Behavior Research Methods
2006, 38 (4), 557-568

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

DAVID E. HUBER
University of Maryland, College Park, Maryland

This article provides important mathematical descriptions and computer algorithms in relation to the responding optimally with unknown sources of evidence (ROUSE) model of Huber, Shiffrin, Lyle, and Ruys (2001), which has been applied to short-term priming phenomena. In the first section, techniques for obtaining parameter confidence intervals and parameter correlations are described, which are generally applicable to any mathematical model. In the second section, a technique for producing analytic ROUSE predictions is described. Huber et al. (2001) averaged many stochastic trials to obtain stable behavior. By appropriately weighting all possible combinations of feature states, an alternative analytic version is developed, yielding asymptotic model behavior with fewer computations. The third section ties together these separate techniques, obtaining parameter confidence and correlations for the analytic version of the ROUSE model. In doing so, previously unreported behaviors of the model are revealed. In particular, complications due to local minima are discussed, in terms of both variance–covariance analyses and bootstrap sampling analyses.

The mathematical techniques described in this article provide important supporting material for applications of the responding optimally with unknown sources of evidence (ROUSE) model of short-term priming (Huber & Cousineau, 2003; Huber & O'Reilly, 2003; Huber, Shiffrin, Lyle, & Quach, 2002; Weidemann, Huber, & Shiffrin, 2005). These descriptions consist of (1) techniques for producing parameter confidence intervals and correlations between parameters in order to better understand model behavior and (2) a new analytic method for obtaining asymptotic behavior for the ROUSE model, which stands in contrast to previously used methods of stochastic sampling (Huber, Shiffrin, Lyle, & Ruys, 2001). This order of presentation is chosen because the parameter confidence techniques are generally applicable to any mathematical model and are presented first for a wider audience. The analytic method for implementing ROUSE has been utilized in several articles, but this is the first detailed description of the technique. The techniques for producing parameter confidence intervals and parameter correlations are not new but are not typically described in the psychological literature (although, for recent discussions of these techniques, see Verguts & Storms, 2004; Visser, Raijmakers, & Molenaar, 2000).

These two sections are linked in that both of them describe important computer techniques in support of a growing number of publications in which the ROUSE model has been utilized (e.g., Huber & O'Reilly, 2003; Huber, Shiffrin, Lyle, & Quach, 2002; Huber et al., 2001; Huber, Shiffrin, Quach, & Lyle, 2002; Ratcliff & McKoon, 2001; Wagenmakers et al., 2003; Weidemann et al., 2005). Providing some degree of integration across the reported techniques, the third section of the article reports the specific application of the parameter confidence techniques to the analytic implementation of the ROUSE model. These results illuminate previously unreported aspects of the model's behavior and identify complications that can arise when a nonlinear model containing local minima is used. The third section provides a direct comparison between a bootstrap analysis and use of the variance–covariance matrix, finding qualitatively similar results for each method, but with slightly different quantitative results. Both of these techniques are estimates, and each has particular complications and limitations that may give rise to these quantitative differences.

This research was supported by NIMH Grants MH063993-04 and MH63993-01, NSF Grant KDI/LIS IBN-9873492, and ONR Grant N00014-00-1-0246. I thank Lewis O. Harvey, Jr., Roger Ratcliff, Lael Schooler, Michael Wenger, and Eric-Jan Wagenmakers for many helpful tips in developing techniques for assessing parameter confidence. Correspondence concerning this article should be addressed to D. E. Huber, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109 (e-mail: dhuber@ucsd.edu).

Parameter Confidence Intervals and Parameter Correlations

When computational models are fit to behavioral data, a range of parameter values often produce fits nearly as

good as the best fit. Under the assumption that the model is the actual generator of the data (i.e., the model is correct), the best-fitting parameters should be thought of as an estimate of the "true" parameters, because the observed data are a sample of the true population. The techniques described below can be used to find the parameter confidence region within which the true parameters lie. Although these techniques have been described in detail elsewhere (Efron & Tibshirani, 1993; Press, 1992), they are rarely employed in the course of data fitting with process models (although see Ratcliff & Murdock, 1976; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). In contrast, these techniques are often employed in relation to more descriptive models, such as *structural equation modeling* (e.g., Gonzalez & Griffin, 2001).

A confidence region is produced by assuming that measurement error (e.g., limited sampling) causes the best-fitting parameters to deviate from their true values. This analysis does not speak to the goodness of one model versus another, otherwise known as model selection. Model selection is beyond the present application, and the interested reader is referred to Pitt, Myung, and Zhang (2002) for a recent discussion of mathematical techniques for model selection.

For the following discussion, refer to Figure 1, which shows a hypothetical *data space* (area defined by all the possible combinations of observed behavior for each condition) that is linked through sampling noise to a corresponding *parameter space* (area defined by all the possible combinations of model parameters). The parameter confidence region is a region in parameter space (the ellipse in the right-hand panel of the figure) in which the true parameters exist with the specified level of confidence. The true parameters correspond to some hypothetically true data set, which is not necessarily observed, due to measurement error. Therefore, the parameters that best fit the observed data are said to be estimates of the true parameters. Because changes in the observed data correspond to changes in the parameter estimates, the confidence region is found by determining the extent to which parameters vary in response to statistically expected variations in the data.

There is typically some residual error for a best fit to a particular data set, termed χ^2_e when a chi-square error measure of accuracy is used. Chi-square is used because it is a maximum likelihood estimator (i.e., it maximizes the probability of the data, given the model). The confidence region is defined as deviations in the data space that are significantly worse than this best fit, as determined by the criterion $\chi^2_c = \chi^2_e + \chi^2_\Delta$, where χ^2_Δ defines the chosen level of confidence. In other words, traditional confidence limits of observed data can be transformed into confidence limits of model parameters. However, the translation from a data confidence region to a parameter confidence region

