Psychonomic Bulletin & Review
2009, 16 (1), 190-203
doi: 10.3758/PBR.16.1.190

Notes and Comment

Comparing time–accuracy curves: Beyond goodness-of-fit measures

Charles C. Liu
University of Melbourne, Melbourne, Victoria, Australia
and Monash University, Melbourne, Victoria, Australia

and

Philip L. Smith
University of Melbourne, Melbourne, Victoria, Australia

The speed–accuracy trade-off (SAT) is a ubiquitous phenomenon in experimental psychology. One popular strategy for controlling SAT is to use the response signal paradigm. This paradigm produces time–accuracy curves (or SAT functions), which can be compared across different experimental conditions. The typical approach to analyzing time–accuracy curves involves the comparison of goodness-of-fit measures (e.g., adjusted-R^2), as well as interpretation of point estimates. In this article, we examine the implications of this approach and discuss a number of alternative methods that have been successfully applied in the cognitive modeling literature. These methods include model selection criteria (the Akaike information criterion and the Bayesian information criterion) and interval estimation procedures (bootstrap and Bayesian). We demonstrate the utility of these methods with a hypothetical data set.

Response time (RT) and response accuracy are the two most common dependent variables in experimental psychology. These two variables are often measured together to assess the possibility of a *speed–accuracy trade-off* (SAT): That is, at a given level of sensitivity, faster responses tend to produce more errors. The SAT phenomenon is often unrelated to the underlying mental processes of interest and, therefore, has been regarded as a nuisance variable in most studies.

One strategy for controlling SAT is to formulate a quantitative model for the decision process. For example, sequential sampling models of decision making account for SAT with a criterion parameter, which determines the amount of information accumulated by the participant before he or she responds (Link, 1992; Ratcliff & P. L. Smith, 2004; P. L. Smith, 2000). This criterion parameter can be estimated simultaneously with other parameters, including those that index the strength and variability of underlying representations. However, full specification of a sequential sampling model requires a number of detailed, and often highly technical, assumptions about the model architecture, and the models are usually fitted to entire distributions of both correct and error RTs. It is not surprising, then, that routine application of sequential sampling modeling is rare in experimental psychology.[1] Here, we will focus on an alternative, and more popular, method known as the *response signal paradigm*.

Response Signal Paradigm

The response signal paradigm is a popular method for controlling SATs (Dosher, 1979; Reed, 1976; Wickelgren, 1977). This method allows the effects of discriminability to be decoupled from those of criterion shifts, because the experimenter specifies the time at which a response must be made. On each trial, observers are instructed to respond as soon as they hear a response signal (an auditory tone), which is presented at one of several deadline lags. At early lags, the observer often responds at close to chance, because there is not enough time to fully integrate the available information. As lag increases, accuracy improves monotonically to an asymptote. This time–accuracy curve is often referred to as the *SAT function*.

When accuracy is measured in d' units, the data are usually fitted by a three-parameter shifted exponential function:

$$\begin{aligned} d'(t) &= \lambda\left[1 - e^{-(t-\delta)/\beta}\right],\ t > \delta, \\ &= 0,\ t \leq \delta, \end{aligned} \tag{1}$$

where λ is the asymptotic accuracy, $1/\beta$ is the rate at which accuracy approaches the asymptote from chance ($d' = 0$), and δ is the time at which accuracy begins to exceed chance.

The response signal paradigm allows us to characterize the effect of experimental manipulations across the entire time course of decision processing. One of the primary reasons for using this paradigm is to compare estimates of processing dynamics across conditions in which asymptotic accuracy varies. In principle, observed differences between conditions may be due to changes in one or more parameters of the shifted exponential function. Such direct inferences are generally not available in other experimental paradigms. For example, SATs are often dismissed when responses in one condition are faster and more accurate than those in another condition. This pattern does not necessarily imply an advantage at asymptote, however, and is also consistent with an advantage due solely to rate, to intercept, or to both (see Dosher, Han, & Lu, 2004, p. 6).

The traditional approach to analyzing time–accuracy curves proceeds as follows. First, a set of nested models is formulated. This set includes a full model that allocates separate asymptote, rate, and intercept parameters to each

C. C. Liu, charles.liu@muarc.monash.edu.au

condition. Other models, which are special cases of the full model, are specified by setting certain parameters to be invariant across conditions. Point estimates for each model are then computed by minimizing a squared-error loss function. The best model is then chosen on the basis of three criteria (e.g., Öztekin & McElree, 2007): (1) the value of an adjusted-R^2 statistic (Reed, 1973), (2) the consistency of the parameter estimates across participants, and (3) evaluation of whether systematic deviations could be accounted for by additional parameters. Once the best model is chosen, the corresponding point estimates are used for inference.

Although this approach has proven, over many years, to be a generally useful one (Dosher, 1981; McElree, 1996, 1998, 2001; McElree & Dosher, 1989, 1993; Wickelgren & Corbett, 1977), we believe that it could be improved upon. First, estimates derived by ordinary least squares will not, in general, correspond to the maximum likelihood estimates, because the variability of d' is not constant across the time–accuracy curve. Second, model selection criteria can be used to quantitatively determine (1) the balance between goodness of fit and parsimony, (2) the consistency of parameter estimates across participants, and (3) whether additional parameters could account for systematic deviations. Finally, the interpretation of point estimates alone can be misleading when there is no measure of uncertainty associated with those estimates. These issues will be expanded upon in the remainder of this article.

Overview

The remainder of this article is divided into five sections. First, we will raise a number of issues that must be addressed before analyzing time–accuracy curves. We also will describe a hypothetical data set in order to demonstrate various analysis methods. Second, the hypothetical data will be analyzed by the traditional method of maximizing adjusted-R^2. Using these adjusted-R^2 statistics, we will compare the full shifted exponential model against simpler nested models. Third, the same models will then be compared using information criteria: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These criteria are preferable for the purposes of model selection because they account for both goodness of fit and model complexity (in terms of number of free parameters) in a principled manner. Finally, we will highlight the importance of parametric uncertainty by calculating interval estimates derived, first, from a bootstrap method and, second, from a Bayesian method. One of the goals of this article is to broaden the range of methods available to experimental psychologists for analyzing response signal data.

ISSUES IN RESPONSE SIGNAL STUDIES

Whenever time–accuracy curves are analyzed, a number of key questions about the nature of the data must first be addressed. These questions include the following: How should individuals within groups be analyzed; what is the dependent variable; what is the independent variable; and what is the functional form of the curve?

Group Analysis

The first question is whether the analysis should be performed on data from each individual participant or on data pooled across participants. When the data are sufficiently precise, it is often preferable to analyze each individual separately; fitting nonlinear functions (such as the shifted exponential) to averaged data may yield distorted estimates that do not reflect the true underlying process for any individual (Brown & Heathcote, 2003; Estes & Maddox, 2005; Myung, Kim, & Pitt, 2000). Because response signal experiments usually involve a small number of participants who complete a large number of trials, analysis of individual curves is the more common practice. The issue of individual differences will be discussed in greater detail.

Dependent Variable (Accuracy)

The second question is what the relevant measure of accuracy should be. In most studies, the accuracy in each condition is calculated in d' units, using the formula

$$d' = z[P_S(C)] + z[P_N(C)], \tag{2}$$

where $z[\cdot]$ denotes the inverse normal (z score) transformation, $P_S(C)$ denotes the proportion of correct responses to signal trials, and $P_N(C)$ denotes the proportion of correct responses to noise trials. When the measure of accuracy is chosen, an important consideration is the sampling variability of that measure. Explicit assumptions about sampling variability are crucial for carrying out the analyses proposed here (see Lee, 2004). For large sample sizes, the sampling distribution of d' is well approximated by a normal distribution with variance calculated as

$$\mathrm{var}(d') = \frac{P_S(C)\left[1 - P_S(C)\right]}{n_S \phi^2\left\{z\left[P_S(C)\right]\right\}} + \frac{P_N(C)\left[1 - P_N(C)\right]}{n_N \phi^2\left\{z\left[P_N(C)\right]\right\}}, \tag{3}$$

where n_S and n_N are the number of signal and noise stimuli, respectively, in each condition and $\phi(\cdot)$ is the standard normal density function evaluated at the specified abscissa (Gourevitch & Galanter, 1967). (Miller [1996] has discussed alternative calculations that are necessary for small sample sizes.) This approximation implies that the sampling variability is not constant across the time–accuracy curve; in fact, sampling variability generally increases as accuracy increases.

