Psychological Bulletin
1993, Vol. 114, No. 3, 510-532

0033-2909/93/$3.00

Methods for Dealing With Reaction Time Outliers

Roger Ratcliff

The effect of outliers on reaction time analyses is evaluated. The first section assesses the power of different methods of minimizing the effect of outliers on analysis of variance (ANOVA) and makes recommendations about the use of transformations and cutoffs. The second section examines the effect of outliers and cutoffs on different measures of location, spread, and shape and concludes using quantitative examples that robust measures are much less affected by outliers and cutoffs than measures based on moments. The third section examines fitting explicit distribution functions as a way of recovering means and standard deviations and concludes that unless fitting the distribution function is used as a model of distribution shape, the method is probably not worth routine use.

Almost everyone who has analyzed reaction time data has been faced with the problem of what to do with outlier response times. Outliers are response times generated by processes that are not the ones being studied. The processes that generate outliers can be fast guesses, guesses that are based on the subject's estimate of the usual time to respond, multiple runs of the process that is actually under study, the subject's inattention, or guesses based on the subject's failure to reach a decision. For almost any theoretical or empirical purpose, it is desirable to eliminate outliers from the data. However, eliminating outliers requires unambiguously identifying them, and that is extremely difficult. The problem is that the distribution of response times from the real processes under study overlaps to a great extent the distribution of outlier response times. So the best we can hope to do is to reduce the effects of potential outliers while eliminating as little as possible of the data of real interest.

In this article, I present a series of simulation studies of response time distributions. The results of these studies offer practical recommendations for how to deal with outliers. The article contains three sections that progress from the purely practical question of how to improve the power of analysis of variance (ANOVA), to the issue of how outliers affect descriptive statistics for response time distributions, and then to the issue of fitting explicit models to distributions that may contain outliers. The three sections of the article move from a pragmatic emphasis on hypothesis testing to the theoretical domain of modeling.

In the first section of the article, distributions of response times with and without outliers are simulated and the power of ANOVA to find significant differences between distributions is investigated as a function of several possible ways to deal with outliers. In the second section, I examine the variability in measures of the shape, spread, and location of reaction time distributions under several schemes for eliminating outliers. Correlations among the different measures show to what extent the measures evaluate the same things. In the third section I present an examination of the use of explicit theoretical distribution functions to describe the shapes of empirical reaction time distributions and to allow the mean and standard deviations of the empirical distributions to be recovered.

Reaction Time Models

To evaluate the effectiveness of various methods for eliminating outliers, simulated reaction time distributions, both with and without outliers, were generated. Reaction time distributions are usually skewed to the right. This means that the distributions rise rapidly and then fall off slowly with a long skewed tail. For the simulations reported in this article, two explicit theoretical distribution functions were used to mimic empirical reaction time distributions. The first was the convolution of the normal and the exponential distributions (called the ex-Gaussian by Luce, 1986). It has been used successfully as a convenient summary of empirical reaction time distributions in a range of experimental paradigms (Heathcoate, Popiel, & Mewhort, 1991; Hockley, 1982, 1984; Hohle, 1965; Ratcliff, 1978, 1979, 1981, 1988a, 1988b; Ratcliff & Murdock, 1976; Ulrich & Miller, 1992). The distribution has three parameters: the mean of the normal, μ, the standard deviation of the normal, σ, and the parameter and mean of the exponential, τ. The expression for the ex-Gaussian is

$$f(t) = \frac{e^{-[(t-\mu)/\tau]+\sigma^2/(2\tau^2)}}{\tau\sqrt{2\pi}} \int_{-\infty}^{\{[(t-\mu)/\sigma]-\sigma/\tau\}} e^{-y^2/2} dy. \quad (1)$$

The mean of the ex-Gaussian is $\mu + \tau$ and the variance is $\tau^2 + \sigma^2$. Because empirically the size of σ is usually not more than one fourth the size of τ, the standard deviation in the distribution is approximately τ (i.e., $\sqrt{[1 + (1/4)^2]} = 1.03$). The parameter σ approximately represents the rise in the left tail of the distribution, and τ approximately represents the fall in the right tail. I have found in fitting this distribution to data that $\mu + \tau$ is

This research was supported by National Institute of Mental Health Grants HD MH44640 and MH00871 to Roger Ratcliff and Air Force Office of Scientific Research Grant 90-0246 (jointly funded by the National Science Foundation) to Gail McKoon. I thank Gail McKoon for extensive comments on this article as well as James Townsend, William Hockley, and an anonymous reviewer.

Correspondence concerning this article should be addressed to Roger Ratcliff, Department of Psychology, Northwestern University, Evanston, Illinois 60208.

a good approximation to the mean of the data. When random samples of data are generated from the theoretical distribution, then if μ is higher than the population μ, τ will most often be lower than the population τ and vice versa (leading to a negative correlation in the parameter estimates).

To use the ex-Gaussian to generate a simulated data point, a random number drawn from a normal distribution is added to a random number drawn independently from an exponential distribution (see Ratcliff, 1979). Repeating this process with different random numbers yields a simulated distribution of response times.

