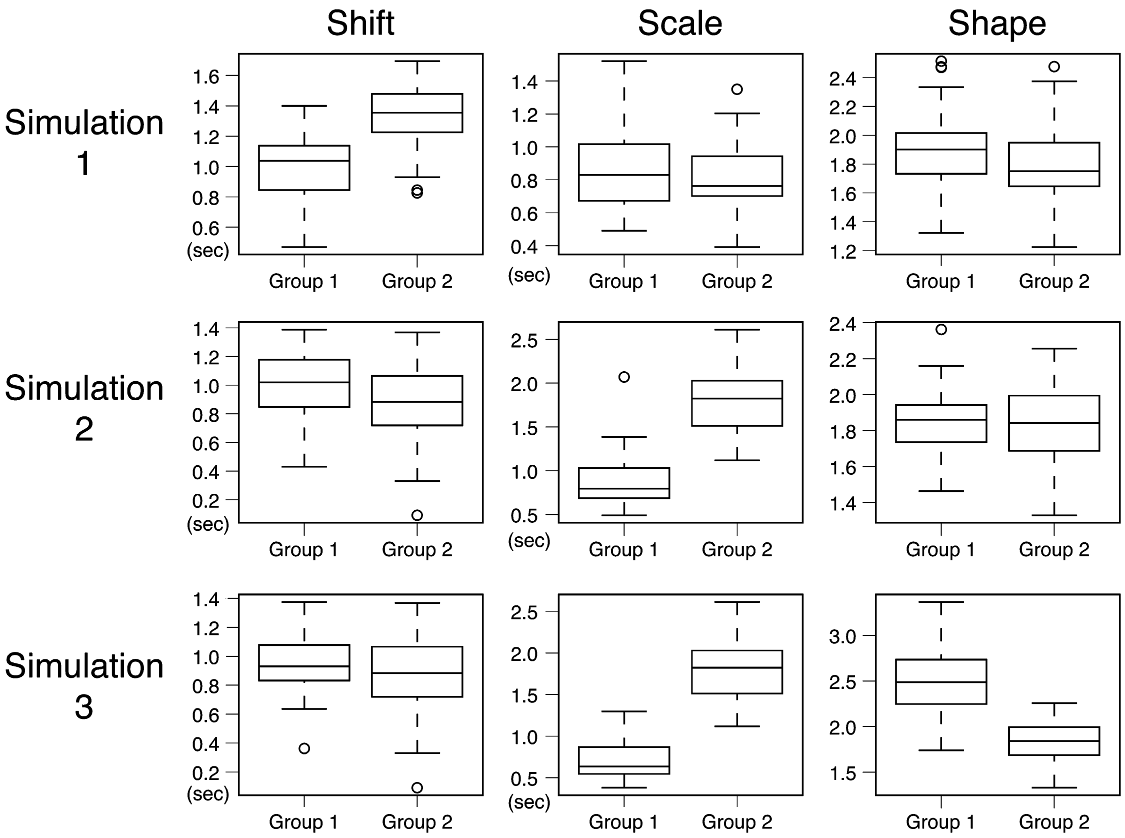

Figure 14. Hierarchical Bayesian estimates of shift, scale, and shape from data generated from an ex-Gaussian.

In Simulation 1, there was only a group difference[22] in the ex-Gaussian location parameter. Hence, hierarchical Weibull analysis should reveal a shift difference but no scale or shape difference across groups. Figure 14 shows HB estimates as box plots, and the upper row shows the results from Simulation 1. Only estimates of shift were affected by group membership.

There are some minor differences in scale and shape across the two groups in Simulation 1. Some random variation is indeed expected, for two reasons. First, although the true population parameters are invariant across conditions, individuals' shapes and scales were randomly sampled. Given that the simulations consist of only 50 individuals, we expect some variation in the distribution of individual true parameters. Second, in the simulations, there were only 50 observations per individual; hence, individuals' samples do not perfectly reflect their true parameter values.

In Simulation 2, there were only group differences in ex-Gaussian scale parameters. Consequently, hierarchical Weibull analysis should reveal a scale difference and no shape or location difference across groups. The middle row shows HB estimates: There is a large scale difference and no shape difference across groups. There is a small shift difference, and the reason for this deviation will be discussed below. In Simulation 3, there were only group differences in ex-Gaussian shape parameters. As has been discussed previously, hierarchical Weibull analysis should reveal a difference in shape and no difference in location. There is no interpretation of scale parameters in this case, since scale (in)variance is fairly meaningless when shape changes. Differences in scale are interpretable only when shape is constant. The bottom panel shows a large effect of group membership on shape estimates, but not on location estimates.

Overall, the hierarchical Weibull provides a robust platform for exploring the effects of manipulations and group membership. The only problem is in measuring shift: Location parameters may not show appropriate invariances when the distribution is misspecified. The Weibull has a true lower bound parameter, below which there is no mass. If the underlying distribution has no such lower bound, the Weibull shift may also change when scale changes. Although the ex-Gaussian lacks a lower bound, real human data certainly have a lower bound (no participant, for example, responds before the experiment begins). We, along with others, argue that the inclusion of a lower bound is sensible. The lower bound is interpretable as a minimum residual, the speed of the fastest possible response (Dzhafarov, 1992; Hsu, 1999).

There is a misspecification that is worthy of careful attention: the possibility of fast guesses (e.g., Yellott, 1969, 1971). Even occasional fast guesses, should they occur, will necessarily have heavy influence on parameter estimation. The reason is that the shift parameter estimate is always below the smallest observed value. If this value is not from the processes under consideration, it is possible to estimate shifts that are too small, scales that are too large, and shapes that are too symmetric. Currently, researchers can take one of two mitigating steps. The first is to use instructions that minimize fast guesses. Green, Smith, and von Gierke (1983), for example, were able to eliminate fast guesses by stressing to their participants the need to wait for stimuli. The second is to trim responses in analysis that could not conceivably be stimulus related. We took this tactic and trimmed responses below 200 msec. The good fit of the Weibull (Figure 11), especially for the fast quantiles, indicates that our trimming was indeed successful in this application. The most elegant solution is to expand the model so that it is a mixture of the guessing state and the Weibull. In this model, the distribution of the guessing state would be specified beforehand—for example, a uniform distribution from 0 to 5 sec. The probability of responses coming from the Weibull or the guessing state would be an additional free parameter. This type of mixture would allow for more robust estimation of Weibull parameters. We are currently developing this type of mixture model for the hierarchical Weibull. It is worth noting that all models that posit a lower bound, such as sequential sampling models, are affected by the presence of fast guesses.

It is worth considering the violation of the model's assumptions about the parent distributions. It turns out that proper specification of these is not critical. The real gain in estimation comes not from proper specification of the majority of the parent distribution's mass, but from the orderly decrease in its tails. The parent distribution has its largest influence when an individual's data lead to extreme estimates. In this case, the individual estimate is typically in the tails of the parent distribution. Consider the previous example in Figure 3, in which the extreme estimates were in the tails of the parent distribution. Consequently, they were adjusted to regions of the parent distribution with more mass. In essence, it is the fact that the parent distribution's tails fall off in a reasonable manner that provides much of the gain. Misspecifying the more central regions will certainly add some bias to estimates. This bias, however, is minimal, considering the gain in accuracy from the shrinkage of extreme estimates.

AN APPLICATION TO A SYMBOLIC DISTANCE EFFECT

In this section, we apply the model to a well-known symbolic distance effect. In the experimental task, participants decide whether a presented digit is greater than or less than 5. In this task, it typically takes longer to make decisions about digits close to 5, such as 6 and 4, than to digits far from 5, such as 2 and 8 (Link, 1990; Moyer & Landauer, 1967; Poltrock, 1989; Smith & Mewhort, 1998). The interpretation typically offered is that when performing this task, participants represent numbers in an analogue fashion (Moyer & Landauer, 1967). The analogue code of 5 is more similar to 4 than to 2. When the codes are similar, it takes longer to obtain an accurate comparison. This finding is one of several semantic and symbolic distance effects that have been used to examine whether there are analogue mental representations (e.g., Kosslyn, 1975; Moyer, 1973). The question we ask is whether the distance-from-5 symbolic distance affects shift, scale, shape, or some combination of these distributional properties. Fifty-four participants each contributed about 40 observations in each of the six conditions.