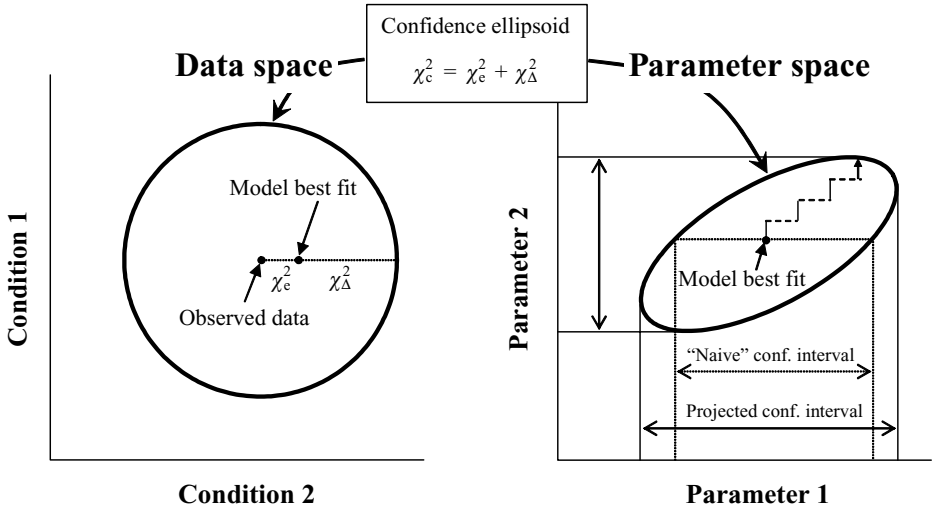

Figure 1. A confidence region ellipsoid shown both in data space and parameter space. For easy 2-D viewing, the hypothetical data consist of two conditions, and the hypothetical model consists of two parameters. The point in the middle is the result of the best-fitting parameters, where chi-squared error is minimized. Moving away from this minimum by a specified increase in chi-squared error defines the confidence region, with the magnitude of the specified increase determining the degree of confidence. In this example, the two parameters are positively correlated, so that increased error due to increases in one parameter is partially offset by increases in the other parameter. A "naive" confidence interval, obtained by varying one parameter in isolation, underestimates the true extent to which a parameter could vary and still remain within the confidence region. Instead, the projection of the ellipsoid onto the parameter axis provides an accurate measure of the confidence interval for that parameter. The projection of the ellipsoid can be found through a series of steps in one parameter (solid lines of the staircase), with error minimization changes in the other parameters at each step (dashed lines of the staircase). Alternately, the projection can be found through the error variance–covariance matrix, as described in the text.

is often nontrivial, and the techniques reported below are designed to address this translation.

Once the best-fitting parameters are determined for the observed data, the confidence region can be found by perturbing the parameters. In the case of only one parameter, increasing and decreasing that parameter results in the specified increase in error, and the confidence region is a line segment. For two parameters, combined changes in the parameters result in the specified increase in error, and the form of the confidence region is an ellipse. In the case of many parameters, the form is an ellipsoid (assuming that error monotonically increases for small changes away from the best-fitting parameters, so that it can be approximated by a quadratic). Within this confidence region ellipsoid, a "naive" confidence interval could be found by changing a single parameter sufficiently to obtain the specified increase in error, while holding the other parameters fixed. However, this method does not take into account covariance between the parameters and will tend to underestimate parameter confidence (see Figure 1). Instead, what is required is the projection of the parameter ellipsoid onto each of the parameters.

There are several commonly accepted methods for identifying the shape of the confidence region that properly incorporate parameter covariance. With the advent of high-speed computers, a popular method is the nonparametric bootstrap (e.g., Efron & Tibshirani, 1993; Wichmann & Hill, 2001), which is achieved by repeatedly sampling with replacement subsets of the observed data. Variations in the best-fitting model parameters to each of these sampled data yield an estimate for the shape and extent of the confidence region. This technique is said to be nonparametric because it does not make any assumptions regarding the manner in which the data are distributed. A related technique is the parametric bootstrap, in which the model is used to generate new data according to the variability inherent in the model. These new data sets are then fit, and as with the nonparametric bootstrap, variations in the parameters determine the confidence region. The advantage of these bootstrap techniques is that they are straightforward and relatively assumption free. The disadvantages are that they can be computationally intensive, and, as will be seen in the third section, are potentially misleading when one is dealing with nonlinear models containing local minima.

Next, a computationally less intensive method is described that directly assesses the shape of the error space as determined from a relatively small number of parameter combinations. In this method, the variance terms (i.e., the diagonal elements) of the chi-square error variance–covariance matrix are directly related to the projections of the confidence ellipsoid onto each parameter. Such a relation is allowed if (1) the distribution of maximum likelihood error estimates (chi-square goodness-of-fit values in this case) as a function of the parameter estimates (i.e., best-fitting parameters) is normally distributed across the sampled data sets (i.e., for parameters producing data within the confidence region) and (2) the parameters are combined in a linear manner or the sample size is sufficiently large as to reject linear alternatives (Press, 1992). The normality assumption will be addressed next, and the issue of linearity will be addressed at the end of the third section.

Bayesian and multinomial models are currently popular computational accounts of behavioral data (e.g., Batchelder & Riefer, 1990; Huber et al., 2001; Ratcliff & McKoon, 1997, 2001; Schooler & Anderson, 1997; Shiffrin & Steyvers, 1997). However, the parameters in these models are probabilities, which are bounded and, therefore, inherently nonnormal. As was stated above, it is required that the best-fit likelihood error is normally distributed as a function of the parameters in order to map the variance–covariance matrix onto parameter confidence. Because parameter changes near the upper and lower bounds of 1 and 0 produce much larger changes in model behavior, as compared with equal-sized parameter changes near .5, probability parameters inevitably produce nonnormal error distributions. However, the assumption of normality can be approximated by transforming the parameters from probabilities to log-odds (i.e., log of the likelihood ratio), which serves to eliminate the boundedness of the probability parameters. Equation 1 converts probability parameters (p) to the log-odds parameters (L), and Equation 2 performs the reverse conversion:

$$L = \log\left(\frac{p}{1-p}\right), \tag{1}$$

and

$$p = \frac{1}{1+e^{-L}}. \tag{2}$$

In employing this transformation, the error minimization routines operate in the L-space. This log-odds transformation works to counteract skew for extreme probabilities that are close to 0 or 1. The outcome of normally distributed likelihood error as a function of the parameters is that the parameters within the confidence region should themselves be normally distributed (Fisher, 1922). Therefore, a check of error normality is achieved by examining the normality of the parameter distributions for values falling within the confidence interval (note that normally distributed error produces normally distributed parameters, but the converse is not necessarily true; nevertheless, nonnormal parameters conclusively demonstrate a violation of the error normality assumption). For the particular example of the ROUSE model reported in the second section, which contains parameters that are probabilities, 10,000 bootstrap samples were performed in order to ascertain that the log-odds transformation reduced skew. Comparing the probability and the log-odds parameter distributions, skew and kurtosis were reduced for all three ROUSE parameters. More specifically, the average skew across behavioral conditions reported in the third section was reduced from .69 to −.04, and the average kurtosis was reduced from 6.7 to 3.3.