Independent Variable (Time)

The third question is whether accuracy should be analyzed as a function of either the deadline lag alone or some measure of total processing time. This dilemma arises because RTs (calculated from response signal onset) are not constant across lags; instead, RTs decrease with increasing lag and remain constant at long lags (see Dosher, 1982, 1984; Reed, 1976). In addition, RTs can vary across conditions. Various mechanisms have been proposed to explain these patterns (Nikolic & Gronlund, 2002; Ratcliff, 2006; Usher & McClelland, 2001), but no definitive consensus has yet emerged. As has been argued by McElree and Dosher (1989), differences in RT across conditions can be controlled for by the use of total processing time (lag + mean RT) as the independent variable. In this way, estimates of processing dynamics remain undistorted.

Functional Form

Finally, a researcher must decide whether to fit the more common shifted exponential function or an alternative, such as the *shifted diffusion* function, derived from the diffusion model (Ratcliff, 1978). According to the shifted diffusion function, the growth of sensitivity over time can be described by the following equation:

$$d' = \frac{A}{\sqrt{1+[R/(t-I)]}},\ t > I,$$
$$= 0,\ t \le I, \tag{4}$$

where A is the asymptotic sensitivity, R is the rate of growth (in time units), and I is the intercept, or the time at which accuracy begins to exceed chance. McElree and Dosher (1989) systematically compared fits of the shifted diffusion and the shifted exponential functions to short-term memory data and found that the exponential function produced slightly better fits. A recent Monte Carlo study has shown that the shifted exponential and shifted diffusion statistically mimic one another (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). These simulations are consistent with empirical results showing that analyses of the two functions often yield similar conclusions (e.g., Gronlund & Ratcliff, 1989; McElree & Griffith, 1995; Mulligan & Hirshman, 1995).

Hypothetical Data

Here, we will present some hypothetical data from a response signal experiment. For concreteness, we will assume that these hypothetical data were obtained from a spatial cuing experiment (e.g., Carrasco & McElree, 2001), although the same ideas could be applied to data from other domains. On a typical trial in this task, a cue is used to direct attention to a particular location. Following a lag that is too short for the eyes to refixate, a target stimulus is presented at either the cued location or an uncued location. Responses to cued targets are often faster and more accurate than responses to uncued targets.

Our hypothetical data are plotted in Figure 1. We assume that these data came from a single participant, so our analysis is at the individual level. We will also make the (unrealistic) assumption that the variability is constant: that each of the observed d' scores was sampled from a normal distribution with a standard deviation of 0.2. (The corresponding standard error bars are plotted in Figure 1.) Following most researchers in this area, the independent variable is defined as the deadline lag plus the mean RT. Finally, although we will analyze only the shifted exponential in this article, all of the methods can be applied in the same way to the shifted diffusion function.

MAXIMIZING GOODNESS OF FIT

In this section, we will attempt to infer, via *parameter estimation*, the true shifted exponential function that generated the hypothetical data. In response signal studies, the conventional method for parameter estimation is to maximize the proportion of variance accounted for by the model. This method is equivalent to least-squares estimation (LSE), which yields the parameter values that minimize the sum-of-squares error between the observations and the predictions. One justification for this square-loss function is that, under certain conditions, parameter values derived from LSE may coincide with those from *maximum likelihood estimation* (MLE). The basic idea behind MLE (see Myung, 2003, for a tutorial) is to seek the parameter values that are most likely to have generated the data. Parameter values from LSE are equivalent to those from MLE when the sampling distributions are normal with constant variance. Because the constant variance assumption will rarely hold in response signal experiments, LSE will yield estimates different from those from MLE. (For normal sampling distributions, MLE will be equivalent to *weighted* LSE, with variances equal to the reciprocal of the weights.)

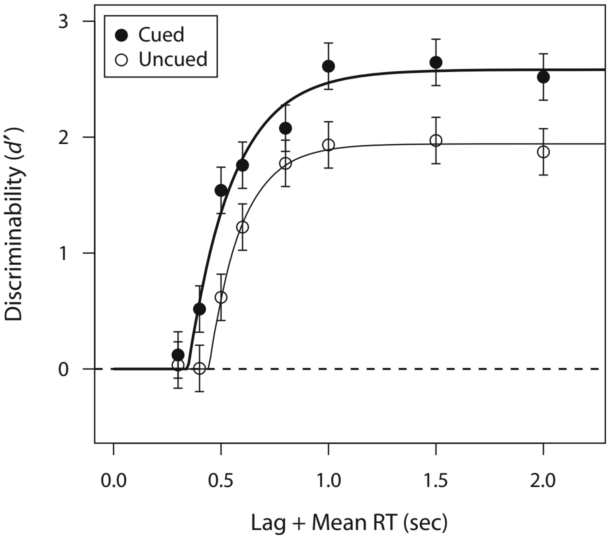

Figure 1. Hypothetical response signal data. The points represent observed d', with standard error bars plotted. The lines are time–accuracy curves derived from the maximum likelihood estimates of the full model. The heavy line represents the cued condition, and the light line represents the uncued condition.

Because we assumed constant variance in our hypothetical data set, the LSE and MLE procedures would give the same results in this case. Nonetheless, we will use MLE here because this procedure will also serve as a foundation for other analysis methods described later.

The likelihood is denoted by $L(\theta)$, which is defined as the probability of the data, given a model with fully specified parameter values.[2] The likelihood is a function of the parameter vector θ, which represents the vector of asymptote, rate, and intercept parameters for the two curves. Given the assumption that the data are normally distributed, the likelihood can be defined as follows:

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(d_i - \hat{d}_i\right)^2}{2\sigma^2}\right), \tag{5}$$

where d_i is the observed d' values, $\hat{d}_i$ is the predicted values, $n = 16$ is the number of data points, and $\sigma = 0.2$ is the standard error. The predicted values, $\hat{d}_i$, are defined by the shifted exponential function in Equation 1 and the parameter vector θ. The maximum likelihood estimates are denoted by $\hat{\theta}$. Computationally, the MLE procedure can be achieved by minimizing the negative likelihood or negative log-likelihood. (Minimization routines are available in common software packages, such as `fminsearch` in MATLAB, or `optim` in R.)

The MLE values are shown in Table 1 and were used to generate the time–accuracy curves in Figure 1. We report the rate estimates in units of time (where a longer time means a slower rate), so that they are on the same scale as the intercept estimates.

The obtained parameter values suggest the following conclusions. First, the cued condition produced higher discriminability (i.e., asymptote) than did the uncued condition. Second, the rate at which discriminability approached asymptote was faster in the uncued condition than in the cued condition. Finally, the minimal retrieval time (i.e., intercept) was shorter in the cued condition than in the uncued condition.

FITTING SIMPLER MODELS

In the previous section, we analyzed the full six-parameter model (2*A*–2*R*–2*I*), in which the two conditions were fit with separate asymptote (*A*), rate (*R*), and intercept (*I*) parameters. In this section, we will consider whether simpler models that are special cases of the full model can fit the data equally well. Each nested model is specified by whether the cued and uncued conditions are fitted with the same or different (asymptote, rate, and intercept) parameters. Including the full model, there were eight candidate models (1*A*–1*R*–1*I*, 2*A*–1*R*–1*I*, 1*A*–2*R*–1*I*, 1*A*–1*R*–2*I*, 2*A*–2*R*–1*I*, 2*A*–1*R*–2*I*, 1*A*–2*R*–2*I*, and 2*A*–2*R*–2*I*). The relationships among the eight models are shown in Figure 2 (see Batchelder & Riefer, 1990, p. 552). The full six-parameter model is shown at the top. The second row shows the five-parameter models. The third row shows the four-parameter models. And the null three-parameter model is shown at the bottom. The directed arrows represent nested relations between models.