Method

Participants. Fifty-four University of Missouri undergraduate students served as participants in partial fulfillment of a require-

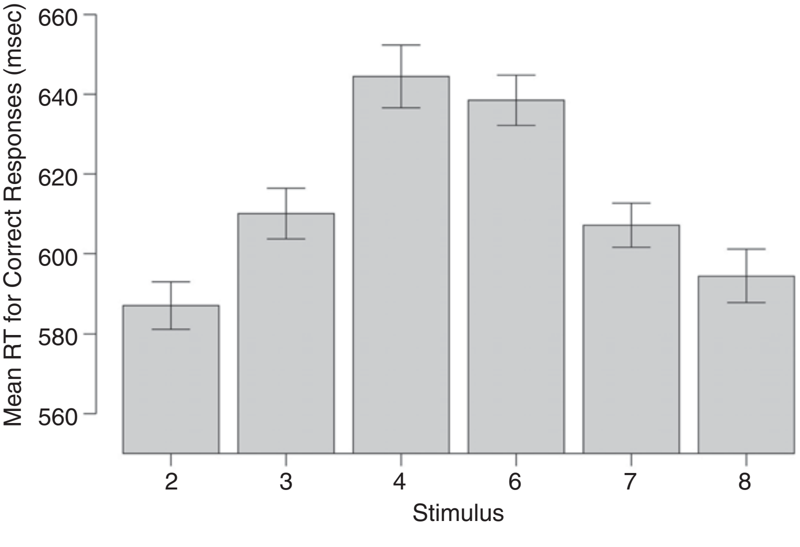

Figure 15. Mean response time (RT) for the six different digits. Error bars denote 95% "within-subject" confidence intervals (Masson & Loftus, 2003).

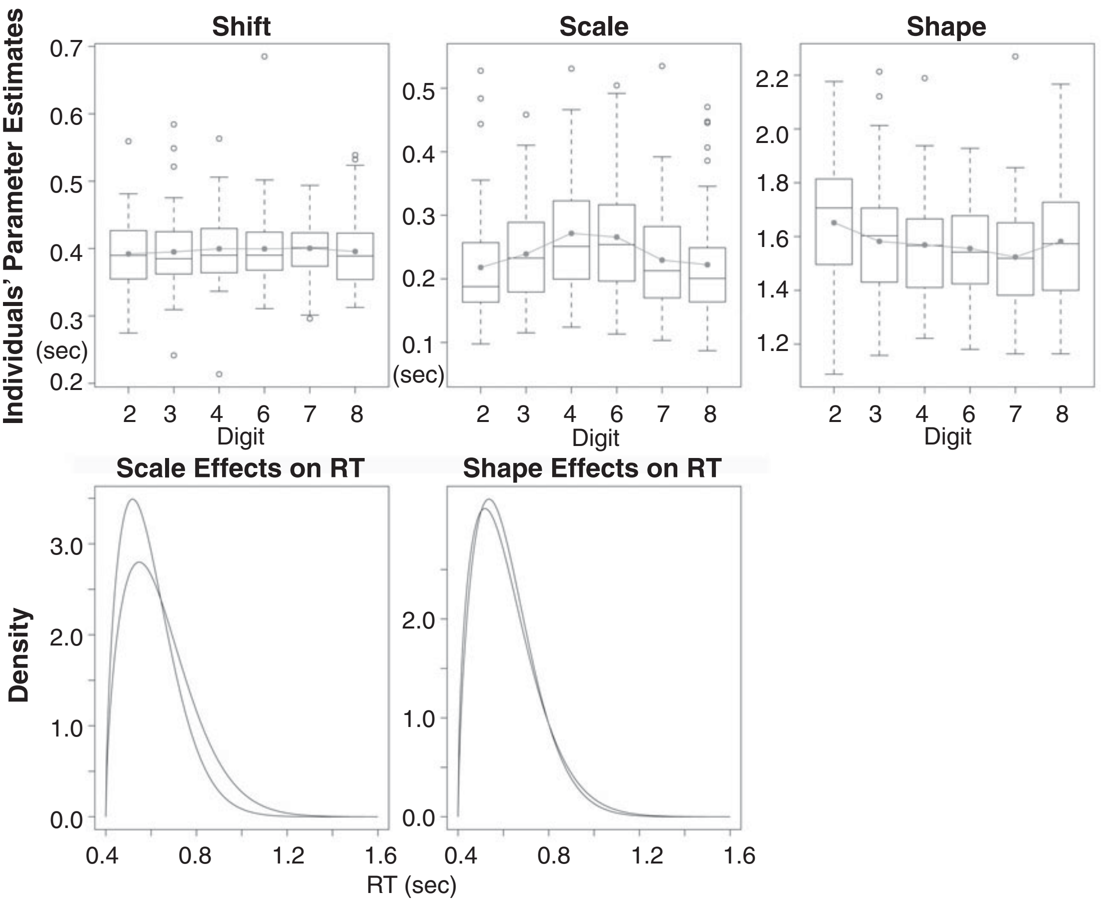

Figure 16. Parameter estimates. The top panel shows box plots of the distribution of parameters across participants. The points and lines denote means. There is no effect of digit for shift, a moderate effect for scale, and an unsubstantial effect for shape. The bottom row shows the effect of scale and shape on response time (RT) distributions. The difference is much larger for scale than for shape.

ment in an introductory psychology course. Two were eliminated for excessive anticipatory responses.

Design. The to-be-classified digit served as the main independent variable. The digit was 2, 3, 4, 6, 7, or 8. The levels were manipulated in a within-subjects design. All of the digits appeared equally often and in a random order.

Procedure. A trial began with a black screen. After 1,000 msec, a digit was presented in the center of the screen in a standard DOS font. The digit remained on the screen until the end of the trial. The participants were instructed to depress the "z" key if the digit was less than 5 and the "/" key if the digit was greater than 5. Following the response, feedback was provided: A pleasant, rising two-tone sequence indicated a correct response, whereas a low-frequency buzz indicated a wrong response. Feedback lasted for 400 msec, after which the next trial followed. Sixty trials made up a block. The participants were instructed to take breaks between blocks. The session consisted of six blocks and took approximately 20 min to complete.

Empirical Analysis

Visual inspection of mean RT as a function of trial number revealed that all noticeable practice effects occurred within the first 25 trials. Hence, these trials were discarded. Additional trials were discarded if (1) the trial followed a break, (2) the response was incorrect, or (3) the response time was less than 200 msec or greater than 2,000 msec. Fewer than 1.8% of the responses were errors, and fewer than 0.7% were outside the window from 200 to 2,000 msec. The mean RT for the six digits is shown in Figure 15. As can be seen, there is a significant 50-msec symbolic distance effect [$F(5,255) = 42.2, p < .05$]. The latency of the decision varies inversely with the distance from five.

Hierarchical Model Analysis

We used the hierarchical Weibull model to estimate each individual's shift, scale, and shape parameters in all six digit conditions. We started with a general model in which there are separate shift, scale, and shape parameters for every participant–digit combination. This model yielded a total of 3 × 52 participants × 6 digit condition = 936 primary parameters. Details of this model are presented in Appendix B. Figure 16 shows the resulting estimates as box plots. The top row contains separate panels for shift, scale, and shape estimates. Within each of these panels, there are box plots of the distribution of the 52 individuals' parameters tabulated by digit condition.

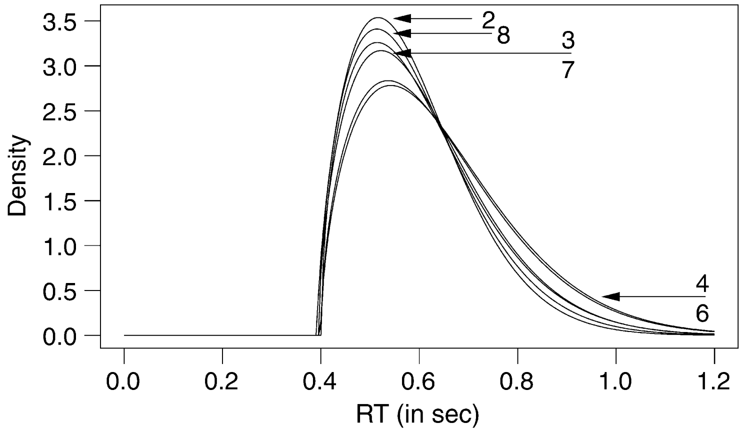

Figure 17. Group-level response time (RT) distribution for the digit conditions. The effect of digit is on the scale of the distribution.