An estimate of the parameter error variance–covariance matrix is found by taking the inverse of one half the Hessian (the Hessian is the matrix of second derivatives for

each parameter with respect to every other parameter; by convention, this matrix is multiplied by .5 to remove the factor of two terms). In the event that the Hessian cannot be calculated analytically or is not produced in the course of parameter optimization, a numerical estimation of the Hessian can be found through a Taylor series finite difference approximation of the second derivatives. Truncation error is minimized by using the centered difference approximation (Haberman, 1998), shown in Equation 3 for diagonal elements and Equation 4 for off-diagonal elements (Equation 4 assumes $\Delta x = \Delta y$):

$$\frac{d^2 f}{dx^2}(x_0) \approx \frac{f(x_0 + \Delta x) - 2f(x_0) + f(x_0 - \Delta x)}{(\Delta x)^2}, \quad (3)$$

and see Equation 4 at the bottom of the page.

In applying this numerical estimate of the Hessian, it may be necessary to adjust the finite difference ($\Delta x = \Delta y$) in order to avoid zero or negative terms along the variance–covariance matrix. These irregularities can occur when the finite difference is too small, resulting in zero second derivative terms due to computer roundoff error (i.e., small steps away from the best-fitting parameters may not produce appreciable error increases). The other requirement is that the Hessian is well conditioned (poorly conditioned matrices may not have a unique inverse). The Hessian is poorly conditioned if the finite difference is too large, so that increased error in some directions becomes much larger than increased error in the other directions. In the applications reported in the third section, these joint constraints were satisfied by progressively trying larger finite differences for the log-odds parameters in increments of .01. The Hessian with the lowest condition number that did not contain negative terms along its inverse was then used to calculate the variance–covariance matrix.

Once a well-conditioned Hessian is found, a confidence deviation for each parameter is determined according to Equation 5 (Press, 1992):

$$\delta_i = \sqrt{C_{ii} \Delta\chi^2}, \quad (5)$$

in which δ_i is the confidence deviation of parameter i, C_{ii} is the diagonal element of the variance–covariance matrix corresponding to parameter i, and $\Delta\chi^2$ is the specified increase in chi-square error in order to define a confidence limit. The δ_is are used to calculate lower and upper confidence limits for the probability parameters by taking the best-fitting log-odds parameter (L_i), adding and subtracting the appropriate δ_i, and then converting back to probability space with Equation 2.

In addition to the upper and lower confidence limits appearing in Table 1, the error variance–covariance matrix allows determination of the parameter correlation matrix. In this matrix, parameters that increase or decrease together in order to maintain a good fit have a positive correlation, whereas parameters that trade off against each other have a negative correlation. Equation 6 (Press, 1992) determined the parameter correlation matrices appearing in Table 2:

$$\rho_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}. \quad (6)$$

In the third section, both the error variance–covariance and bootstrap techniques will be applied to the ROUSE model described next. Table 1 shows the 99.99% confidence limits from the error variance–covariance analyses, as compared with appropriate quantiles from 10,000 bootstrap samples, and Table 2 shows the error variance–covariance correlations, as compared with Pearson's rho from the bootstrap samples.

An Analytic Method for Producing ROUSE Predictions

Huber et al. (2001) presented a new theory of short-term priming called ROUSE, for responding optimally with unknown sources of evidence. This theory was developed in order to explicate an otherwise confusing pattern of data obtained in perceptual identification experiments with two-alternative forced choice testing (2AFC). Experimental variables were observed to switch the direction of prime-induced bias, leading the authors to adopt the more theoretically neutral term, *preference*, instead of *bias* (Huber, Shiffrin, Lyle, & Quach, 2002; Huber et al., 2001; Huber, Shiffrin, Quach, & Lyle, 2002; Weidemann et al., 2005).

Before describing the new method for obtaining analytic ROUSE predictions, it is necessary to describe the basics of stochastic simulations with ROUSE. ROUSE is more clearly explained within the context of the forced choice priming paradigm (the present application considers repetition priming, for which the largest preference changes were observed). In this task, identification of a briefly flashed and masked target word is tested through a forced choice between target and foil (see Figure 2). Immediately prior to the target flash, two prime words are presented that allow priming of neither, one, or both of the choice words. The corresponding conditions are termed *neither primed*, *target primed*, *foil primed*, and *both primed*. Comparisons of the target-primed and foil-primed conditions allow measurement of the magnitude and direction of preference effects, and comparison of the both-primed condition with the neither-primed condition allows measurement of unbiased perceptual effects.

ROUSE consists of two mechanisms: source confusion, which corresponds to unknown sources of evidence, and discounting, which corresponds to responding optimally. Source confusion refers to feature activations arising from

$$\frac{d^2 f}{dxdy}(x_0, y_0) \approx \frac{f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0 - \Delta x, y_0 + \Delta y) + f(x_0 - \Delta x, y_0 - \Delta y) - f(x_0 + \Delta x, y_0 - \Delta y)}{4(\Delta x)^2}. \quad (4)$$

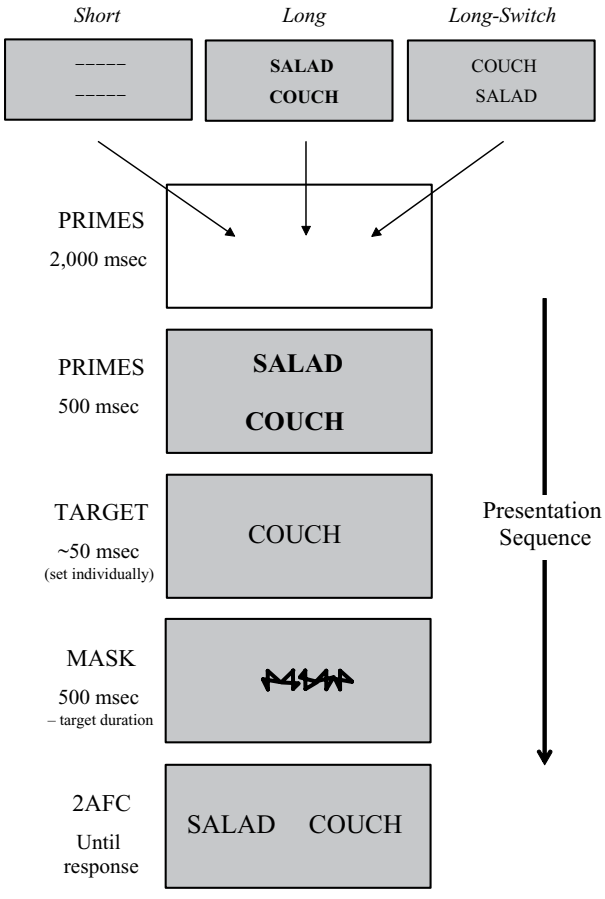

Figure 2. The sequence of visual displays used by Huber, Shiffrin, Quach, and Lyle (2002), Experiment 1. The task of the participant is to choose the word that matches the briefly flashed target. The particular words in the figure provide an example of repetition priming for the both-primed condition. Varying the relationship between the prime words and the choice words, so that only the target, only the foil, or neither choice word is primed, produces the target-primed, foil-primed, and neither-primed conditions. 2AFC, two-alternative forced choice.