For each model, we computed the adjusted-R^2 statistic (Dosher et al., 2004, p. 13):

$$\text{adj } R^2 = 1 - \frac{\sum_{i=1}^{n}(d_i - \hat{d}_i)^2/(n-k)}{\sum_{i=1}^{n}(d_i - \bar{d})^2/(n-1)}, \tag{6}$$

Table 1
Maximum Likelihood Estimates for Full Model

Condition	Asymptote (d')	Rate (sec)	Intercept (sec)
Cued	2.58	0.209	0.346
Uncued	1.94	0.152	0.443

Note—The rate estimates are reported in time units. The rate estimates, unlike the intercept estimates, suggest an advantage for the uncued condition over the cued condition. This outcome is likely to be due to *overfitting* and *parameter trade-off*. See the text for a discussion.

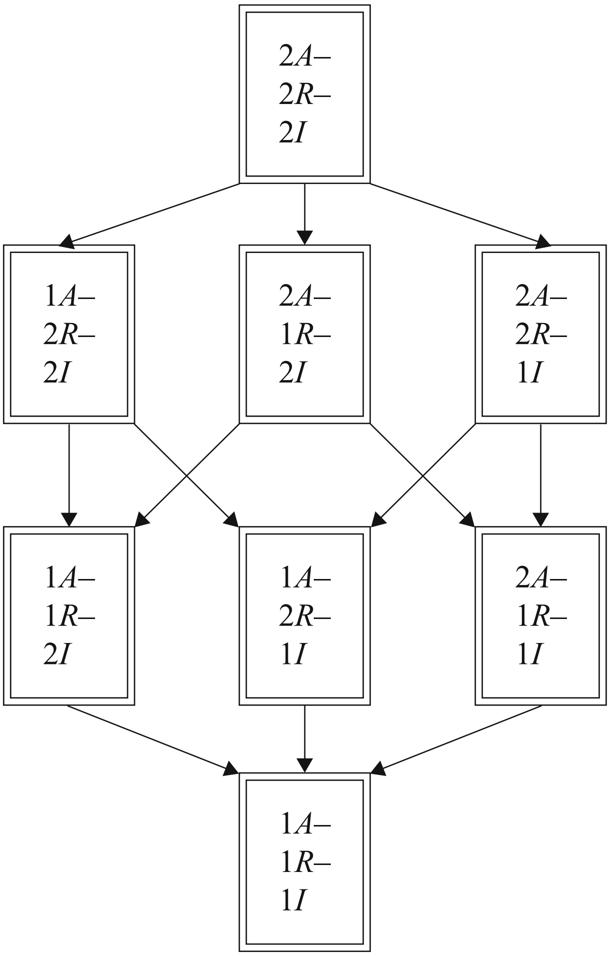

Figure 2. Nested hierarchy for the eight candidate models. Directed arrows indicated nested relations between models. *A*, asymptote; *R*, rate; *I*, intercept.

where d_i is the observed d' values, $\hat{d}_i$ is the predicted values, $\bar{d}$ is the mean, n is the number of data points, and k is the number of free parameters. These statistics are reported in Table 2.

The best-fitting model was the full six-parameter model (2*A*–2*R*–2*I*), and the worst-fitting model was the *null* three-parameter model (1*A*–1*R*–1*I*). For any pair of nested models (i.e., where one is a special case of the other), the model with more parameters yielded a better fit. The null model can be rejected because there is a substantial margin between the fit of the null model and those of the other models. The four best-fitting models were the four models (2*A*–1*R*–1*I*, 2*A*–2*R*–1*I*, 2*A*–1*R*–2*I*, and 2*A*–2*R*–2*I*) in which different asymptote parameters were allocated to the two conditions.

We can conclude from these analyses that there is likely to be a difference in asymptotes between the two conditions. It is more difficult, however, to judge whether there is also a difference in rates, a difference in intercepts, or both, because these models yielded similar values of adjusted-R^2. Although adjusted-R^2 is a useful measure of goodness of fit, the adjustment for additional parameters is a reflection of computing unbiased estimates of mean

Table 2
Model Comparison Statistics for Eight Candidate Models

Model	Adj-R^2	Deviance	AIC	BIC
1*A*–1*R*–1*I*	.900	14.0	20.0	22.3
2*A*–1*R*–1*I*	.977	−13.0	−4.97	−1.88
1*A*–2*R*–1*I*	.961	−6.73	1.27	4.36
1*A*–1*R*–2*I*	.951	−2.79	5.21	8.30
2*A*–2*R*–1*I*	.983	−15.0	−4.96	−1.10
2*A*–1*R*–2*I*	.990	−17.9	−7.93*	−4.06*
1*A*–2*R*–2*I*	.965	−7.24	2.76	6.62
2*A*–2*R*–2*I*	.992*	−18.3*	−6.31	−1.67

Note—The asterisk denotes the "best" model according to the model comparison statistic. *A*, asymptote; *R*, rate; *I*, intercept.

squared error, rather than being a principled penalty for model complexity (Anderson-Sprecher, 1994). A principled penalty for model complexity is necessary to avoid *overfitting* (Myung, 2000; Pitt & Myung, 2002). This refers to fitting of random noise that is specific to a single set of data but unrelated to the true underlying process.

One strategy to avoid overfitting is to conduct null hypothesis significance tests. However, there are at least two limitations to such a strategy (see Wagenmakers, 2007, for many more limitations). First, classical hypothesis tests are restricted to comparisons of nested models; nonnested models cannot be directly compared. Second, hypothesis tests do not provide an intuitive measure of evidence for, or against, either hypothesis. In the next section, we will use two model section criteria that can overcome these limitations.

MODEL SELECTION

To strike an appropriate balance between goodness of fit and parsimony, some researchers rely on the AIC or the BIC. Together, these criteria have been used in many domains of psychology (see Wagenmakers & Farrell, 2004, for applications of the AIC, and Wagenmakers, 2007, for applications of the BIC). The AIC and BIC for a given model can be calculated as follows:

$$\text{AIC} = -2\log L(\hat{\theta}) + 2k, \tag{7}$$

$$\text{BIC} = -2\log L(\hat{\theta}) + k\log n, \tag{8}$$

where $L(\hat{\theta})$ is the maximized likelihood, k is the number of parameters, and n is the number of observations (in our case, $n = 16$) entering the likelihood calculation. Models with lower values on these criteria are preferred to models with higher values. Note that both criteria are calculated as the sum of two terms. The first term is known as the *deviance* and is a measure of goodness of fit, where a higher deviance implies a worse fit. (The classical definition of *deviance* includes a constant term that represents the fit of a saturated model. The omission of this constant does not affect the model comparisons in this case.) The second term for each criterion is a measure of model complexity: Models with more parameters suffer a greater penalty. These criteria implement a form of *Occam's razor*; that is, the selected model should use the fewest number of free parameters to provide a good fit to the data.

The deviance values for the eight candidate models are reported in Table 2, where higher deviance values imply a worse fit. These values reveal the same ordering of models as that of adjusted-R^2, although on a different scale. The difference in deviance values between any two nested models (the G^2 statistic) can be used to conduct hypothesis tests. When the simpler model is "true," the difference in deviance values between the two nested models has an asymptotic χ^2 distribution, with degrees of freedom equal to the difference in the number of parameters. Thus, the simpler model can be rejected when the observed difference in deviance exceeds the relevant critical value. For example, the 2*A*–1*R*–2*I* model fits better than the 2*A*–1*R*–1*I* model [$G^2(1) = 4.95$, $p < .05$], but there is no difference between the 2*A*–1*R*–2*I* model and the full model [$G^2(1) = 0.38$, $p = .54$]. These hypothesis tests suggest that the 2*A*–1*R*–2*I* model is the most parsimonious model that adequately fits the data. The AIC and BIC analyses below give similar conclusions but also provide a measure of uncertainty across models.

Table 2 reports the AIC values for the same models: The AIC values are on the same scale as the deviance values but are shifted upward by an amount that is proportional to the number of parameters in each model. The addition of the complexity penalty means that the full model is no longer the best model. The model with the lowest AIC is the 2*A*–1*R*–2*I* model. This is also the model with the lowest BIC (see Table 2). We can conclude from these information criteria that the two conditions differ in asymptotes and intercepts, but not rates.