The distribution of the shift parameters (top left) shows that shift hardly varies with digit. This indicates that peripheral processes are unaffected by digit condition. For the scale parameter (top-middle panel), there is a clear dependence on digit with larger scales for digits closer to 5. The effect of digit on shape is less clear. The change in mean shape across digits is sizable in terms of participant variability but fairly small in substantial significance (from 1.65 to 1.52). The bottom row of panels makes this point clear. The first panel shows the difference between two Weibull distributions that vary in scale (shift fixed at 0.4 sec, shape fixed at 1.6). The two scale values (0.218 and 0.272 sec) were chosen because they were the most extreme mean scales (from digit conditions 2 and 4, respectively). As can be seen, there is a moderate difference between these distributions. The middle panel is the comparable plot for shape. The scales and shifts are fixed (shift of 0.4 sec, scale of 0.24 sec). The shape values (1.52, 1.65) were chosen because they are the most extreme mean shapes (from digit conditions

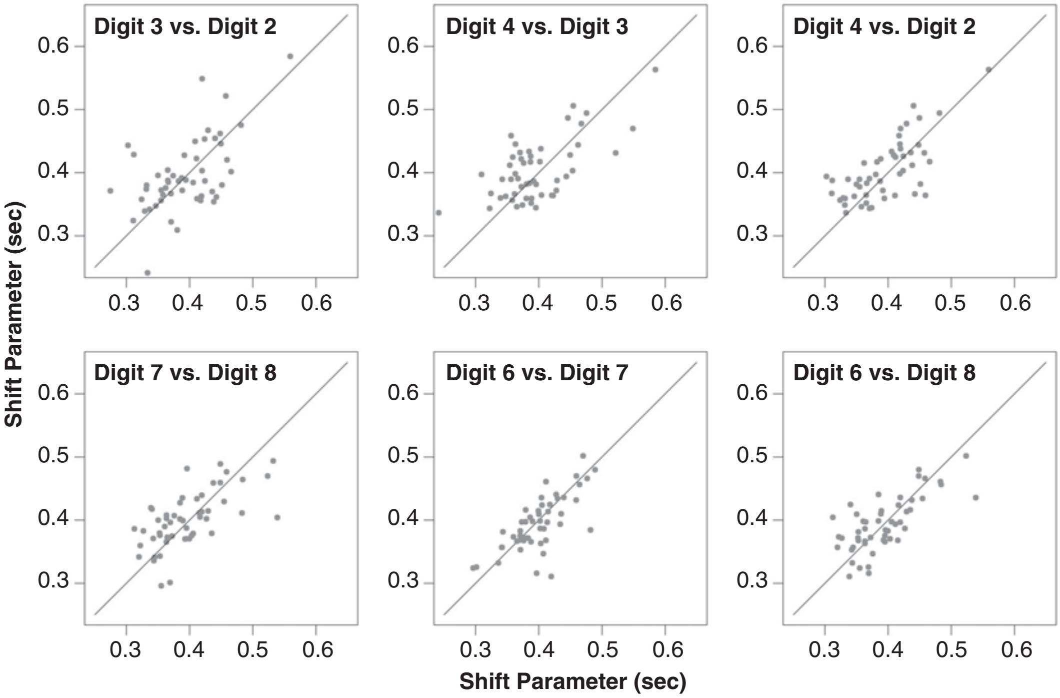

Figure 18. Scatterplots for the shift parameter. Each point corresponds to a particular participant. There is a high degree of correlation, indicating that the participants who had a high shift in one condition tended to have a high shift in another.

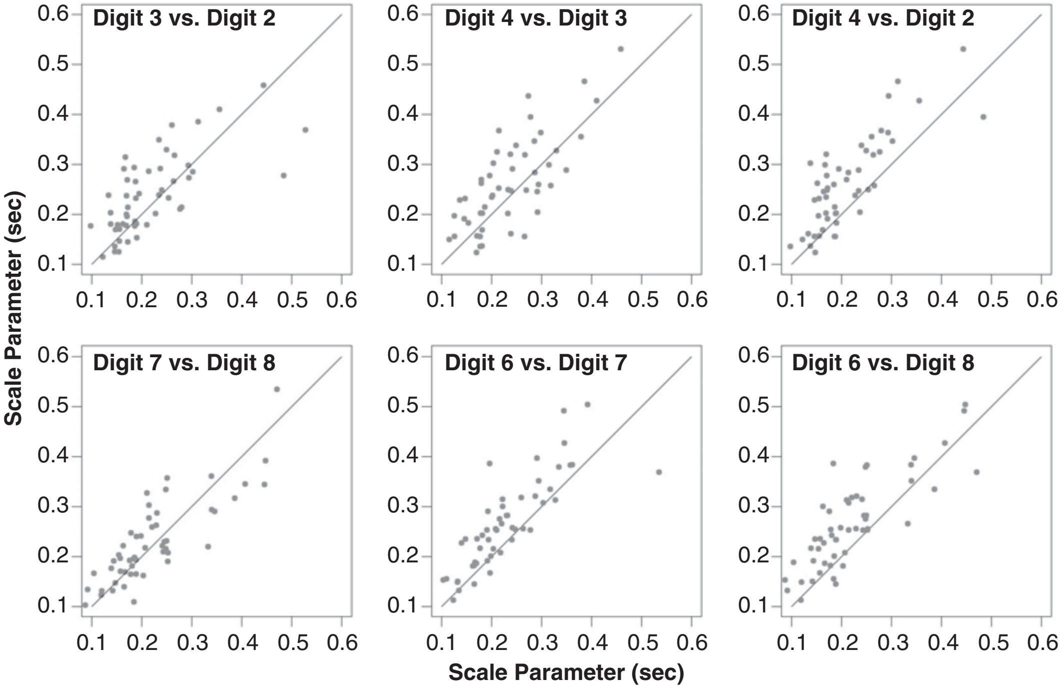

Figure 19. Scatterplots for the scale parameter. Each point corresponds to a particular participant.

7 and 2, respectively). As can be seen, there is very little effect of the variability in shape across digit condition on RT.

It is possible to generate group-level RT distributions from model-based parameter estimates. The method is straightforward. We evaluate the Weibull density function using parameters averaged across participants (indicated by the lines in the top row of panels in Figure 16). The resulting group-level RT distributions are shown in Figure 17.

A more refined analysis of participant and condition effects is shown in Figures 18–20. Each figure is composed of a series of scatterplots. Figure 18 is for the shift parameter, and each scatterplot shows a participant's shift in one condition as a function of his or her shift in another condition. Each point corresponds to a particular participant. As can be seen, there is a high degree of correlation, indicating that the participants who had a high shift in one condition tended to have a high shift in another. The fact that points cluster near the diagonal indicates that there is little systematic effect of condition. One way of characterizing the shift estimates is that there are large participant-specific main effects, as well as more modest participant × condition interactions. There is no digit condition main effect.

Figure 19 shows scatterplots for the scale parameter. Once again, there is a fair amount of correlation, indicating the presence of participant-specific main effects. In addition, main effects of condition are evident in this plot. The y-axis of the plots always represents a digit closer to 5, whereas the x-axis always represents a digit further from 5. The fact that most points cluster above the diagonal indicates greater scale values for digits closer to 5. Figure 20 shows scatterplots for the shape parameter. Here, there is little systematic variation. The interpretation is that there are only participant × condition interactions but no discernible main effects. This plot confirms the conclusion that the systematic effects of digit condition are not in the shape parameter. It suggests that all the participant × item combinations may have similar if not identical shape-parameter values.

An Additive Hierarchical Model

In the original model, there is a separate parameter estimate for each participant × condition combination. The graphical analyses indicate that a more parsimonious model may be obtained by modeling a main effect of symbolic distance on scale. In this section, we will implement such a model. The goal is to provide an additive model that provides a means of estimating the main effects of symbolic distance, as well as testing whether this main effect is statistically significant.