the prime that are mistaken for the target. These feature activations result from a stochastic process, requiring averaging across a large number of simulated trials in order to obtain stable predictions. Instead, the new analytic technique calculates the asymptotic result of this stochastic process. Discounting refers to the assignment of a lowered level of evidence to features that may have been activated by the primes, rather than by the target. The amount of discounting, as compared with the amount of source confusion, can cause a preference for or against choosing primed words.

First, the stochastic nature of source confusion will be described. As can be seen in Figure 3, presentation of the primes activates primed choice word features with a probability of α, the target flash activates target features with a probability of β, and the pattern mask and other sources of noise activate any choice word feature with a probability of γ. Allowing for degrees of prime similarity less than identity, the parameter ρ determines the proportion of features that are shared between a prime and a primed choice word, and only these primed choice word features are potentially activated with a probability of α. In this manner, the combination of prime similarity (ρ) and source confusion (α) results in more active features in primed choice words and, therefore, a preference for primed choice words. In addition, this results in both-primed deficits due to variability (i.e., across simulated trials, there will be occasions on which the primes activate more features in the foil than in the target).

Next, the discounting of features will be described. Knowing which features might have been primed, an optimal decision assigns those features a lower level of evidence (i.e., discounting). If the prime could have been the source of activation, the probability that the target was the source of activation is lessened, resulting in less confirmatory evidence for primed features. Figure 4 shows the feature likelihood ratios for the various possibilities. In this 2×2 contingency table, features exist in either an active or an inactive state, and a prime may or may not have been a potential source of activation (i.e., the feature is or is not shared with a prime). It is assumed that the decision process does not know the actual activation probabilities and must use estimates of these parameters (designated by the prime symbol). Each activation probability requires an actual activation parameter and an estimate of that parameter. The actual parameters stochastically determine feature activation, and the estimates of the parameters determine the strength of evidence assigned to each feature. The crucial feature likelihood ratio is found in the lower right hand cell of Figure 4, which represents the discounted level for an active feature that may have been activated by a prime.

If one assumes feature independence, the likelihood ratios for all the features in a choice word are multiplied to provide a likelihood ratio for that choice word. These choice word likelihood ratios are then compared, and the larger is chosen. In the event of a tie, a choice is made randomly. In this decision process, the value of discounting (α'), as compared with the degree of source confusion (α), determines whether there is a preference for or against choosing primed choice words.

Huber et al. (2001) simulated ROUSE stochastically, with a particular number of activated and shared features randomly determined across many simulated trials. In the data fits reported by Huber, Shiffrin, Lyle, and Quach (2002), Huber, Shiffrin, Quach, and Lyle (2002), and Weidemann et al. (2005), an analytic version of ROUSE was employed, providing stable asymptotic model behavior in much less computational time.

The analytic version of ROUSE is achieved by calculating the probability of occurrence for all possible feature combinations of inactive (OFF), active-unprimed (ON), and active-primed (DIS) feature states across both the target and the choice words (active features that have been primed are discounted; hence, the term DIS). Only these three feature states are considered, because the two terms for inactive features are identical (see Figure 4). This feature combination probability space is tallied for those feature combinations resulting in the correct choice, yield-

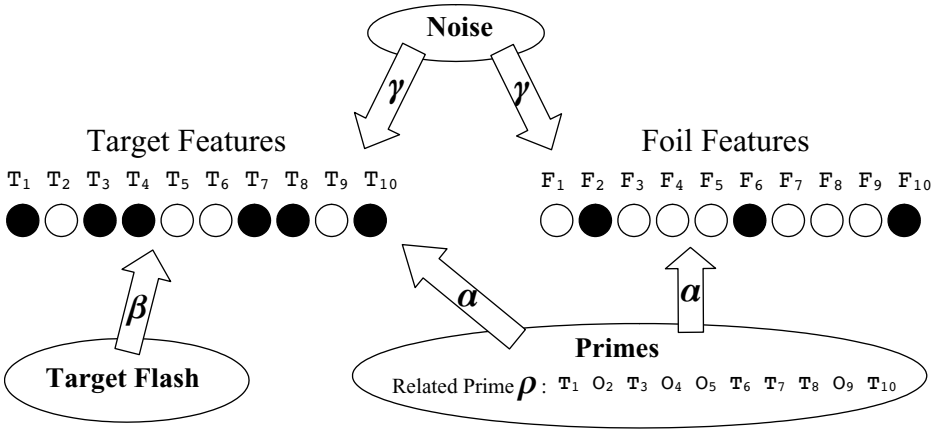

Figure 3. Source confusion through stochastic feature activation in ROUSE. For the perceptual identification paradigm, three sources of feature activation exist. With a probability of β, every feature in the target word is activated. Features shared between a prime and a choice word are activated with a probability of α, and a mediating parameter, ρ, probabilistically determines which features are shared. Noise activation is applied to all features with a probability of γ. Ten features per word are shown, although 20 features per word were used in the simulations. Because it is unknown by which source a feature has become active, this situation results in source confusion and, therefore, a preference to choose prime-related words.

ing an analytic accuracy prediction. This is functionally equivalent to simulating an infinite number of stochastic trials and is similar to the application of Luce's choice rule (Luce, 1963) for determining accuracy.

Because features are assumed to be independent, the basic calculation for the analytic version of ROUSE is a binomial distribution for *unprimed* choice words (features are either ON or OFF) and a trinomial distribution for *primed* choice words (features could additionally exist in the DIS state). These distributions are used to determine the probabilities of particular combinations of observed features for each choice word, and then these probabilities are multiplied in order to determine the probability of particular feature combinations across both choice words.