The AIC was derived under a philosophical framework different from that of the BIC. The AIC (Akaike, 1973; Bozdogan, 2000; Burnham & Anderson, 2002) for a model was derived as an estimate of the relative expected Kullback–Leibler divergence between the fitted model and the unknown true model. The BIC (Raftery, 1995; Schwarz, 1978; Wasserman, 2000) for a model is directly related to the marginal likelihood, which is computed by averaging the likelihood over the parameter space for a model. According to Wagenmakers and Farrell (2004),

> The BIC assumes that the true generation model is in the set of candidate models, and it measures the degree of belief that a certain model is the true data-generating model. The AIC does not assume that any of the candidate models is necessarily true, but rather calculates for each model the Kullback–Leibler discrepancy, which is a measure of the distance between the probability density generated by the model and reality. . . . A formal comparison in terms of performance between AIC and BIC is very difficult, particularly because AIC and BIC address different questions. (pp. 194–195)

To apply these two criteria, however, one need not accept the stringent assumptions embodied in their respective frameworks. Instead, one can adopt a more pragmatic approach that was suggested by Penny, Stephan, Mechelli, and Friston (2004). It is well known in the model selection literature (Kass & Raftery, 1995) and is evident from Equations 7 and 8 that the AIC tends to favor more complex models, whereas the BIC tends to favor simpler models, whenever the sample size, n, exceeds seven. When the same model is selected by both the AIC and the BIC (as is the case for this hypothetical data set), one may be more confident that the best model was selected.

Although it is beyond the scope of this article to discuss the relative merits of the AIC and BIC (see, e.g., Kuha, 2004), both criteria are useful alternatives to the adjusted-R^2 statistic for comparing time–accuracy curves. To our knowledge, there is no principled justification for the penalty implied by the adjusted-R^2 statistic (as there is for the AIC and BIC). Myung (2000, p. 196) made a similar comment about the *root-mean squared deviation* (RMSD):

$$\mathrm{RMSD} = \sqrt{\textstyle\sum_{i=1}^{n}(d_i - \hat{d}_i)^2 \,/\, (n-k)}, \tag{9}$$

which is a statistic equivalent to the adjusted-R^2. Both the adjusted-R^2 and the RMSD have been shown to perform poorly in model selection scenarios (Myung, 2000; Myung & Pitt, 1997; Raftery, 1995).

Model Weights

Some researchers have suggested that raw differences between models in AIC or BIC values can be transformed into *model weights* (Glover & Dixon, 2004; Wagenmakers, 2007; Wagenmakers & Farrell, 2004). The model weights, w_k, can be computed as

$$w_k = \frac{\exp\left\{-\frac{1}{2}\Delta_i(IC)\right\}}{\sum_{k=1}^{K}\exp\left\{-\frac{1}{2}\Delta_k(IC)\right\}}, \tag{10}$$

where IC denotes the value of the information criterion (either the AIC or the BIC) and $\Delta_i(IC)$ is the difference between the IC value of model i and the lowest IC value obtained for the set of K candidate models. These model weights (plotted in Figure 3) are expressed as probabilities—ranging from 0 to 1—that a particular model is the best model.

The likelihood ratio LR_{ij} (Glover & Dixon, 2004) in favor of model i over model j can be simply calculated as

$$\begin{aligned}\mathrm{LR}_{ij} &= \frac{w_i}{w_j},\\ &= \frac{\exp\left\{-\frac{1}{2}\Delta_i(IC)\right\}}{\exp\left\{-\frac{1}{2}\Delta_j(IC)\right\}}\end{aligned} \tag{11}$$

One advantage of using model weights over null hypothesis significance testing is that model weights are able to quantify the degree of evidence in favor of simpler models.[3] In this case, the 2*A*–1*R*–2*I* model is 2.2 times as likely as the 2*A*–2*R*–2*I* model according to the AIC weights and 3.3 times as likely according to the BIC weights. Furthermore, the model weights (unlike hypothesis tests) can be used to directly compare nonnested models.

The parameter estimates for the 2*A*–1*R*–2*I* model are shown in Table 3. The estimates from the 2*A*–1*R*–2*I* model are slightly different from those from the full model in Table 1. Because the information criteria favored the 2*A*–1*R*–2*I* model, which assumes the same rate for both

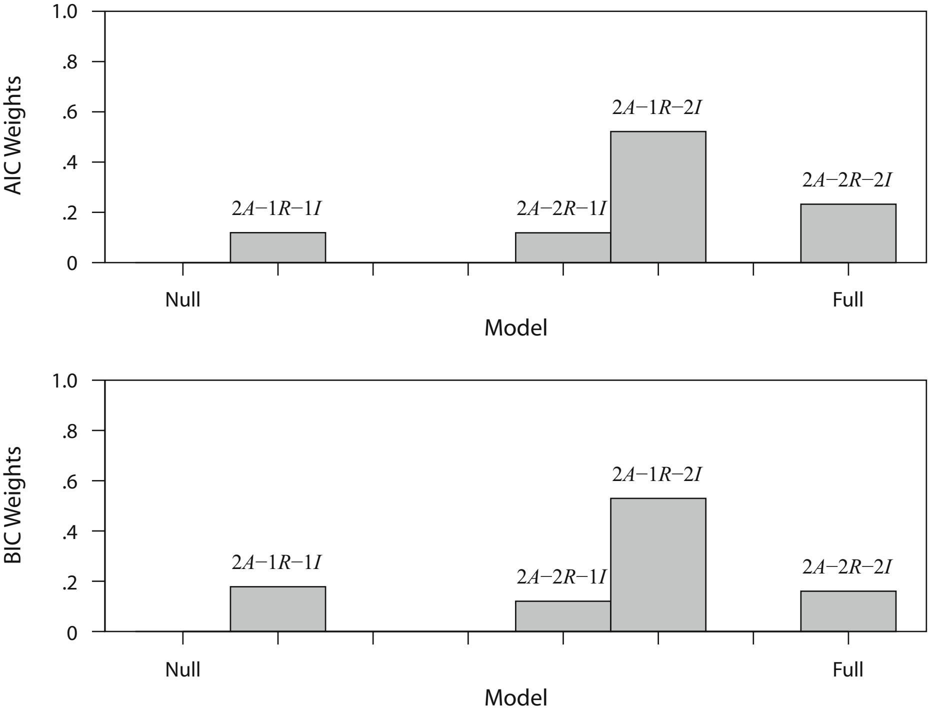

Figure 3. Akaike information criterion (AIC) weights (top panel) and Bayesian information criterion (BIC) weights (bottom panel) for eight candidate models. Weights can be interpreted as the probability that a model is the best model, according to the relevant criterion.

Table 3
Maximum Likelihood Estimates for 2*A*–1*R*–2*I* Model

Condition	Asymptote (d')	Rate (sec)	Intercept (sec)
Cued	2.55	0.189	0.354
Uncued	1.98	0.189	0.425

Note—The rate estimates are reported in time units. The model constrains the rates to be identical in the cued and uncued conditions.

conditions, the apparent rate differences in the full model may be attributed to overfitting.

Group Data

The model selection criteria could also be applied when data are collected from multiple participants (as is often the case in response signal experiments), to quantify the consistency of an effect across participants. Assuming that the data from N individuals are statistically independent, the N likelihood ratios from all individuals can be multiplied to produce a *group likelihood ratio* (GLR), as suggested by Stephan and Penny (2007):

$$\mathrm{GLR}_{ij} = \prod_{n=1}^{N} \mathrm{LR}_{ij}^{n}, \tag{12}$$

where LR_{ij}^{n} denotes the likelihood ratio for model i over model j for individual n. To complement the GLR, Stephan and Penny recommended the reporting of an *average likelihood ratio* (ALR), which is the geometric mean of the individual likelihood ratios:

$$\mathrm{ALR}_{ij} = (\mathrm{GLR}_{ij})^{1/N}. \tag{13}$$

ALR_{ij} can be interpreted on the same scale as an individual likelihood ratio: Greater than 1 implies a preference for model i over model j, and less than 1 implies a preference for model j over model i, on average, across all participants.

There are, however, a number of limitations to the strategy outlined above. First, it is assumed that all participants should be fitted by the same model. Second, the potential problem of parameter trade-off is not adequately addressed. Finally, the direction of the experimental effect is not accounted for. We will describe each of these limitations briefly below.