We provide an additive model with main effects of participants and conditions. Ideally, this additive model would be placed on the scale parameter, but the analysis of this model appears intractable.[23] To meet the goal, we adopt an alternative parameterization of the Weibull. Previously, the Weibull was parameterized with shift, scale, and shape. But, in our additive model, we parameterized the Weibull in terms of shift, rate, and shape. The details of the change of parameterization are given in Appendix C. In cases in which there is no systematic

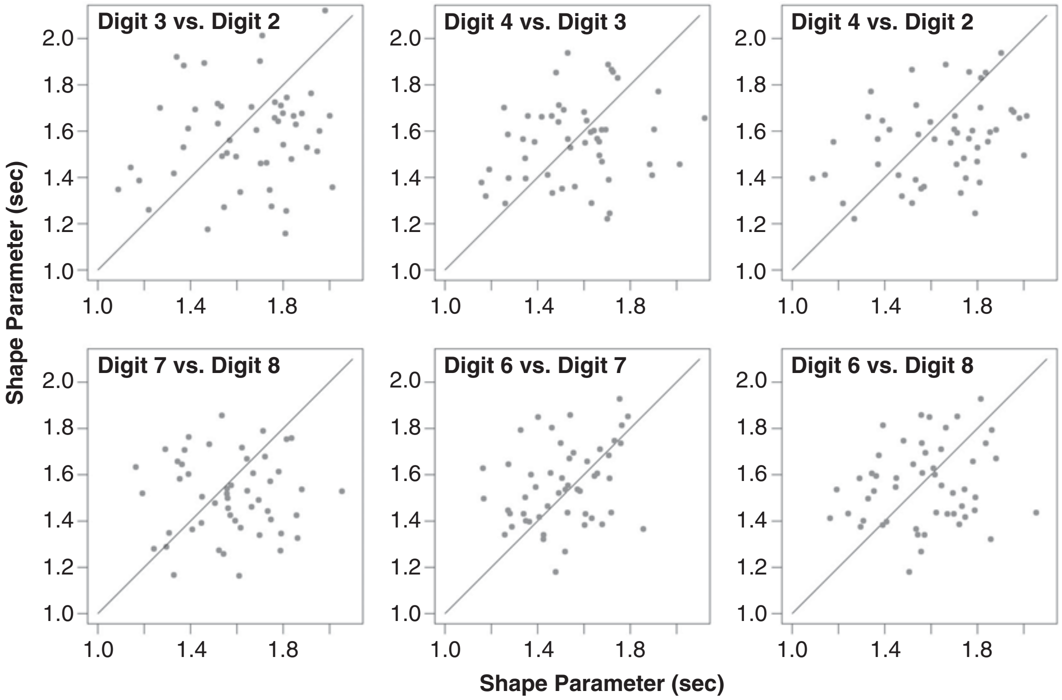

Figure 20. Scatterplots for the shape parameter. Each point corresponds to a particular participant.

change in shape across conditions, rate is inversely related to scale. The additive model is placed on rate, rather than on scale, and is given by:

$$\text{Log(Rate)} = \text{Grand Effect} + \text{Participant Effect} + \text{Condition Effect} + \text{Noise}.$$

The key concept is that there is a main effect of digit condition on the logarithm of rate. This main effect, which is separate from and added to the participant-specific effect, serves as our parameter of interest. The complete model is provided in Appendix C. This model is one of a family of additive models we have presented for psychological data (Lu, 2004; Lu, Sun, Speckman, & Rouder, 2005), and further statistical details have been presented there. Peruggia, Van Zandt, and Chen (2002) presented a similar approach with the two-parameter Weibull (shift set to zero), in which they also placed a linear model on logarithm of the rate.

Before discussing the analysis, we will justify the choice of placing an additive model on the logarithm of the rate parameter, rather than on the rate parameter itself. Once again, the reason for doing so is computational tractability. It may prove quite difficult to estimate other additive models. Fortunately, an additive model on the logarithm of rate is indicated by the previous model analysis. Figure 21 shows the values of the logarithm of rate for the different conditions. The plots in the figure are analogous to those previously displayed for shift, scale, and shape parameters. An additive model is indicated because the points tend to fall on a line parallel to the diagonal. The distance of this line from the diagonal indicates the size of the condition effect.

Figure 22 shows the resulting parameter estimates for the main effect of digit. The points are the estimates (mean of the posterior distributions), and the error bars are 95% credible intervals. Credible intervals in Bayesian statistics are analogous to confidence intervals in classical statistics. As can be seen, rate is affected by numerical distance; it is largest (quickest RTs) for the digits far from 5 and smallest (slowest RTs) for digits near 5.

To test the statistical significance of the distance main effect, we will employ the Bayes factor method used by Lu (2004; Lu et al., 2005). The Bayes factor approach has been discussed extensively in the statistical literature (Kass & Raftery, 1995; Meng & Wong, 1996) and has been imported to the psychological literature (e.g., Pitt et al., 2002). The Bayes factor is the odds that one model is true relative to another, given the data (and the priors). In our case, the first hypothesis is that there are nonzero main effects, and the second one is that all main effects are zero. The odds for the first hypothesis relative to the second are 4.1×10^{14}; hence, we can safely conclude that the main effects are significant. Details of Bayes factor computation can be found in Lu et al.; the additive model discussed here corresponds to their unstructured model.

Discussion

The analyses above provide a locus for the symbolic distance effect: It is largely in scale (rate) and not in shape. The results are not consistent with a theory that postulates that the effect is due to a processing change, such as the insertion of recheck stages for numbers close to 5. Some caution is necessary for an outright rejection

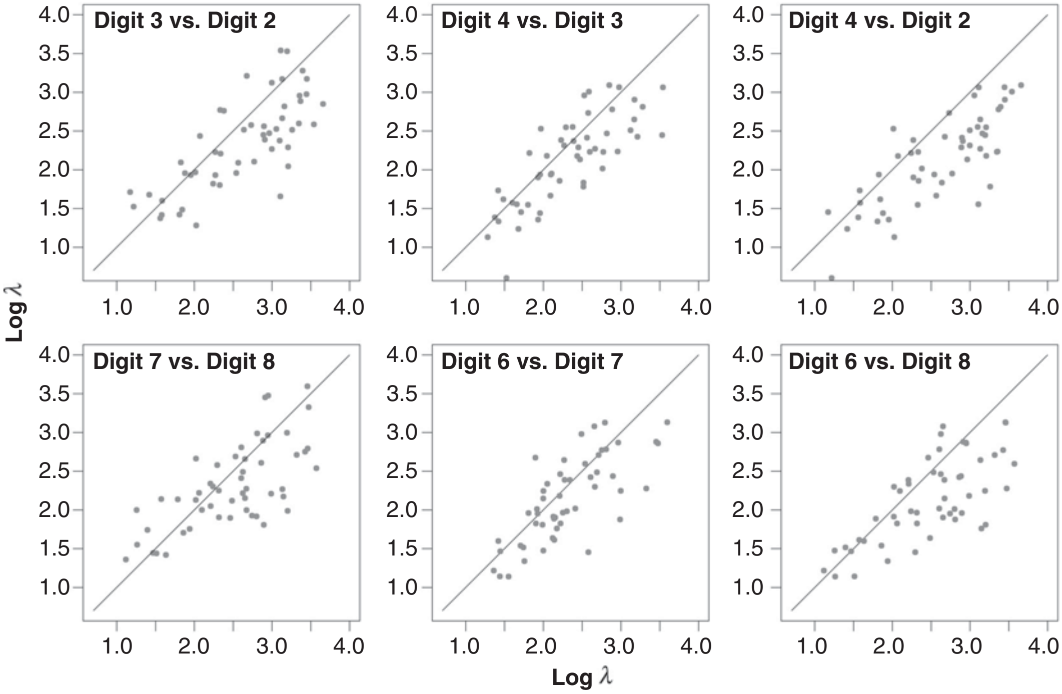

Figure 21. Scatterplots for logarithm of the rate parameter. Each point corresponds to a particular participant.

of theories of architecture change, such as rechecking. The present analysis reveals that if shapes vary systematically across conditions, they do so slightly. It may be possible to construct an architecture change theory that yields a sufficiently small shape effect, so as not to be contradicted by the present analysis.

Symbolic distance is not typically modeled as an architecture change; instead, it is modeled as a diffusion process or random walk (e.g., Link, 1990; Poltrock, 1989; Schwarz, 2001; Smith & Mewhort, 1998). Changes in scale are broadly consistent with either a change in drift rate or a change in bounds. For a small change in drift rate, such as the type needed to produce the 50-msec effects observed here, the changes in shape are minimal (see Figure 9). The results indicate that the effect can be accounted for parsimoniously with a model that postulates that symbolic distance affects scale (speed) of processing, rather than functional architecture.

GENERAL DISCUSSION

Summary

In this article, we have presented a framework for estimating the shift, scale, and shape of RT distributions. The framework is parametric, hierarchical, and Bayesian. It is suited for cases in which there are several participants but only a few observations per participant per condition. The main advantage of a hierarchical model is that it allows for the pooling of information across several participants. The Weibull hierarchical model has four advantages. (1) It allows for superior parameter estimation. (2) The Weibull parameters can be interpreted at several levels, including that of a process-oriented stage model. At this level, differences in shift index differences in peripheral processes, differences in scale index differences in central processing speed, and differences in shape index differences in processing architecture. (3) The hierarchical Weibull model fits the data well. (4) The model is fairly robust to misspecification. The main reason we prefer the Bayesian approach is tractability. Although statistical analysis with Bayesian methods is not simple, it is feasible. We do not know how to analyze the models presented here with classical methods.