For unprimed choice words, there are only two possible feature states: OFF features and ON (not discounted) fea-

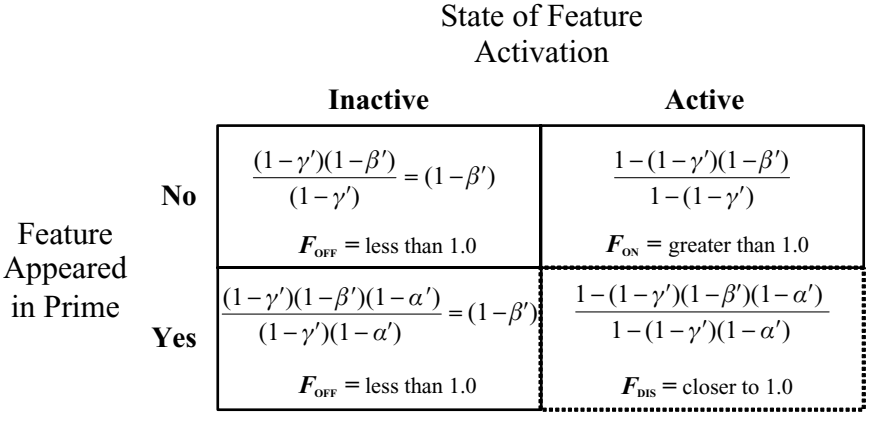

Figure 4. Discounting through feature likelihood ratios in ROUSE. The evidence provided by each feature in the decision process can be calculated as a likelihood ratio. These ratios are the probability that a feature exists in its observed state of activation given that it is part of the target divided by the probability given that the feature is part of the foil. The important contingencies are for inactive or active features that did or did not appear in a prime (i.e., may or may not have been activated by a prime). The likelihood ratio for active features that appeared in a prime (i.e., the lower right-hand cell) represents a discounted level of evidence, and with increasing estimates of prime activation, α' (i.e., increased levels of discounting), this ratio approaches the neutral value of 1.0. The three unique feature likelihood ratios, labeled F_{OFF}, F_{ON}, and F_{DIS}, correspond to the OFF, ON, and DIS feature states referred to in the description of the analytic version of ROUSE, and these are the feature likelihood ratios that appear in Equation 9.

tures. Because the target flash is an additional source of activation in the case of the features of the target word, the probability parameters for the binomial are different for a target than for the foil. For the features of an *unprimed foil*, the probability of a feature remaining OFF is $(1 - \gamma)$, because noise must have failed to activate the feature. The probability of such a feature existing in an ON state is one minus this probability (γ). For the features of an *unprimed target*, the probability of a feature remaining OFF is $(1 - \gamma)(1 - \beta)$, because both noise and the target flash must have failed to activate the feature, and the probability of a feature existing in an ON state is one minus this probability. Knowing these state probabilities (which determine p_{OFF} and p_{ON}), the probability of observing each possible combination of feature states across the N features of an unprimed choice word is calculated with Equation 7. These N features are broken down into n_{OFF} features in the OFF state (each occurring with a probability of p_{OFF}) and n_{ON} features in the ON state (each occurring with a probability of p_{ON}):

$$p(\text{unprimed word}) = \binom{N}{n_{OFF}} \left(p_{OFF}\right)^{n_{OFF}} \left(p_{ON}\right)^{n_{ON}}. \quad (7)$$

Similar procedures apply for primed choice words, although a trinomial distribution (Equation 8) must be used because features may additionally exist in the DIS state. In addition, calculation of the probabilities for primed choice words must also include ρ, the probability that a feature is shared with a prime (note that probabilistic use of prime similarity is appropriate only for situations in which the actual proportion of primed features varies from trial to trial). A *primed foil* that is ON (not discounted) is not shared with a prime (otherwise, it would be discounted) and could have been activated only by noise. Therefore, such a feature occurs with a probability of $(1 - \rho)\gamma$. Using similar logic, it follows that the probability of a DIS feature is $\rho[1 - (1 - \alpha)(1 - \gamma)]$, since the prime is an additional source of activation for such a feature. This equation follows by considering that the feature must be in a prime (ρ), and it cannot be the case that both the prime $(1 - \alpha)$ and noise $(1 - \gamma)$ failed to activate the feature. Finally, the probability of an OFF feature is one minus the sum of the other two probabilities. For a feature contained in a *primed target*, the target flash is an additional source of activation for the active states (ON and DIS), and therefore, the probability of an ON feature is $(1 - \rho)[1 - (1 - \gamma)(1 - \beta)]$, the probability of a DIS feature is $\rho[1 - (1 - \alpha)(1 - \beta)(1 - \gamma)]$, and the probability of an OFF feature is one minus the sum of the two other probabilities. Once these feature state probabilities are determined, Equation 8 (at the bottom of the page) calculates the probability of observing each possible combination of feature states in a primed choice word.

For each possible combination of feature states for a given choice word, as occurs with a probability determined by Equation 7 or 8, there is an associated word likelihood ratio that is calculated using the estimates of the activation probabilities (i.e., the product of the associated feature likelihood ratios). These word likelihood ratios are most easily found by taking the log of the product, turning the situation into a sum of the log feature likelihood ratios. As can be seen in Equation 9 at the top of the next page, the log of the OFF, ON, and DIS feature likelihood ratios appearing in Figure 4 are multiplied by the particular numbers of OFF, ON, and DIS features, producing the log word likelihood ratio (W) for the assumed combination of feature types (the terms, F_{OFF}, F_{ON}, and F_{DIS}, are adopted for the three unique values appearing in Figure 4).

In each priming condition, all possible feature state combinations for the target (i.e., different values for n_{ON}, n_{OFF}, and n_{DIS}) are compared with all possible feature state combinations for the foil. For each of these comparisons in which the target likelihood ratio (W_{TAR}) exceeds the foil likelihood ratio (W_{FOIL}), the target and foil probabilities of the particular feature state combinations (as determined by Equations 7 and 8) are multiplied, yielding the probability of the particular combination of feature state types across both target and foil ($p_{TAR}p_{FOIL}$). This resultant value is added to the tally of total correct. If the target likelihood ratio is exactly equal to the foil likelihood ratio, half the probability of such an occurrence of feature states is added to the tally (this assumes random guessing in a case of no differential information). In this way, an analytic expression for accuracy is calculated for each priming condition, as can be seen in Equation 10 at the bottom of the next page.

Huber, Shiffrin, Quach, and Lyle (2002, Experiment 1) did not use similar choice words, and so a fixed number of diagnostic features is used in the next section (it is assumed that the 20 features of each choice word are unique). In other experiments, choice word similarity was manipulated, in which case the analytic procedures described above are repeated for each possible number of diagnostic features (i.e., each possible N), with the results weighted according to the probability of occurrence of that number of diagnostic features. If one assumes that choice word similarity varies probabilistically across stimuli, the probability of observing a particular number of diagnostic features is determined through a binomial distribution with one minus the choice word similarity parameter determining the probability of a diagnostic feature.