The strategy above involves fitting a particular model to all of the participants, so it assumes that all the participants exhibited the same pattern of experimental effects. Although this is a reasonable initial assumption, such an analysis can be sensitive to outliers. Consider the scenario in which we have 11 participants, where the data from 10 individuals are best fit (according to the AIC or BIC) by a 2*A*–2*R*–1*I* model, whereas the data from a single individual are best fit by a 2*A*–2*R*–2*I* model. It is possible that both the GLR and the ALR will favor the 2*A*–2*R*–2*I* model, because the likelihood ratio for the single individual overwhelms that of the other 10 individuals. The more reasonable conclusion (that the outlier shows a pattern qualitatively different from the others) can be realized by relaxing the assumption that all participants should be fitted by the same model.

In cases in which a few individuals exhibit patterns very different from those for all the other individuals, one can still use the information criteria to select the optimal partitioning. Some participants may exhibit an experimental effect (and, therefore, require more parameters), whereas other participants may show no experimental effect (and, therefore, require fewer parameters). The entire sample of participants may then be split into qualitatively different groups. The likelihood ratios can be combined across participants in the same way as was described above, but without the constraint that all the participants should be fitted by the same model. For example, Lee and Webb (2005) used the BIC in a similar fashion to identify the number of clusters of participants who showed qualitatively different responses.

Another limitation of these information criteria is that they do not directly address possible parameter trade-offs. In particular, it can often be difficult to statistically distinguish between differences in rate from differences in intercept. To overcome this difficulty, a new variable can be computed by summing the rate and intercept parameters in time units (see, e.g., Carrasco, Giordano, & McElree, 2004). This yields a composite measure of processing "speed," $\delta + \beta$, which indexes how quickly the curve rises from stimulus onset.

A third limitation is that the information criteria do not account for the direction of experimental effects. For example, suppose that all the participants who performed a spatial cuing task are best fitted by the 2*A*–1*R*–1*I* model that requires two asymptote parameters. Despite this apparent consistency, it is possible that although most of the participants exhibit a cuing advantage (i.e., higher asymptote in the cued condition than in the uncued condition), some participants may exhibit a cuing disadvantage (i.e., higher asymptote in the uncued condition than in the cued condition).[4] Given such a scenario, one could use a *t* test to determine whether the mean difference in asymptotes across all participants is different from zero (McElree, 1998), where the dependent measure is the difference in asymptote estimates from each participant.

Although this *t* test can be used to examine a significant effect at the group level, potential significant effects at the individual level may also be informative. In addition, point parameter estimates can be misleading when there is no measure of uncertainty associated with those estimates. These issues will be addressed in the next section.

ACCOUNTING FOR PARAMETRIC UNCERTAINTY

There are two common approaches to testing nested hypotheses (Kass, 1993). In the first approach (known as *model selection*), a set of constrained models is formulated, and the goal is to find the most parsimonious model that adequately fits the data. Using this approach, the 2*A*–1*R*–2*I* model was selected. In the second approach (known as *estimation*), only the full model is assumed to be true. Nested hypotheses can be tested by examining the difference between the relevant parameter estimates, with some measure of uncertainty that is expressed as an interval. Very small differences in parameter estimates could be regarded as "nonsignificant." Inferences based on interval estimates avoid overfitting (as compared with those based on point estimates), because the intervals express a range of possible values.

One argument in favor of estimation over model selection is that in some cases, the simpler models are not plausible. For example, consider the 2*A*–1*R*–2*I* model, which assumes that the two rates are identical; in other words, the difference in rates is *precisely* zero. In contrast, the full model assumes that the difference in rates is not precisely zero but may be so small that it is *practically* zero. The plausibility of this assumption will, of course, vary from one domain to the next. For example, one of the strengths of process models, such as the diffusion model, is that they assume *parameter invariance* (Ratcliff, 2006): Parameters that represent psychological variables should not be free to vary across conditions, unless it is meaningful to do so.[5] In contrast, parameter invariance need not hold for descriptive models, such as the shifted exponential function. Instead, it may be more reasonable to assume that different conditions could potentially have any effect on the model parameters, even if very small. To demonstrate the estimation approach to hypothesis testing, we will fit the full model (2*A*–2*R*–2*I*) only.

Interval Estimates: Bootstrap

We will return to fitting the full model but will include a measure of uncertainty for the parameter estimates that were reported in Table 1. There are many methods available for calculating interval estimates for parameters (Huber, 2006; Verguts & Storms, 2004). In this section, we will use an intuitive and flexible method known as the *parametric bootstrap*. Wichmann and Hill (2001) showed, in greater detail, how the parametric bootstrap can be used to analyze psychometric functions. Carrasco, Giordano, and McElree (2006) have recently applied the bootstrap to analyze time–accuracy curves.

The basic idea behind the parametric bootstrap is to produce a distribution of parameter estimates by refitting the model to a large number of replicate data sets sampled from the original MLEs (see Wagenmakers et al., 2004, for a related bootstrap procedure). We first took the original MLEs of the full model (shown in Table 1) and generated 1,000 replicate data sets from these parameter values. These replicate data sets were produced by adding sampling error to the predicted time–accuracy points. (In this demonstration, the standard error was 0.2, the known standard deviation in our simulated data set. In practice, when the standard deviation is unknown, an estimate based on Equation 3 would be used.) As in the original data set, each replicate data set comprised 8 points on each time–accuracy curve. We then treated each replicate data set as if it were observed in some hypothetical study and found the corresponding MLE. The 1,000 resampled estimates of the parameter vector could then be used to express the uncertainty about the original parameter estimates. For example, a 95% central confidence interval is obtained for each parameter by reporting the 2.5 and 97.5 percentiles of the bootstrap estimates. (This is known as the *percentile method* for calculating confidence intervals; more sophisticated methods have been presented in Efron & Tibshirani, 1993, and Davison & Hinkley, 1997.) These interval estimates are shown in Table 4. The intervals indicate that there is considerable uncertainty associated with the original parameter estimates, especially those of the rates.

Table 4
Bootstrap Interval Estimates

Condition	Asymptote (d')	Rate (sec)	Intercept (sec)
Cued, 97.5%	2.86	0.390	0.400
Cued, 2.5%	2.21	0.097	0.258
Uncued, 97.5%	2.41	0.489	0.540
Uncued, 2.5%	1.70	0.048	0.328

Note—The 95% confidence intervals are represented by the 2.5 and 97.5 percentiles of the bootstrap estimates.

To compare the cued and uncued conditions, we computed a 95% confidence interval for the difference in parameter estimates between the two conditions. For each of the 1,000 bootstrap samples, the asymptote, rate, and intercept estimates in the cued condition were subtracted from the corresponding estimates in the uncued condition. This procedure produced a bootstrap distribution for the difference between the two conditions on each of the three parameters (see Figure 4).

The results in Figure 4 can be interpreted as follows. The asymptote in the cued condition was higher than that in the uncued condition, and the size of this accuracy advantage could be anywhere from very small up to quite large (up to 1 d'). The difference in rate estimates was close to zero. Given that the interval is relatively wide, no strong conclusions can be made about the difference in rates. The intercept in the cued condition was shorter than that in the uncued condition, and this advantage could be anywhere from small to large (up to 220 msec). The meaning attached to the magnitude of such differences must be interpreted within the particular theoretical context: What may be considered a small effect in one domain may be considered a large effect in another domain. Finally, the inferences made in this section are similar to those made in the previous section, where the 2*A*–1*R*–2*I* model was considered the best model. Given that the interval for the rate difference was

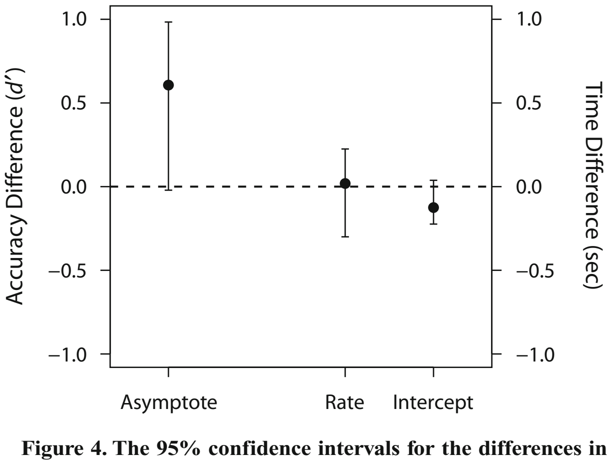

Figure 4. The 95% confidence intervals for the differences in parameters between cued and uncued conditions. The original differences in maximum likelihood estimates are represented by points.

relatively wide, however, it may have been premature to assume that the two rates were precisely identical.