Application of the method to a symbolic distance effect revealed a clear locus for the effect in scale. The ensuing interpretation is that increasing numerical distance increases the rate of processing but does not change the form or architecture of the underlying process.

The question of addressing shape need not be done in the context of the Weibull, the ex-Gaussian, or any other parametric form. It can be done in a nonparametric manner by studying higher order distributional properties, such as skew or interquartile skew. We recommend parametric, rather than nonparametric, analysis for increased power. The hierarchical implementation presented here increases the accuracy of parameter estimates in a principled manner.

The General Benefit of Hierarchical Models in Cognition

We believe that there are several domains in which hierarchical models can be of service. The strength of this

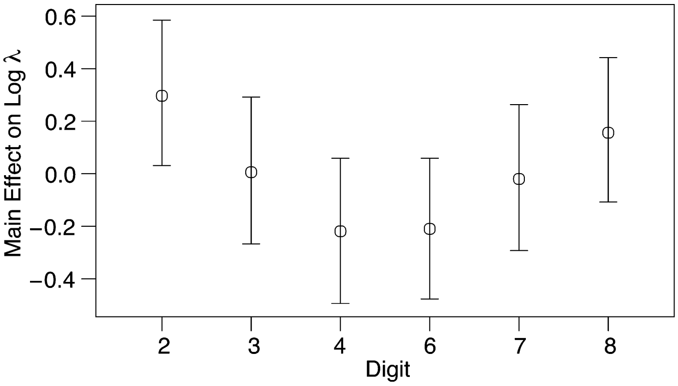

Figure 22. Main effect of digit condition on rate. Error bars denote 95% credible intervals from the posterior distributions of parameter estimates.

approach is that it provides a principled and powerful method of pooling data across disparate individuals. We have illustrated how they can be used to assess manipulation effects on RT distribution properties in cases in which participants are assumed to have their own unique shift, scale, and shape. Below, we will mention a few other domains in which future hierarchical approaches may prove valuable.

Learning is one such domain. There are several theoretical and methodological issues involved in describing the speeding of tasks from practice. Although the power law of practice has been widely believed to describe this effect, it has recently come under considerable scrutiny (e.g., Heathcote et al., 2000). The power law states that RT decreases as a power function of the number of practice trials. Other alternatives are that practice follows an exponential (Heathcote, Brown, & Mewhort, 2002), follows a mixture of power laws (Rickard, 1997), or makes a sharp transition (Haider & Frensch, 2002). A number of authors have pointed out that averaging RT across individuals will distort the shape of empirical-learning curves, and these distortions tend to artificially favor a power law interpretation (Brown & Heathcote, 2003; Estes, 1956; Haider & Frensch, 2002; Heathcote et al., 2000; Myung, Kim, & Pitt, 2002). Hence, studying participant-averaged curves provides a misleading picture of how each individual's RT changes with practice.

We have presented a set of hierarchical models of learning similar to that for symbolic distance (Lu et al., 2005). Each individual's RT on each trial is described by a three-parameter Weibull, with parameters drawn from parent distributions. Practice is postulated to affect the scale parameter, and the form of this effect, whether as a power function or as an exponential, is assessed. We are in the process of analyzing practice effects in an alphabet arithmetic task (Logan, 1988) with these models (Rouder, Lu, Sun, Speckman, & Morey, 2003).

Hierarchical models also may be useful in domains in which there is variability over items, such as in verbal learning or memory experiments. There has long been a concern about unaccounted variability from stimulus items in ANOVAs. Clark (1973) argued that unaccounted variance from items in a memory test could inflate the true Type I error rate. Fortunately, Wickens and Keppel (1983) showed that this type of variability represents only a minor concern in well-counterbalanced experiments. The situation, however, is not as sanguine with regard to nonlinear models, such as the process dissociation procedure (Jacoby, 1991) or other sequential stage type multinomial models (e.g., Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). Curran and Hintzman (1995) have pointed out that variability across items or individuals can greatly bias estimation and inference (see also Ashby, Maddox, & Lee, 1994; Luce, 1959). Their critique was aimed at Jacoby's process dissociation model, a model that seeks to isolate the effects of conscious recollection from automatic forms of recognition, such as feelings of familiarity. Curran and Hintzman's critiques center on latent covariation in psychological processing. They speculate that participants who are better at recalling items from conscious recall may be better at recalling items from familiarity. Likewise, items that are more easily consciously recalled may give rise to greater feelings of familiarity. They show that these types of covariation will bias estimates.

Unfortunately, Curran and Hintzman's (1995) critique is applicable to just about every nonlinear model. For example, Rouder and Batchelder (1998) proposed a multinomial model for separating storage and retrieval factors in memory. The effectiveness of the model, however, is undermined if items that are more easily stored are also more easily retrieved. Psychologists, in general, have not sufficiently addressed the possibility of the deleterious effects of unaccounted variability and correlation in nonlinear models.

We believe that hierarchical modeling provides an attractive means of accounting for and assessing variability and correlation at several different levels. In particular, correlation among parameters across individuals or items may be modeled at the level of parent distributions. For example, in the case of the process dissociation procedure, correlations in recollectability and familiarity across items or participants may be modeled by assuming that the participant and the item effects are sampled from a correlated bivariate distribution. Theoretically, the degree of correlation would be a free parameter; one would obtain not only corrected estimates, but an estimate of correlation as well.

Although the hallmark of experimental psychology is rigorous control, there are always sources of variability that cannot be reduced. When these sources occur simultaneously at different levels, researchers can gain better statistical control with hierarchical models. Bayesian analysis is well suited to hierarchical models and can often provide a tractable means of inference. In this article, we have postulated a hierarchical model to account for both within-subjects and between-subjects variability in RT. We believe that the same modeling approach may be applicable across a number of domains within cognitive and perceptual psychology.

REFERENCES

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, & Cognition*, **27**, 514-544.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**, 144-151.

Ashby, F. G., Tien, J.-Y., & Balakrishnan, J. D. (1993). Response time distributions in memory scanning. *Journal of Mathematical Psychology*, **37**, 526-555.

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

Atkinson, R. C., Holmgren, J. E., & Juola, J. F. (1969). Processing time as influenced by the number of elements in a visual display. *Perception & Psychophysics*, **6**, 321-326.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 340-357.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, **128**, 32-55.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, **6**, 57-86.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, **35**, 11-21.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, **100**, 432-459.

Cheng, R. C. H., & Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B*, **45**, 394-403.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, **12**, 335-359.

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

Cousineau, D., Goodman, V. W., & Shiffrin, R. M. (2002). Extending statistics of extremes to distributions varying in position and scale and the implications for race models. *Journal of Mathematical Psychology*, **46**, 431-454.

Curran, T. C., & Hintzman, D. G. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 531-547.

Dey, D., Ghosh, S., & Mallick, B. (2000). *Generalized linear models: A Bayesian perspective*. New York: Dekker.

Dolan, C. V., van der Maas, H. L. J., & Molenaar, P. C. M. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods, Instruments, & Computers*, **34**, 304-323.

Dzhafarov, E. N. (1992). The structure of simple reaction time to step-function signals. *Journal of Mathematical Psychology*, **36**, 235-268.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, **2**, 312-329.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

Estes, W. K. (1956). The problem of inference from curves based on grouped data. *Psychological Bulletin*, **53**, 134-140.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, **66**, 271-288.

Gelfand, A., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Green, D. M., Smith, A. F., & von Gierke, S. M. (1983). Choice reaction time with a random foreperiod. *Perception & Psychophysics*, **34**, 195-208.

Haider, H., & Frensch, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 392-406.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, **108**, 356-388.

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

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185-207.

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

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

Hockley, W. E. (1984). Analysis of reaction time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 598-615.

Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics*. New York: Macmillan.

Hohle, R. H. (1965). Inferred components of reaction time as a function of foreperiod duration. *Journal of Experimental Psychology*, **69**, 382-386.

Hsu, Y. F. (1999). *Two studies on simple reaction times: I. On the psychophysics of the generalized Pieron's law. II. On estimating minimum detection times using the time estimation paradigm.* Unpublished doctoral dissertation, University of California, Irvine.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, **30**, 513-541.

Jiang, Y. (2002). *Statistical approaches to forming group RT distributions*. Unpublished master's thesis, University of Missouri, Columbia.

Jiang, Y., Rouder, J. N., & Speckman, P. L. (2004). A note on the sampling properties of the Vincentizing (quantile averaging) procedure. *Journal of Mathematical Psychology*, **48**, 186-195.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

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

Kosslyn, S. M. (1975). Information representation in visual images. *Cognitive Psychology*, **7**, 341-370.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Lee, P. M. (1997). *Bayesian statistics: An introduction*. New York: Wiley.