A Specific Application of the Techniques From the First and Second Sections

Huber, Shiffrin, Quach, and Lyle (2002) found that prime exposure duration (Figure 1) causes a change in the magnitude of the preference for primed choice words.

$$p(\text{primed word}) = \binom{N}{n_{OFF}} \binom{N - n_{OFF}}{n_{ON}} \left(p_{OFF}\right)^{n_{OFF}} \left(p_{ON}\right)^{n_{ON}} \left(p_{DIS}\right)^{n_{DIS}}. \quad (8)$$

$$\log(W) = n_{\text{OFF}} \log(F_{\text{OFF}}) + n_{\text{ON}} \log(F_{\text{ON}}) + n_{\text{DIS}} \log(F_{\text{DIS}}). \quad (9)$$

This result is shown in Figure 5 and will serve as a test bed for the reported techniques. In the *short* condition, the two primes were presented for 500 msec in boldface font (this is the same as the passive-priming condition found in Huber et al., 2001). In the *long* condition, the two primes were presented for 2,500 msec in boldface font. In the *long-switch* condition, the two primes were presented for 2,000 msec in regular face font, followed by a swap of position and a final presentation for 500 msec in boldface font (this mimics the prime presentations that occurred in the active-priming conditions of Huber et al., 2001). The key finding was a reduction in the difference between the target-primed and foil-primed conditions in moving from short, to long, and then to long-switch, despite robust both-primed deficits for all three prime durations.

Because the sample size was large for this experiment (528 total data points for each of the conditions), chi-square fitting error increases rapidly with changes in the ROUSE parameters, even though the model behavior changes only slightly. In other words, small changes in the parameters produce small changes in the absolute difference between observed and predicted data, but these small differences are easily rated as significant increases in error, due to the statistical power employed in the experiment. Therefore, a large increase in chi-square was used in defining the parameter confidence region in order to produce sizable confidence limits, making it easy to visually compare the confidence limits of each of the parameters. This was achieved through an increase in chi-square of 23.5 (i.e., $\Delta\chi^2 = 23.5$; calculated on the basis of four degrees of freedom for the four priming conditions), corresponding to a 99.99% chance that the true parameters lie within the confidence region.

The best-fitting ROUSE behavior shown in Figure 5 was produced using the analytic method for producing ROUSE predictions and the Simplex algorithm (Nelder & Mead, 1965) for optimizing parameters so as to minimize chi-square error (the best-fitting parameters are shown in boldface in the upper rows of Table 1). The number of features in a choice word was fixed at 20. As was discussed by Huber et al. (2001), both the noise activation probability (γ) and the estimate of noise activation (γ') were fixed at .02, and the estimate of target perception (β') was set to its true value. Because the three prime durations were likely to induce different levels of target perception (β), due to changes in forward masking and attentional differences with primes of different durations, prime activation (α), due to changes in the level of source confusion, and estimates of prime activation (α'), due to changes in the availability of prime information, these three parameters were allowed different values for each prime duration, resulting in three free parameters for each set of four priming conditions. If one takes into consideration the number of free parameters, relative to the number of conditions, this is not a critical test of ROUSE, but it is useful for demonstrating the reported techniques and for investigations of the model's behavior more generally. Other work has provided more diagnostic tests of ROUSE, as compared with other theoretical accounts of priming data (e.g., Huber, Shiffrin, Lyle, & Quach, 2002).

In order to ascertain the reliability of the techniques for producing a parameter confidence region, both the error variance–covariance analysis (the upper rows of Table 1) and a bootstrap analysis (the lower rows of Table 1) were performed. In this situation with subject-averaged accuracy data, there is essentially no difference between a parametric bootstrap (sampling model predictions) and a nonparametric bootstrap (sampling observations). Sampling from either the model or the data is the result of a single Bernoulli process for each experimental condition, and furthermore, the best fit to the data is nearly perfect, and so the Bernoulli process suggested by the data is nearly identical to the one produced by ROUSE. To produce variability in accord with the error variance–covariance chi-square confidence region, the bootstrap was performed by sampling 10,000 times from appropriate binomial distributions for each of the experimental conditions. These binomial distributions were defined by using the observed accuracy probabilities and the observed number of data points. With 10,000 samples, the 99.99% confidence interval is defined by the range of the bootstrap parameters when the quantile method is used.[1]

As can be seen in Table 1, the upper and lower 99.99% confidence limits associated with the β parameters are close to the best-fitting parameter values. In contrast, the confidence limits associated with α and α' parameters deviate more from the best-fitting values. These patterns are observed regardless of the technique for assessing parameter confidence. Because the number of features and the effect of noise (γ) were set to fixed constants, only the activation of features due to the target flash (β) determines the overall performance level across all four priming conditions, and so β is greatly constrained by the data. The data also highly constrain the relationship between α and α' (i.e., the amount of discounting, as compared with the amount of source confusion); for underestimates of α, target-primed is greater than foil-primed (e.g., the short condition) whereas for overestimates of α, target-primed is less than foil-primed (e.g., the long-switch condition). However, the absolute magnitudes of α and α' are not as important for producing these situations as are the relative magnitudes of α and α'. For this reason, a range of α and α' values produce reasonable fits.

$$p(c) = \sum_{W_{\text{TAR}} > W_{\text{FOIL}}} p_{\text{TAR}}\, p_{\text{FOIL}} + \frac{1}{2} \sum_{W_{\text{TAR}} = W_{\text{FOIL}}} p_{\text{TAR}}\, p_{\text{FOIL}}. \quad (10)$$