The procedure that we have demonstrated here is known as the parametric bootstrap. An alternative is the *nonparametric* bootstrap, in which replicate data sets are generated by resampling, with replacement, from the original set of correct and incorrect responses (e.g., hits and false alarms). As an illustration of how parametric and nonparametric analyses might diverge, consider the following example adapted from Wichmann and Hill (2001). Assume that 2 observers perform the same response signal task. Although the hit rates and false alarm rates across lags are slightly different for the 2 observers, the maximum likelihood fits to the two data sets happen to yield identical parameter estimates. In this situation, the parametric bootstrap would produce identical estimates of parametric uncertainty for both observers. In contrast, the nonparametric bootstrap would not produce identical estimates for both observers, given the individual differences in the raw data. According to Wichmann and Hill, "the choice of parametric versus nonparametric analysis concerns how much confidence one has in one's hypothesis about the underlying mechanism [parametric] that gave rise to the raw data, as against the confidence one has in the raw data's precise numerical values [nonparametric]" (p. 1316). Although we have demonstrated only the parametric approach here, it should be noted that the nonparametric bootstrap could also be used to produce interval estimates for parameters, and the latter has the advantage of accounting for individual differences at the level of the raw data.

Interval Estimates: Bayesian

All of the methods discussed so far have relied on maximizing likelihood. Likelihood forms an essential component of statistical inference because it expresses, in an objective manner, the information that the data provide about the parameters. Nonetheless, there is often important subjective information about the parameters even before the data are observed. If this prior information is ignored, one may end up with unreasonable parameter estimates.[6] We first will offer a simple example to demonstrate how parameter estimates can be misleading when they are based on likelihood alone.

Consider a participant who has completed a block of 10 trials in a binary choice task. For simplicity, assume that these responses are independent and identically distributed. We wish to estimate θ_{true}, the participant's true rate of responding correctly, given the data. Suppose that the participant responds correctly on all 10 trials. The binomial likelihood function $L(\theta)$ for these data is

$$L(\theta) = \binom{n}{r} \theta^{r} (1-\theta)^{n-r}, \tag{14}$$

where $n = r = 10$. This function (plotted in Figure 5A) gives the probability of observing 10 correct responses out of 10 trials, for different values of θ. The MLE is the sample proportion ($10/10 = 1$) and is marked by the cross in Figure 5A; of all the possible values of the true rate, $\theta = 1$ is the most likely value, given these data.

This estimate implies either that the task can be performed with perfect accuracy or that the true error rate is so low that it cannot be detected with a small sample size. Most researchers would probably make the latter attribution and would consider the MLE to be implausible. Moreover, the likelihood function itself cannot be used to make simple inferences, such as "what is the probability that the true rate is greater than chance (.5)?" These limitations can be overcome using Bayesian methods (see Lee & Wagenmakers, 2005).

Bayesian analysis begins with a prior distribution, denoted here by $\pi(\theta)$, which represents the uncertainty about the true rate before the data are observed. Two different prior distributions were considered and are plotted as dashed lines in Figure 5B. The straight horizontal line is the uniform distribution, which is a standard *noninformative* prior for rates. This distribution expresses the prior belief that all possible values of the true rate (between 0 and 1) are equally likely. The curved line is an informative prior distribution that expresses the prior belief that the true rate cannot be 0 or 1 but lies somewhere in between. Although this prior is informative, it nevertheless represents a high degree of prior uncertainty about the true rate. Bayesian parameter estimation then combines the prior distribution $\pi(\theta)$ with the likelihood $L(\theta)$ to produce a posterior distribution:

$$\pi(\theta \mid D) \propto \pi(\theta) L(\theta), \tag{15}$$

where D denotes the data. The posterior distribution represents the uncertainty about the true rate after the data have been observed (see the Appendix).

The two resulting posterior distributions look very different from each other. The posterior arising from the uniform prior (Figure 5C) is simply proportional to the likelihood. This fact can be easily deduced from Equation 15: $\pi(\theta)$ is a constant for the uniform prior. Thus, the posterior mode in Figure 5C is identical to the MLE of 1. In contrast, the posterior mode arising from the informative prior (Figure 5D) is slightly biased away from 1, which appears to be a more plausible estimate than is the MLE.

When reporting point estimates from a Bayesian analysis, one is not restricted to the posterior modes. For example, the posterior means are also valid point estimates, and here they offer a very intuitive interpretation (see the Appendix). The posterior mean from the uniform prior (plotted as a circle in Figure 5C) is calculated as $(10 + 1)/(10 + 2) = 11/12$. The posterior mean from the informative prior (plotted as a circle in Figure 5D) is calculated as $(10 + 2)/(10 + 4) = 12/14$. Compared with the sample proportion (10/10), these posterior means can be interpreted as observed proportions that are calculated with additional prior observations. For example, the posterior mean from the uniform prior is equivalent to having observed one correct response out of two trials, before the data were collected. These calculations are very similar to those used to adjust perfect scores to avoid infinite d' estimates in signal detection experiments (Snodgrass & Corwin, 1988). The Bayesian framework offers a meaningful, alternative justification for these adjustments.

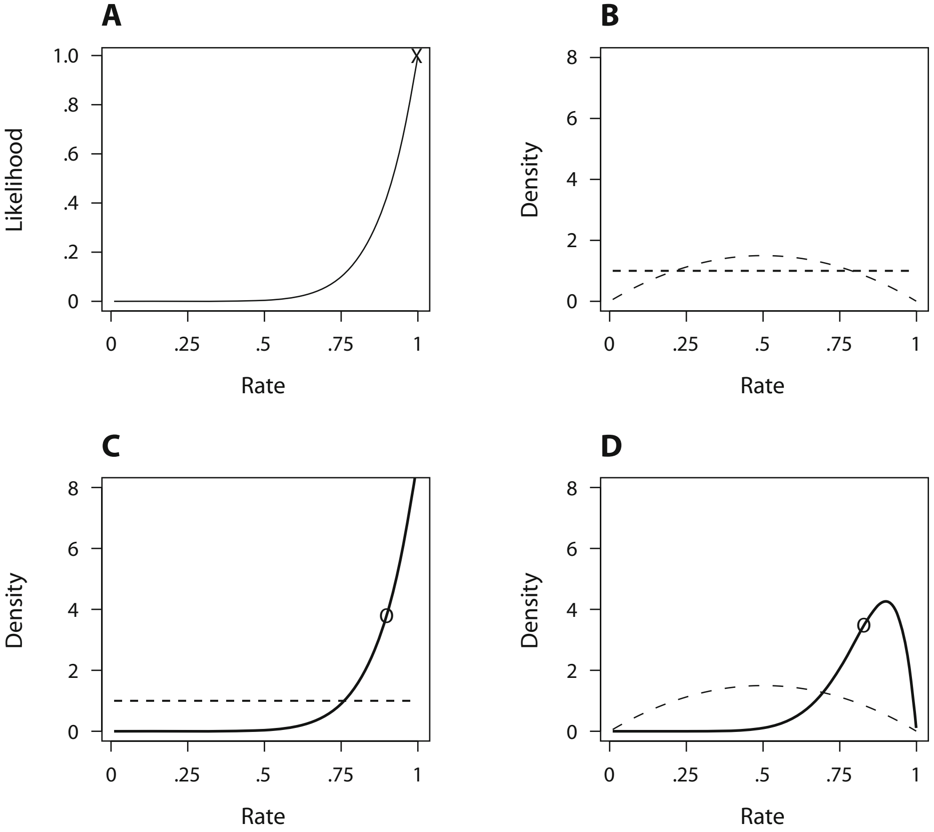

Figure 5. Estimation of the true proportion of correct responses. (A) Likelihood function. The cross marks the maximum likelihood estimate. (B) Prior distributions. The straight line is the uniform prior, and the curved line is the informative prior. (C) Posterior distribution (solid line) given the uniform prior (dashed line). The circle marks the posterior mean. (D) Posterior distribution (solid line) given the informative prior (dashed line). The circle marks the posterior mean.

Although we have reported only point estimates above, the entire posterior distributions can be used for inferences about the true rate. For example, the probability that the true rate is greater than .5 is over 99% for both posterior distributions. With respect to the hypothetical response signal data, we will report 95% (Bayesian) confidence intervals for the differences in asymptote, rate, and intercept parameters.