Lehmann, E. L. (1991). *Theory of point estimation*. Pacific Grove, CA: Wadsworth & Brooks.

Link, S. W. (1975). The relative judgement theory of two choice response time. *Journal of Mathematical Psychology*, **12**, 114-135.

Link, S. W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology*, **34**, 2-41.

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

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

Lu, J. (2004). *Bayesian hierarchical models and applications in psychological research*. Unpublished doctoral dissertation, University of Missouri, Columbia.

Lu, J., Sun, D., Speckman, P. L., & Rouder, J. N. (2005). *Hierarchical models for practice-effects in response time*. Manuscript submitted for publication.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

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

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, **57**, 203-220.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, **88**, 375-407.

Meng, X., & Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, **6**, 831-860.

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, **13**, 1080-1084.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, **215**, 1519-1520.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, **47**, 90-100.

Myung, I. J., Kim, C., & Pitt, M. A. (2002). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, **28**, 832-840.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

Peruggia, M., Van Zandt, T., & Chen, M. (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. In C. Gatsonis, R. E. Kass, A. Carriquiry, A. Gelman, D. Higdon, D. K. Pauler, & I. Verdinelli (Eds.), *Case studies in Bayesian statistics* (Vol. 6, pp. 319-334). New York: Springer-Verlag.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, **72**, 346-354.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

Poltrock, S. E. (1989). A random walk model of digit comparison. *Journal of Mathematical Psychology*, **33**, 131-162.

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

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

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

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model analysis of letter masking. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 127-140.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, **126**, 288-311.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measure of cognitive processes. *Psychological Review*, **95**, 318-339.

Rouder, J. N. (2000). Assessing the roles of change discrimination and luminance integration: Evidence for a hybrid race model of perceptual decision making in luminance discrimination. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 359-378.

Rouder, J. N. (2001). Testing evidence accrual models by manipulating stimulus onset. *Journal of Mathematical Psychology*, **45**, 334-354.

Rouder, J. N. (2004). Modeling the effects of choice-set size on the processing of letters and words. *Psychological Review*, **111**, 80-93.

Rouder, J. N., & Batchelder, W. H. (1998). Multinomial models for measuring storage and retrieval processes in paired-associate learning. In C. Dowling, F. Roberts, & P. Theuns (Eds.), *Recent progress in mathematical psychology* (pp. 195-225). Hillsdale, NJ: Erlbaum.

Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., & Morey, R. D. (2003, November). *A statistical model of skill acquisition*. Poster presented at the 20th Annual Meeting of the Psychonomic Society, Vancouver.

Rouder, J. N., Ratcliff, R., & McKoon, G. M. (2000). A neural network model of priming in object recognition. *Psychological Science*, **11**, 13-19.

Rouder, J. N., & Speckman, P. L. (2004). An evaluation of the Vincentizing method of forming group-level response time distributions. *Psychonomic Bulletin & Review*, **11**, 419-427.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, **68**, 589-606.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, **84**, 1-66.

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

Sheu, C.-F., & O'Curry, S. L. (1998). Simulation-based Bayesian inference using BUGS. *Behavior Research Methods, Instruments, & Computers*, **30**, 232-237.

Smith, D. G., & Mewhort, D. J. K. (1998). The distribution of latencies constrains theories of decision time: A test of the random-walk model using numeric comparison. *Australian Journal of Psychology*, **50**, 149-156.

Speckman, P. L., & Rouder, J. N. (2004). A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. *Psychonomic Bulletin & Review*, **11**, 574-576.

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 461-479.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, **153**, 652-654.

Tanner, M. A. (1993). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. Berlin: Springer-Verlag.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, **51**, 599-606.

Theeuwes, J. (1994). Stimulus-driven capture and attentional set: Selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 799-806.

Thomas, E. A. C., & Ross, B. (1980). On appropriate procedures for combining probability distributions within the same family. *Journal of Mathematical Psychology*, **21**, 136-152.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

Townsend, J. T., & Nozawa, G. (1995). On the spatio-temporal properties of elementary perception: An investigation on parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, **39**, 321-359.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97-136.

Tversky, A., & Kahneman, D. (1990). Judgements under uncertainty: Heuristics and biases. In P. K. Moser (Ed.), *Rationality in action:*

Contemporary approaches (pp. 171-188). New York: Cambridge University Press.

Ulrich, R., & Miller, J. (1994). Effects of truncation of reaction time analysis. *Journal of Experimental Psychology: General*, **123**, 34-80.

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

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

Vickers, D. (1980). Discrimination. In A. Welford (Ed.), *Reaction times* (pp. 25-72). London: Academic Press.

Vincent, S. B. (1912). The function of vibrissae in the behavior of the white rat. *Behavioral Monographs*, **1**(Whole No. 5).

Wang, X. H., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, **26**, 109-128.

Wickens, T. D., & Keppel, G. (1983). On the choice of design and test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning & Verbal Behavior*, **22**, 296-309.

Yellott, J. I. (1969). Probability learning with noncontingent success. *Journal of Mathematical Psychology*, **6**, 541-575.

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

NOTES

1. To assess whether the displayed variation in distributional properties reflects differences in underlying true properties or sample noise, we fit a Weibull distribution to individuals' RT data. Two models were fit: a general one in which individuals had their own free parameters, and a restricted one in which a distributional property was equated across pairs of participants. To test whether the difference in shift in the left panels is significant, both participants' shifts were equated. The resulting log-likelihood test revealed that the shifts are statistically different [$G^2(1) = 53.2, p < .05$]. Analogous tests were performed on the difference in scale (center panels) and shape (right panels) [$G^2(1) = 113.7, p < .05$, and $G^2(1) = 26.1, p < .05$, for scale and shape, respectively].

2. The density of the three-parameter Weibull is given by

$$f(t \mid \psi,\theta,\beta) = \frac{\beta(t-\psi)^{\beta-1}}{\theta^{\beta}} \exp\left(-\frac{(t-\psi)^{\beta}}{\theta^{\beta}}\right)$$

for $\theta, \beta > 0$ and $\psi \le t$.

3. Concepts of shift, scale, and shape can be given precise meanings. Let the density of a random variable exist everywhere and be expressed as $f(t \mid \Theta_1, \ldots, \Theta_n)$, where $\Theta_1, \ldots, \Theta_n$ are the parameters. Let

$$z = \frac{t-\theta_1}{\theta_2}.$$

We refer to the density as being in location–scale form if there exists some function g such that

$$f(t \mid \Theta_1, \ldots, \Theta_n) = \Theta_2^{-1}\, g(z \mid \Theta_3, \ldots, \Theta_n). \qquad (1)$$

If Equation 1 holds, then Θ_1 is referred to as the location parameter, and Θ_2 as the scale parameter. Parameters Θ_3 through Θ_n are the shape parameters. Many random variables have densities that can be expressed in location–scale form. For the normal, for example, $\Theta_1 = \mu$, $\Theta_2 = \sigma$, and $g(z) = (2\pi)^{-1} \exp(-z^2/2)$. The location parameter corresponds to the mean, the scale parameter corresponds to the standard deviation, and there are no shape parameters. For the exponential, $\Theta_1 = 0$, $\Theta_2 = \lambda^{-1}$, and $g(z) = e^{-z}$. The scale parameter corresponds to the inverse of rate, and there are no shape or location parameters. The three-parameter Weibull can be expressed in location-scale form; $\Theta_1 = \psi$, $\Theta_2 = \theta$, and $g(z \mid \beta) = \beta z^{\beta-1} \exp(-z^{\beta})$. In general, the location parameter may correspond to any of a number of characteristics of a distribution, including the mean, mode, or point at which a distribution first attains mass. For the Weibull, the location parameter defines where the distribution first attains mass. We refer to the location parameter as the *shift* parameter, since it aptly describes how changes in location affect the density function.

There is an asymmetric relationship between location, scale, and shape parameters and the central moments of a distribution. Changes in location certainly imply changes in mean, but not in moments of higher order than the mean. Changes in scale certainly imply changes in variance, but not in moments of higher order than variance. Changes in scale, in general, may imply changes in the mean too. For example, the exponential has a single parameter, the rate, which is a scale parameter. Increasing the rate not only decreases the variance, but also decreases the mean as well. Changes in shape, in general, imply changes in moments higher than variance.

4. We use the term *parent distribution* because it emphasizes the hierarchical relations among levels of variability. Parent distributions are also known as *latent-trait distributions* (as in item response theory). The equivalence of the terms is realized by conceptualizing parameters as personal traits.