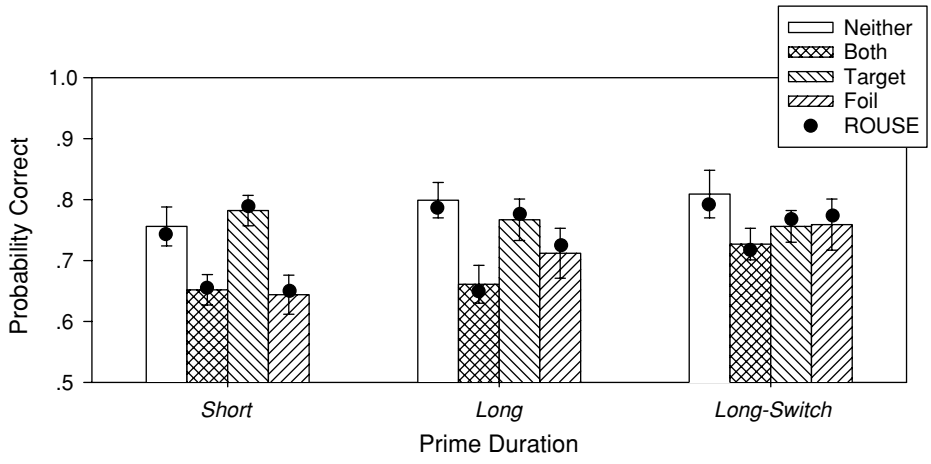

Figure 5. Observed and simulated results from Huber, Shiffrin, Quach, and Lyle (2002, Experiment 1). The *short* duration presented primes for 500 msec and is the same as the passive priming conditions in Huber, Shiffrin, Lyle, and Ruys (2001), and the *long-switch* prime duration mimicked the active priming conditions by presenting the primes for 2,000 msec, followed by switching the positions of the two primes and presenting them for an additional 500 msec. The *long* prime duration presents the two prime words for 2,500 msec without a position switch. The error bars are plus and minus one standard error of the observed means. The filled circles are the result of analytic ROUSE with best-fitting parameters.

The error variance–covariance and bootstrap analyses are in agreement for the confidence limits on β, but for the α and α' parameters, the confidence limits for the bootstrap version are somewhat larger. In the particular case of α' for the short and long-switch conditions, the bootstrap analysis produced a very different median best-fitting parameter, as compared with the parameter that best fit the observed data. Figure 6 helps explicate these differences by plotting the error landscape in ROUSE, demonstrating the role of local minima. Both panels of Figure 6 plot the relationship between α (source confusion) and α' (discounting) for long-switch prime presentations. The left panel plots the 10,000 bootstrap samples (collapsing over β), and the right panel plots chi-square error based on the observed data as α and α' are varied while β is held fixed at its global best-fitting value. Better fits to the observed data are shown as darker shades of gray.

Table 1
Huber, Shiffrin, Quach, and Lyle (2002, Experiment 1): Best-Fitting ROUSE Parameters With Lower and Upper 99.99% Confidence Limits, Shown to the Left and Right of the Boldface Best-Fitting Parameters

	Condition		
Parameter	Short	Long	Long-Switch
α (prime actual)	.039 **.074** .135	.069 **.144** .276	.023 **.049** .101
	.028 **.072** .238	.069 **.142** .283	.019 **.048** .138
α' (prime estimate)	.034 **.059** .099	.102 **.135** .175	.056 **.075** .099
	.011 **.048** .216	.050 **.137** .266	.018 **.046** .161
β (target flash)	.032 **.045** .064	.042 **.057** .077	.045 **.059** .077
	.034 **.045** .058	.043 **.057** .071	.047 **.059** .070
$\Sigma\chi^2$ (error)	.728	1.510	2.247

Note—For each parameter, the top row shows the error variance–covariance analyses, and the bottom row the bootstrap analyses. The boldface values for the bootstrap are calculated using the quantile method at .00005, .5 (the median), and .99995. The sum of chi-square error in the bottom row is the best fit to the observed data.

Figure 6 reveals that there are several local minima (several modes) in the parameter space, mostly lying along the diagonal where α and α' are approximately equal. Only the long-switch condition is displayed, but similar multimodal landscapes exist for the short and long prime presentations. The multimodal nature of this error landscape is due to the accrual of differing numbers of prime-activated features as α increases (e.g., moving from a situation in which the prime is likely to produce two additional active features to a situation in which the prime is likely to produce three additional active features). With each somewhat discrete change in the number of prime-activated features, the amount of discounting (α') must also increase if the observed preference is to remain constant.

There are potential complications for both the bootstrap technique and the error variance–covariance technique, due to these local minima. The error variance–covariance (exemplified by the right-hand panel) is calculated over the global minimum for the observed data (the darkest patch), and so this analysis does not include the neighboring local minima. Because the bootstrap analysis is sensitive to these nearby minima, due either to noise in the sampling process or to a failure to find the global minimum in the course of parameter optimization, the confidence limits for the bootstrap analysis are larger for the α and α' parameters. In the case of the long-switch condition (as well as for the short condition), there was a further complication revealed by the presence of two nearby regions of the parameter space that were nearly identical in their goodness of fit. For both these regions, α is approximately .05, whereas α' can be either .05 or .075. For the observed data, setting α' at .075 produced a slightly better fit (hence,

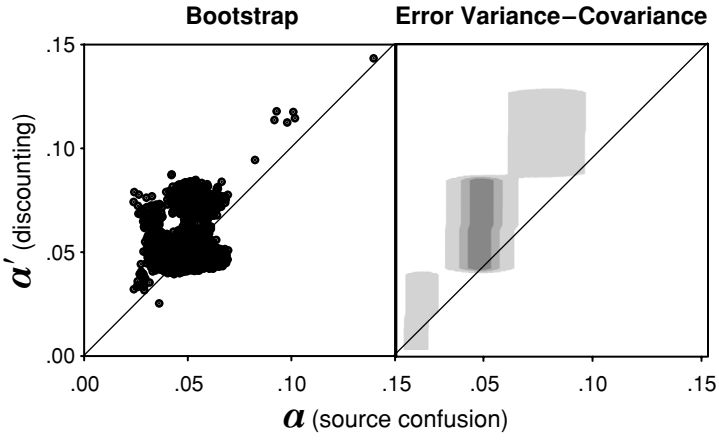

Figure 6. Source confusion (α) and discounting (α') parameters for the long-switch condition as produced by 10,000 bootstrap samples (left panel) and the error landscape calculated from the observed data (right panel). The bootstrap samples collapse over values of β, and the error landscape was calculated with β fixed at its best-fitting value. These graphs demonstrate that ROUSE is multimodal in the parameter space, with several local minima lying along the diagonal (i.e., discounting matched to the level of source confusion). Because of these local minima, the error variance–covariance confidence limits underestimate the range of acceptable parameter values, and a bootstrap analysis is potentially misleading if care is not taken to make sure the global best fit is found for every sample (see the text for details).

this is where the error variance–covariance analysis was centered), but statistically expected variations in the data are commonly fit by setting α at .05 (a similar situation occurred for the α' parameter in the short condition).