When the posterior density cannot be derived analytically, simulation methods are available to approximate the posterior. A significant proportion of modern Bayesian data analysis relies on a set of simulation methods known as Markov chain Monte Carlo, or MCMC. The purpose of these algorithms is to explore the parameter space of a model in such a way that parameter values are sampled in direct proportion to the posterior probability (or density) of those parameter values. As with all Monte Carlo methods, the precision of the estimates can be improved by increasing the number of Monte Carlo samples. Below, we will report estimates that are based on 10,000 samples. A detailed discussion of these methods is beyond the scope of this article; introductions for experimental psychologists are presented in Kuss, Jäkel, and Wichmann (2005), Lee (2008), and Rouder and Lu (2005). Here, we used the tools available in free software packages: WinBUGS 1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000) and the Bayesian Output Analysis package (B. J. Smith, 2005).

We first assume an independent uniform prior, with subjectively determined upper and lower bounds, for each of the six parameters in our model. The priors for the two asymptote parameters are uniform between 0 and 4 d', the priors for the two rate parameters are uniform between 10 and 400 msec, and the priors for the two intercept parameters are uniform between 200 and 600 msec. These upper and lower bounds are chosen to avoid implausible parameter estimates.

On the basis of background knowledge, experimenters will often have an intuitive sense of the range of parameter values that they would regard as reasonable. The Bayesian approach allows experimenters to quantify this intuition. By specifying the range of the prior parameter space, any samples from the posterior that lie outside this range will, effectively, be ignored. Although the specification of Bayesian priors requires subjective judgment, similar decisions are commonly made in other data analysis contexts (see Rouder & Lu, 2005, for a discussion). For example, researchers often exclude from further analyses any responses, or participants, that are considered to be

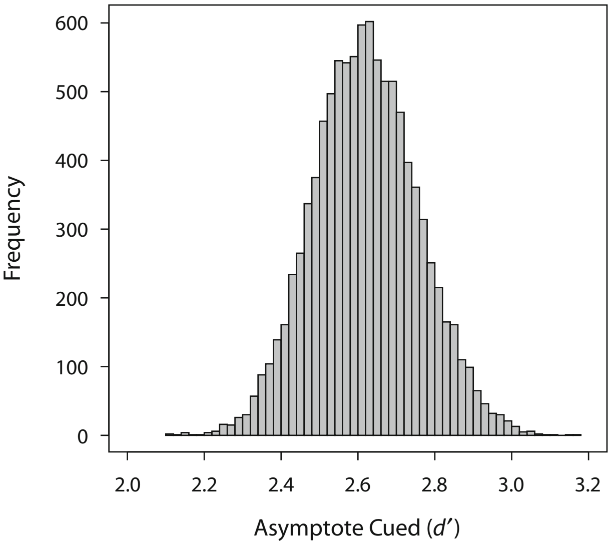

Figure 6. Histogram of 10,000 samples of the asymptote parameter in the cued condition.

outliers. The criteria used to judge the presence of outliers are clearly subjective. In fact, such criteria may be reinterpreted within the Bayesian framework as representing prior knowledge about the expected distribution of the data.[7]

Figure 6 shows a histogram of 10,000 samples for the asymptote parameter in the cued condition. Point and interval estimates can be based on these 10,000 samples for any of the parameters or for any function of the parameters. For example, if we subtract the asymptote of the cued condition from that of the uncued condition at each iteration, we obtain a posterior distribution of the difference in asymptotes.

The 95% Bayesian confidence intervals for the difference in asymptote, rate, and intercept are shown in Figure 7. The posterior median for the difference between the cued and the uncued conditions on each parameter is plotted as a point. These Bayesian intervals offer at least two advantages, when compared with the bootstrap intervals reported previously. The first advantage is that the Bayesian intervals are more precise than the bootstrap intervals, because prior information was incorporated into the former. This increase in precision was possible even with rather vague prior distributions. The bootstrap intervals were calculated from a distribution of MLEs, and many of these estimates may have been unreasonable. The second advantage is that the Bayesian intervals are easier to interpret than the bootstrap intervals. Recall that the bootstrap distribution is a hypothetical sampling distribution of MLEs. The probabilities associated with bootstrap confidence intervals refer to *coverage* probabilities—that is, how often the confidence intervals will cover the true parameter values in repeated sampling. For the nominal coverage probability to be valid, one is forced to "accept" all parameter values within an interval and "reject" those outside the interval. Probabilities associated with Bayesian confidence intervals do not refer to coverage probabilities but to degrees of belief. One can infer, for example, that there is a 50% probability that the true parameter value lies above (or equivalently, below) the posterior median. Intuitive probabilistic inferences such as these are possible within the Bayesian framework only.

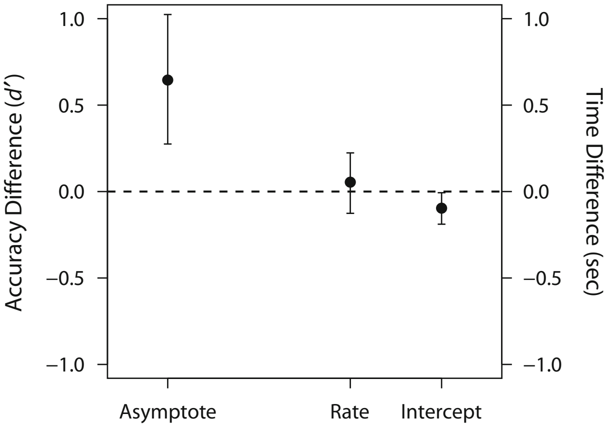

Figure 7. The 95% Bayesian confidence intervals for the difference in parameters between cued and uncued conditions. The posterior medians are represented by points.

CONCLUSIONS

In this article, we have raised a number of important issues that must be addressed when researchers wish to compare time–accuracy curves. Although time–accuracy curves were the primary focus here, similar issues arise in other domains as well. Some of these domains include forgetting or retention curves (Rubin & Wenzel, 1996), learning curves (Heathcote, Brown, & Mewhort, 2000), psychometric functions (Kuss et al., 2005), and multinomial processing models (Batchelder & Riefer, 1990).

One of the key issues is whether to use model selection or parameter estimation for hypothesis testing. The choice between these two approaches is not trivial and depends on the ultimate purpose of the analysis (see Chechile, 1977, p. 183). On the one hand, if a researcher is interested specifically in how two curves might differ in any way, estimation of the full model may be preferred. In this case, interval estimates of the differences between parameters can protect against overfitting. On the other hand, if a researcher wishes to identify the most parsimonious model that adequately describes the data, model selection may be preferred. In this case, researchers should use a principled information criterion that balances goodness of fit and complexity. A strong assumption of the model selection approach is that some parameter values may be constrained to be identical. In the cognitive modeling literature, this principle is known as parameter invariance (Ratcliff, 2006; see also Busemeyer & Wang, 2000). As was suggested earlier, whether such assumptions are justified will depend on the researcher's goals.

For model selection purposes, the criteria described here (AIC and BIC) are not the only criteria available, although they are the most commonly used. One criticism that has been raised against both AIC and BIC is that they are sensitive to only one aspect of model complexity (the number of parameters) but insensitive to functional form. For example, Pitt, Myung, and Zhang (2002) gave several illustrations of models that differ in complexity (which can be regarded, loosely, as how well a model fits random data), even though they have the same number of parameters. Other criteria that account for functional form complexity include cross-validation and minimum description length methods (see Myung, 2000; Pitt & Myung, 2002).

For parameter estimation, both bootstrap and Bayesian methods have their relative advantages and disadvantages. For example, bootstrap methods are often easier to implement than Bayesian MCMC methods (see Wagenmakers et al., 2004, for a discussion). Nonetheless, Bayesian methods offer an attractive alternative approach to statistical inference. For example, the Bayesian approach extends naturally to hierarchical or multilevel designs, in which time–accuracy curves are obtained from multiple participants. In experimental psychology, hierarchical Bayesian modeling has been successfully applied to signal detection models (Rouder & Lu, 2005), RT distributions (Rouder, Lu, Speckman, Sun, & Jiang, 2005), and many other domains (Lee, 2008). Future comparisons of time–accuracy curves could also benefit from the application of hierarchical Bayesian methods.