5. It is possible to place a hierarchical prior on the shift parameter. We do not do so for computation simplicity. A hierarchical model on shift would entail using a Metropolis–Hastings step in Gibbs sampling. Implementation of this step would present a number of additional computational concerns.

6. Participants had to identify in which of four locations an asterisk was presented by depressing one of four keys. Each participant responded to 120 such trials. The last 80 correct observations between 200 and 1,200 msec served as data.

7. The term *quantile* is used by statisticians to generalize *median* and *percentile*. The *p*th quantile of a distribution is the value for which a probability of observing an observation below the value is p. We refer to p as the *probability* of the quantile—for example, the .1 quantile is the same as the 10th percentile.

8. Previously, we have shown that Heathcote et al. (2002) maximized a function that was not the likelihood of the parameters, given the sample quantiles (Speckman & Rouder, 2004). However, we note that in some cases, maximizing their function works better than maximizing the likelihood itself. Here, we competitively test Heathcote et al.'s (2002) function in estimating Weibull parameters.

9. Heathcote et al.'s (2002) method involves constructing bins and assigning counts to them. We followed their algorithms for doing this. For 80 observations, there are 81 bins. The first bin has a right-hand bound at the smallest observation and is assigned half a count. The intermediate 79 bins are defined by successive order statistics and are assigned a single count each. The last bin has a left-hand bound at the largest observation and is assigned half a count. Other bin constructions are possible with regard to the extreme bins. Alternative construction will have a marginal effect on estimates.

10. The *i*th ranked observation serves as a sample quantile for $p = i / (M + 1)$, where M is the number of observations. This formula for estimating sample quantiles is commonly used in statistical software packages (e.g., SAS's Proc-Univariate).

11. There are a few variants of the Vincentizing procedure, as has been discussed by Heathcote (1996). In this article, we average sample quantiles across individuals. See Heathcote or Van Zandt (2000) for a discussion of alternative averaging methods.

12. The phrase *sufficiently large sample sizes* in the context of estimating group-level parameters refers to both a sufficient number of samples per individual and a sufficient number of individuals.

13. Extreme values are extreme enough to dominate mean absolute error as well. For example, for the ML individual estimates with 20 observations, mean absolute error was on the order of 5,000.

14. The terms *central* and *peripheral* do not necessarily refer to locations in the body or the brain. It is known that central processes are located in the cortex, but peripheral processes may be cortical or subcortical. Peripheral processes include processes that may occur outside of the brain—for example, the processes of encoding sensory stimulation or producing motor processes.

15. We analyzed the accuracy–stress condition of Ratcliff and Rouder's (1998) data because this condition yielded the largest variation in RT distributions as a function of luminance.

16. In Ratcliff and Rouder's (1998) experiments, 3 participants observed about 10,000 trials each. With such large individual sample sizes and few participants, ML and HB methods yield near identical estimates.

17. Per condition, 10,000 observations were simulated.

18. Colonius (1995) provides a critique of this justification. His main concern is that the minimum of random variables may degenerate to a constant well before the Weibull shape is observed. More recently, Cousineau, Goodman, and Shiffrin (2002) have shown that this degeneracy is avoided if the constituent distributions are not identically distributed but have variability in the location and scale parameters.

19. The four degrees of freedom come about as follows. The data are partitioned into eight bins, yielding seven overall degrees of freedom. The model has three parameters; hence, the total degrees of freedom left over for fit is four.

20. Estimation error bound intervals are obtained by adding ±1.96 × $\sqrt{\{F(x)[1 - F(x)]/80\}}$, where $F(x)$ is the theoretical cumulative distribution function. The assumption behind this formula is that the sample distribution of the cumulative distribution function is normal, instead of binomial. Although this approximation is quite good for most of the range of probabilities, it is less valid at the extremes. The small deviations of the error bounds above one and below zero are a consequence of the normal approximation.

21. The density of the ex-Gaussian is expressed in location–scale form as $f(t; \mu, \sigma, \eta) = \sigma^{-1}g(z, \eta)$, where $z = (t - \mu) / \sigma$ and

$$g(z,\eta)=\frac{\exp\left(z\eta^{-1}\right)+.5\eta^{-2}}{\eta}\Phi\left(z-\eta^{-1}\right).$$

Φ is the cumulative distribution function for the standard normal. In accordance with note 2, μ, σ, and η are location, scale, and shape parameters, respectively.

22. In the simulations, each individual's parameters were sampled from parent distributions. For Simulation 1, the group difference was in location, with $\mu_1 \sim$ uniform(1.2, 1.8) and $\mu_2 \sim$ uniform(1.5, 2.1). For both groups, $\sigma \sim$ uniform(0.1, 0.3) and $\eta \sim$ lognormal (0.69, 0.2). (In the lognormal, the first and second parameters are the mean and variance of the normal before exponentiation, respectively.) For Simulation 2, the group difference was in scale, with $\sigma_1 \sim$ uniform(0.1, 0.3) and $\sigma_2 \sim$ uniform(0.3, 0.5). For both groups, $\mu \sim$ uniform(1.2, 1.8) and $\eta \sim$ lognormal(0.69, 0.2). For Simulation 3, the group difference was in shape, with $\eta_1 \sim$ lognormal(0, 0.2), and $\eta_2 \sim$ lognormal(1.39, 0.2). For both groups, $\mu \sim$ uniform(1.2, 1.8), and $\sigma \sim$ uniform(0.1, 0.3).

23. We spent several months attempting to devise a suitable hierarchical model on scale. None of our attempts yielded a set of priors capable of the robust estimation. This failure provided the motivation to use the shift, rate, shape parameterization of the Weibull.

APPENDIX A

The basic model is described in this Appendix. A more detailed treatment can be found in Rouder, Sun, et al. (2003). Each participant provides a series of observations. Let y_{ij} denote the RT of Participant i on Trial j ($1 \leq i \leq I$; $1 \leq j \leq J_i$). Each observation is assumed to be independent and identically distributed from a three-parameter Weibull distribution with density

$$f\left(y_{ij} \mid \psi_i,\theta_i,\beta_i\right)=\frac{\beta_i\left(y_{ij}-\psi_i\right)^{\beta_i-1}}{\theta_i^{\beta_i}}\exp\left[-\frac{\left(y_{ij}-\psi_i\right)^{\beta_i}}{\theta_i^{\beta_i}}\right], \text{ for } y_{ij}>\psi_i. \tag{A1}$$

Parameters ψ_i, θ_i, and β_i denote the shift, scale, and shape of the ith participant, respectively. The parameters are assumed to come from a prior distribution. The prior on each individual's shift parameter ψ_i is a uniform distribution from zero to some large number A. The precise value of A is unimportant as long as it is greater than the minimum value for the ith participant. The other two parameters, shape and scale, are assumed to be samples from a two-stage hierarchical prior distribution, whose first stage prior is given by

$$\left(\beta_i \mid \eta_1,\eta_2\right) \overset{iid}{\sim} \text{Gamma}\left(\eta_1,\eta_2\right) \text{ restricted to } \beta_i>0.01, \tag{A2}$$

$$\left(\theta_i^{-\beta_i} \mid \xi_1,\xi_2\right) \overset{iid}{\sim} \text{Gamma}\left(\xi_1,\xi_2\right), \tag{A3}$$

where Gamma(η_1, η_2) denotes the gamma distribution with density

$$f\left(t \mid \eta_1,\eta_2\right)=\eta_2^{\eta_1}t^{\eta_1-1}\exp\left(-\eta_2 t\right)/\Gamma\left(\eta_1\right) \text{ for } t>0.$$

For the first stage, the prior for the shape parameter β_i is a gamma distribution with parameters η_1 and η_2 restricted to the range $\beta_i > 0.01$. This somewhat unusual restriction is a technical one needed to ensure that posterior moments exist for θ_i. This choice of priors is quite general, in the sense that it is flexible and can model a reasonably broad class of prior distributions. Importantly, the choice is convenient and tractable, in that all of the Gibbs sampling can be done without a more computationally extensive Metropolis–Hastings step.

The parameters (ξ_1, ξ_1, η_1, η_2) of the first-stage prior serve as *hyperparameters*. These hyperparameters describe how the shape and scale vary across individuals within the population. The second-stage prior is given by mutually independent distributions:

$$\xi_k \sim \text{Gamma}(a_k, b_k),\ k = 1, 2, \tag{A4}$$

$$\eta_k \sim \text{Gamma}(c_k, d_k),\ k = 1, 2. \tag{A5}$$

The prior values used in fitting are as follows: $a_1 = 2.0$, $b_1 = 0.1$, $a_2 = 2.0$, $b_2 = 2.85$, $c_1 = 1.0$, $d_1 = 0.02$, $c_2 = 2.0$, $d_2 = 0.04$. These priors are somewhat informative and were chosen on the basis of our general experience with RT distributions. In particular, our goal was to have broad coverage of shape in the range from 1 to 5 (these values are typical in Logan, 1992) and broad coverage of scale in the range of 0 to 0.4 sec.