It appears that the bootstrap analysis is more informative because it captures the other modes of the parameter space. However, the existence of local minima can be particularly problematic for a bootstrap analysis. Unlike the error variance–covariance analysis, which produced positive parameter correlations, initial attempts at bootstrap sampling produced negative correlations. This occurred because the error minimization routines often failed to discover the global best fit when the global best fit varied to nearby modes (starting values for the parameters were chosen on the basis of the parameters that best fit the observed data, and these starting values prevented the optimization routines from discovering the global minimum). By first determining more appropriate starting values for the parameters, this pitfall was avoided in the reported bootstrap analyses. To determine these starting values, β was optimized to fit the sampled neither-primed condition (which does not depend on α or α'). Next, α was optimized to fit the sampled both-primed condition (which does not depend on α' for repetition priming). Finally, α' was optimized to fit the sampled target-primed and foil-primed conditions (which critically depend on the relationship between α and α'). With these new starting values, all three parameters were then optimized simultaneously to fit all four conditions. In the absence of these procedures, the bootstrap analysis was misleading, producing greatly reduced confidence intervals and negative, rather than positive, parameter correlations.

The parameter correlation matrices in Table 2 are in agreement with the discussion above. The correlations for the error variance–covariance analyses were calculated according to Equation 6, and the correlations for the bootstrap analyses are traditional Pearson's rho values. In either case, all of the parameter correlations are positive and, in most cases, have similar quantitative values. The positive relationship between α and α' was discussed above and is explained by the need to match the level of discounting to the level of source confusion in order to produce a particular magnitude of preference. The positive correlations between β versus α and β versus α' are also sensible. If statistically expected variations in the data produce overall better accuracy, this can be accommodated by increases in β. However, with more target-activated features due to higher β, there needs to be more prime-activated features (α), in order to produce sizable priming effects, and more discounting (α'), in order to accommodate more prime activation features.

The strong correlations between parameters highlight the nonlinearities inherent in ROUSE. A linear reparameterization of ROUSE would more adequately meet the requirements necessary for relating the error variance–covariance matrix to confidence values and parameter correlations. In particular, this could be accomplished by assuming that performance is related to three additive terms: a positive term for the amount of target perception (similar to β), a negative term for the amount of prime-induced noise (similar to α), and a term for the amount of preference (which would take on positive or negative values). The last term would be similar to the comparison of α with α', representing the combined effect of source

Table 2
Huber, Shiffrin, Quach, and Lyle (2002, Experiment 1): Parameter Correlation Matrices

	Short			Long			Long-Switch		
Parameter	α	α'	β	α	α'	β	α	α'	β
α	1.000	.679	.597	1.000	.829	.483	1.000	.485	.444
	1.000	.529	.380	1.000	.424	.402	1.000	.170	.376
α'	.679	1.000	.393	.829	1.000	.276	.485	1.000	.215
	.529	1.000	.176	.424	1.000	.314	.170	1.000	.159
β	.597	.393	1.000	.483	.276	1.000	.444	.215	1.000
	.380	.176	1.000	.402	.314	1.000	.376	.159	1.000

Note—For each parameter, the top row shows the error variance–covariance analyses, and the bottom row shows Pearson's rho calculated from the bootstrap parameter values.

confusion and discounting. Such a linear alternative could handle these data, although it would do so in a purely descriptive manner. More specifically, Huber, Shiffrin, Lyle, and Quach (2002) found that the efficacy of discounting is modulated by experimental factors affecting overall performance levels and prime similarity. These results were a priori predictions of ROUSE but would be inexplicable for a linear alternative model.

REFERENCES

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability Vol. 57). Washington, DC: Chapman & Hall/CRC.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London: Series A*, **222**, 309-368.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, **6**, 258-269.

Haberman, R. (1998). *Elementary applied partial differential equations: With Fourier series and boundary value problems* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Huber, D. E., & Cousineau, D. (2003). A race model of perceptual forced choice reaction time. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 687-692). Mahwah, NJ: Erlbaum.

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science*, **27**, 403-430.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Quach, R. (2002). Mechanisms of source confusion and discounting in short-term priming: 2. Effects of prime similarity and target duration. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 1120-1136.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, **108**, 149-182.

Huber, D. E., Shiffrin, R. M., Quach, R., & Lyle, K. B. (2002). Mechanisms of source confusion and discounting in short-term priming: 1. Effects of prime duration and prime recognition. *Memory & Cognition*, **30**, 745-757.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. E. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-187). New York: Wiley.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

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

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

Press, W. H. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). Cambridge: Cambridge University Press.

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, **104**, 319-343.

Ratcliff, R., & McKoon, G. (2001). A multinomial model for short-term priming in word identification. *Psychological Review*, **108**, 835-846.

Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review*, **83**, 190-214.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, **32**, 219-250.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**, 145-166.

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

Visser, I., Raijmakers, M. E. J., & Molenaar, P. C. M. (2000). Confidence intervals for hidden Markov model parameters. *British Journal of Mathematical & Statistical Psychology*, **53**, 317-327.

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

Wagenmakers, E.-J., Zeelenberg, R., Huber, D. [E.], Raaijmakers, J. G. W., Shiffrin, R. M., & Schooler, L. J. (2003). REMI and ROUSE: Quantitative models for long-term and short-term priming in perceptual identification. In J. S. Bowers & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 105-123). London: Oxford University Press.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (2005). Confusion and compensation in visual perception: Effects of spatiotemporal proximity and selective attention. *Journal of Experimental Psychology: Human Perception & Performance*, **31**, 40-61.

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

NOTE

1. The quantile method for producing confidence intervals from bootstrap samples is nonparametric, making no assumptions regarding the shape of the parameter distribution. However, this method is appropriate only when taking at least 1,000 bootstrap samples, which is the case for the reported analyses. In addition, the quantile method may be biased, producing a confidence interval that is somewhat shifted, as compared with the original parameters fit to the observed data. Instead, a biased-corrected method can be employed by determining an offset factor for this bias, analyzing the proportion of bootstrap parameters lying below the original parameters. For the reported analyses, such proportions were within a few percentage points of .5 in nearly all the conditions, suggesting little or no bias (if the proportion is exactly .5, the quantile method

is equivalent to the biased-corrected method). A related measure of bias is a comparison of the median of the bootstrap samples with the original parameters. For the bootstrap analyses, Table 1 reports median bootstrap parameters, and for the error variance–covariance analyses, Table 1 reports the original parameters that best fit all the empirical data. With two exceptions, these values are within .002 of each other, suggesting little or no bias for most cases. For these reasons, and because the biased-corrected method includes an assumption of normality in employing the *z*-transform, the biased-corrected method was not used. See Mooney and Duval (1993) for a detailed discussion of the various techniques for producing confidence intervals from bootstrap samples.

(Manuscript received April 9, 2003;
revision accepted for publication July 19, 2005.)