AUTHOR NOTE

This research was supported by a Melbourne Research Scholarship to C.C.L. and an Australian Research Council grant to P.L.S. Preparation of the manuscript was supported by a MUARC publication award. We thank Michael Lee, Brian McElree, David Huber, and Adam Morris for many helpful comments. Correspondence concerning this article should be addressed to C. C. Liu, Accident Research Centre, Monash University, Building 70, Melbourne, VIC 3800, Australia (e-mail: charles.liu@muarc.monash.edu.au).

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Anderson-Sprecher, R. (1994). Model comparisons and R^2. *American Statistician*, **48**, 113-117.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, **97**, 548-564.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, **44**, 62-91.

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, **44**, 171-189.

Carrasco, M., Giordano, A. M., & McElree, B. (2004). Temporal performance fields: Visual and attentional factors. *Vision Research*, **44**, 1351-1365.

Carrasco, M., Giordano, A. M., & McElree, B. (2006). Attention speeds processing across eccentricity: Feature and conjunctive searches. *Vision Research*, **46**, 2028-2040.

Carrasco, M., & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences*, **98**, 5363-5367.

Chechile, R. A. (1977). Likelihood and posterior identification: Implications for mathematical psychology. *British Journal of Mathematical & Statistical Psychology*, **30**, 177-184.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

Dosher, B. A. (1979). Empirical approaches to information processing: Speed–accuracy trade-off functions or reaction time. *Acta Psychologica*, **43**, 347-359.

Dosher, B. A. (1981). The effect of delay and interference: A speed–accuracy study. *Cognitive Psychology*, **13**, 551-582.

Dosher, B. A. (1982). Sentence size, network distance and sentence retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 173-207.

Dosher, B. A. (1984). Degree of learning and retrieval speed: Study time and multiple exposures. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 541-574.

Dosher, B. A., Han, S. M., & Lu, Z. L. (2004). Parallel processing in visual search asymmetry. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 3-27.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, **12**, 403-408.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman & Hall/CRC.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791-806.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, **32**, 25-33.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 846-858.

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

Huber, D. E. (2006). Computer simulations of the ROUSE model: An analytic simulation technique and a comparison between the error variance–covariance and bootstrap methods for estimating parameter confidence. *Behavior Research Methods*, **38**, 557-568.

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society*, **42**, 551-560.

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

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, **33**, 188-229.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, **5**, 478-492.

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, **48**, 310-321.

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

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, **112**, 662-668.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605-621.

Link, W. A. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Erlbaum.

Link, W. A., & Barker, R. J. (2006). Model weights and the foundations of multimodel inference. *Ecology*, **87**, 2626-2635.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, **52**, 362-375.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics & Computing*, **10**, 325-337.

McElree, B. (1996). Accessing short-term memory with semantic and phonological information: A time-course analysis. *Memory & Cognition*, **24**, 173-187.

McElree, B. (1998). Attended and nonattended states in working memory: Accessing categorized structures. *Journal of Memory & Language*, **38**, 225-252.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 817-835.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, **118**, 346-373.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General*, **122**, 291-315.

McElree, B., & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 134-157.

Miller, J. (1996). The sampling distribution of *d′*. *Perception & Psychophysics*, **58**, 65-72.

Mulligan, N., & Hirshman, E. (1995). Speed–accuracy trade-offs and the dual process model of recognition memory. *Journal of Memory & Language*, **34**, 1-18.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.

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

Myung, I. J., Kim, C., & Pitt, M. A. (2000). 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.

Nikolic, D., & Gronlund, S. D. (2002). A tandem random walk model of the SAT paradigm: Response times and accumulation of evidence. *British Journal of Mathematical & Statistical Psychology*, **55**, 263-288.

Öztekin, I., & McElree, B. (2007). Proactive interference slows recognition by eliminating fast assessments of familiarity. *Journal of Memory & Language*, **57**, 126-149.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, **22**, 1157-1172.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, **6**, 421-425.

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, **25**, 111-163.

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

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, **53**, 195-237.

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

Reed, A. V. (1973). Speed–accuracy trade-off in recognition memory. *Science*, **181**, 574-576.

Reed, A. V. (1976). List length and the time course of recognition in human memory. *Memory & Cognition*, **4**, 16-30.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, **12**, 573-604.

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

Rubin, D. B., & Wenzel, A. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, **103**, 734-760.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Smith, B. J. (2005). *Bayesian Output Analysis Program (BOA), Version 1.1.5* [Online]. Iowa City: University of Iowa. Available at www.public-health.uiowa.edu/boa.

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

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research*, **44**, 1297-1320.

Smith, P. L., & Wolfgang, B. J. (2004). The attentional dynamics of masked detection. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 119-136.

Smith, P. L., Wolfgang, B. J., & Sinclair, A. J. (2004). Mask-dependent attentional cuing effects in visual signal detection: The psychometric function for contrast. *Perception & Psychophysics*, **66**, 1056-1075.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.

Stephan, K. E., & Penny, W. D. (2007). Dynamic causal models and Bayesian selection. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (pp. 577-585). Amsterdam: Elsevier.

Thomas, R. D. (2006). Processing time predictions of current models of perception in the classic additive factors paradigm. *Journal of Mathematical Psychology*, **50**, 441-455.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550-592.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, **40**, 61-72.

Verguts, T., & Storms, G. (2004). Assessing the informational value of parameter estimates in cognitive models. *Behavior Research Methods, Instruments, & Computers*, **36**, 1-10.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, **39**, 767-775.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, **14**, 779-804.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, **11**, 192-196.

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

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92-107.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, **27**, 359-397.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, **63**, 1314-1329.

Wickelgren, W. A. (1977). Speed–accuracy trade-off and information processing dynamics. *Acta Psychologica*, **41**, 67-85.

Wickelgren, W. A., & Corbett, A. T. (1977). Associate interference and retrieval dynamics in yes–no recall and recognition. *Journal of Experimental Psychology: Human Learning & Memory*, **3**, 189-202.

NOTES

[1] Application of sequential sampling modeling may increase with the aid of recently developed software packages, such as fast-dm (Voss & Voss, 2007) and DMAT (Vandekerckhove & Tuerlinckx, 2008).

[2] For a continuous sampling distribution, the probability of the data is zero. However, given that all observations are measured to finite preci-

sion (and as long as the measurement precision is high), the probability of the data can be approximated by the probability density.

[3] For an insightful discussion of model weights and their interpretation, see Link and Barker's (2006) comments on the AIC and Weakliem's (1999) comments on the BIC.

[4] These reversals have been observed in the spatial cuing task under certain conditions (see P. L. Smith, Ratcliff, & Wolfgang, 2004; P. L. Smith & Wolfgang, 2004; P. L. Smith, Wolfgang, & Sinclair, 2004).

[5] Parameter invariance is a feature of a psychological theory. A related concept is that of *parameter selective influence* (Thomas, 2006), which refers to an experimental manipulation whereby the effect of the manipulation is selective to a single parameter or subset of parameters within a model.

[6] Alternatively, prior specification can also be justified from an *objective Bayesian* viewpoint (see Lee & Wagenmakers, 2005).

[7] Concerns about the subjectivity of Bayesian priors are legitimate. If there is controversy about the appropriate range of the true parameter values, researchers should rerun their analyses, using a range of reasonable priors to ensure that the resulting inferences are robust (Liu & Aitkin, 2008).

APPENDIX

In Bayesian analyses, a prior distribution is said to be *conjugate* to the likelihood when the resulting posterior distribution is in the same family as the prior distribution. The conjugate prior for a binomial likelihood is the beta(a_0, b_0) distribution, where a_0 and b_0 are prior parameters. The beta distribution is given by:

$$\text{beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \text{ where } \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

The uniform prior distribution is beta(1, 1), and the informative prior distribution is beta(2, 2). Given any beta prior distribution, Gill (2002, p. 68) showed that the posterior distribution is beta($a_0 + r$, $b_0 + n - r$). Thus, the corresponding posterior distributions for the uniform and the informative priors are the beta(11, 1) and beta(12, 2) distributions, respectively. The mean of the beta distribution is

$$\frac{a_0 + r}{a_0 + b_0 + n},$$

which is used to calculate the posterior means.

(Manuscript received December 10, 2007;
revision accepted for publication May 29, 2008.)