APPENDIX A (Continued)

The marginal prior distributions over shape and scale are shown in Figure A1. Included in this figure are two representative posterior distributions as well. Two points are evident: First, the priors are broadly distributed over a range of plausible values, and second, the posterior is fairly narrow and centered away from the mass of the prior. This is an initial indication that the posterior is not unduly influenced by the choice of prior. Rouder, Sun, et al. (2003) provide a more detailed analysis in which they manipulate the priors. The conclusion is the same; with these choices, the posterior distribution mostly reflects the influence of the data, rather than the specification of the prior. Parameter estimation is done through Monte Carlo Markov chain methods (see Gelfand & Smith, 1990, for a review, or Lee, 1997, for a basic introduction to Bayesian methods). Rouder, Sun, et al. (2003) have provided a detailed discussion of the implementation, as well as a discussion of the burn-in period and chain convergence.

APPENDIX B

This Appendix describes the general model for the numerical similarity experiment. The model is a simple extension of that in Appendix A. Let y_{ijk} denote the kth RT for the jth participant in the ith condition. Each observation is assumed to be independent and identically distributed from a three-parameter Weibull distribution:

$$y_{ijk} \sim \text{Weibull}(\psi_{ij}, \theta_{ij}, \beta_{ij}). \tag{B1}$$

Priors on parameters are

$$\psi_{ij} \overset{iid}{\sim} \text{Uniform}\left[0, \min_k \left(y_{ijk}\right)\right] \tag{B2}$$

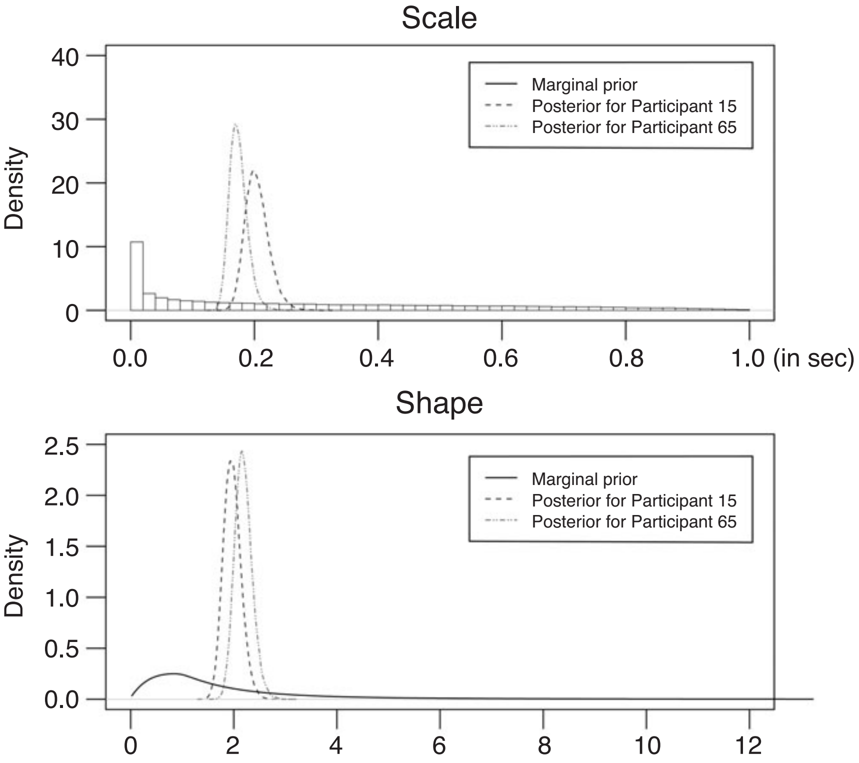

Figure B1. Prior and posterior densities of the scale (θ) and shape (β) parameters for 2 representative participants. Reprinted with permission from Rouder, Sun, Speckman, Lu, and Zhou (2003).

APPENDIX B (Continued)

$$\left(\beta_{ij} \mid \eta_1, \eta_2\right) \overset{iid}{\sim} \text{Gamma}\left(\eta_1, \eta_2\right) \text{ restricted to } \beta_{ij} > 0.01, \tag{B3}$$

$$\left(\theta_{ij}^{-\beta_{ij}} \mid \xi_1, \xi_2\right) \overset{iid}{\sim} \text{Gamma}\left(\xi_1, \xi_2\right), \tag{B4}$$

$$\xi_l \overset{iid}{\sim} \text{Gamma}\left(a_l, b_l\right),\ l = 1, 2, \tag{B5}$$

$$\eta_l \overset{iid}{\sim} \text{Gamma}\left(c_l, d_l\right),\ l = 1, 2. \tag{B6}$$

The hierarchical priors are random effects models in which all participant–digit combinations are assumed to come from a common parent distribution. Values of (a, b, c, d) were the same as those discussed in Appendix A.

APPENDIX C

This Appendix describes a restricted model for the numerical similarity experiment. Let y_{ijk} denote the kth RT for the jth participant in the ith condition. Each observation is assumed to be independent and identically distributed from a three-parameter Weibull distribution:

$$y_{ijk} \sim \text{Weibull}(\psi_{ij}, \lambda_{ij}, \beta_{ij}), \tag{C1}$$

where $\lambda_{ij} = \theta_{ij}^{-\beta_{ij}}$.

To model the effects of symbolic distance in rate, we assume $\log \lambda_{ij} = \mu + \alpha_j + \gamma_i + \epsilon_{ij}$. The rationale is exactly the same as that in ordinary linear models. Parameter α_j is the random effect for the jth participant, γ_i is the effect for the ith condition, and ϵ_{ij} is included for extra unexplained variance.

The priors on the parameters are:

$$\psi_{ij} \overset{iid}{\sim} \text{Uniform}\left[0, \min_k\left(y_{ijk}\right)\right], \tag{C2}$$

$$\left(\beta_{ij} \mid \eta_1, \eta_2\right) \overset{iid}{\sim} \text{Gamma}\left(\eta_1, \eta_2\right) \text{ restricted to } \beta_{ij} > 0.01, \tag{C3}$$

$$\log \lambda_{ij} = \mu + \alpha_j + \gamma_i + \epsilon_{ij}, \tag{C4}$$

$$\eta_l \overset{iid}{\sim} \text{Gamma}\left(c_l, d_l\right),\ l = 1,2, \tag{C5}$$

$$\mu \sim \text{N}\left(u_0, s_\mu\right), \tag{C6}$$

$$\left(\alpha_i \mid \delta_\alpha\right) \overset{iid}{\sim} \text{N}\left(0, \delta_\alpha\right), \tag{C7}$$

$$\delta_0 \sim \text{IG}\left(a_0, b_0\right), \tag{C8}$$

$$\delta_\alpha \sim \text{IG}\left(a_2, b_2\right), \tag{C9}$$

$$\left(\gamma_j \mid \delta_\gamma\right) \overset{iid}{\sim} \text{N}\left(0, \delta_\gamma\right), \tag{C10}$$

$$\delta_\gamma \sim \text{IG}\left(a_3, b_3\right), \tag{C11}$$

where $\text{N}(u, s)$ denotes the normal distribution with mean u and variance s, and $\text{IG}(a_0, b_0)$ denotes the inverse gamma distribution, whose density function is

$$\left[\delta_0 \mid a_0, b_0\right] = \frac{1}{\delta_0^{a_0+1}\Gamma(a_0)} b_0^{a_0} \exp\left(-b_0 / \delta_0\right), \quad \delta_0 > 0.$$

To estimate this model, it is necessary to choose values for parameters of the priors on parent distributions. We chose $u_0 = 0.0$, $s_0 = 50.0$, $a_0 = -.5$, $b_0 = 0.0$, $a_2 = -.5$, $b_2 = 0.0$, $a_3 = -.5$, $b_3 = 0.0$, $c_1 = 2.0$, $d_1 = .02$, $c_2 = 2.0$, $d_2 = .04$. The priors associated with these choices are all fairly noninformative. Details about these choices, the prior distributions, and the estimation of this model are available in Lu et al. (2005).

(Manuscript received July 2, 2003;
revision accepted for publication May 24, 2004.)