The other theoretical distribution function used for the research presented here was the inverse Gaussian or Wald distribution. It is the distribution of finishing times derived from a one-boundary diffusion process (the continuous version of the random walk). This distribution has been shown to be an acceptable model of reaction time distributions (e.g., Luce, 1986; Ratcliff, 1978, 1988b). The expression for the inverse Gaussian is

$$f(t) = \sqrt{\left(\frac{\lambda}{2\pi(t-t_{er})^3}\right)} e^{-[\lambda(t-\theta-t_{er})^2]/[2\theta^2(t-t_{er})]}, \quad (2)$$

where the parameters of the model are θ, λ, and t_{er}, and $\theta + t_{er}$ is the mean of the distribution and the standard deviation is $\sqrt{(\theta^3/\lambda)}$. An algorithm for generating random numbers from the inverse Gaussian is available (Chhikara & Folks, 1989, p. 53). Another form of this distribution written in terms of the diffusion process (Ratcliff, 1978) is

$$f(t) = \frac{z}{\sqrt{2\pi s^2(t-t_{er})^3}} e^{-[z-u(t-t_{er})]^2/[2s^2(t-t_{er})]}. \quad (3)$$

The translation between the forms is: $\theta = z/u$ and $\lambda = z^2/s^2$.

There are a number of other theoretical functions that have been used to fit reaction time distributions such as the lognormal, gamma, and so forth (see Luce, 1986, Appendix B; Ratcliff & Murdock, 1976; see also Ulrich & Miller, 1992). These other functions are not considered in this article. However, if some aspect of the results appeared to be critically dependent on specific distributional assumptions, then it would be straightforward to examine these other distributions in ways parallel to those reported here.

Types of Outliers

There are two major types of reaction time outliers: short and long. There may also be spurious reaction times that overlap the center of the distribution of normal responses, but these are impossible to identify. Short reaction times occur with some frequency in experiments in which mean reaction time is short (e.g., choice reaction time; Swensson, 1972), but in other situations where reaction time is longer (e.g., more than 500 ms) and subjects are cooperating, they are infrequent (e.g., Ratcliff & Murdock, 1976). When they do occur, with uncooperative subjects or subjects tested by an experimenter showing lack of interest, they are usually easy to spot (e.g., the reaction time distributions are bimodal, the accuracy of the fastest responses is at chance, or both) and so they are not considered here. However, long spurious reaction times are almost certainly always present, usually overlapping with long genuine reaction times. Separating genuine from spurious response times is by no means a simple issue. Some models might predict a high tail in the reaction time distribution, and what might be outliers for one model, might not be outliers for another model. However, to some extent, these theoretical considerations are independent of the experimenter's desire to improve the power of experimental analyses.

To illustrate the difference in detectability of short and long outliers, four example distributions are shown in Figure 1. Response times to represent the real process, the one under study, were produced from an ex-Gaussian distribution with parameters μ = 400 ms, σ = 40 ms, and τ = 200 ms. One thousand randomly drawn response times from this distribution are plotted in the histogram in the top left corner. To produce a mixture of the real process and outliers, 800 random numbers were generated from the ex-Gaussian distribution for the real process, and 200 random numbers were generated from the same ex-Gaussian distribution with μ = 600 ms (bottom left), μ = 200 ms (top right), and μ = 800 ms (bottom right). For outliers in the right tail, the histograms show that it is difficult to detect whether or not there is a mixture even when μ = 800 ms, a mean of about 2 standard deviations above the mean of the real distribution. In contrast, when the outliers are in the left tail, the distribution becomes either markedly more symmetrical or humped in the left tail. Response times in the left tail are much more easily identified as problematic compared with noncontaminated distributions than response times in the right tail. In practice, outliers in the left tail show up in one or a few subjects' conditions and comparisons with distributions from other subjects or conditions can be used to identify them. These results can be summarized as follows: Short outliers stand alone; long outliers hide in the tail.

I should stress that even with the outlier distribution displaced 2 standard deviations above the mean, it is almost impossible to determine whether there are outliers in the tail. In the example in Figure 1, it is difficult to decide whether the hump in the right tail is spurious or the product of a mixture of two real processes. In practice, outliers are even more difficult to spot than in this simple case. The outliers' distribution most likely has a much greater standard deviation than the distribution of the real process, giving a long and elevated tail, because outliers often come from spurious events such as loss of attention, daydreaming, distraction, and so forth. Determining whether any particular group of extreme reaction times contains mostly real responses or mostly outliers is extremely difficult. Therefore, finding methods that minimize the effects of suspect observations is an important aim of statistical methods in the analysis of reaction time. In fact, the goal for our models and empirical research should be to account for the middle 85–95% of the observations in our reaction time distributions; these are the data that are most likely to come from the real processes under consideration and also most likely to be critical in testing hypotheses and models.

The statistics literature provides a number of tests for outliers, tests that determine whether there are one, two, or some undetermined number of outliers (e.g., Barnett & Lewis, 1978;

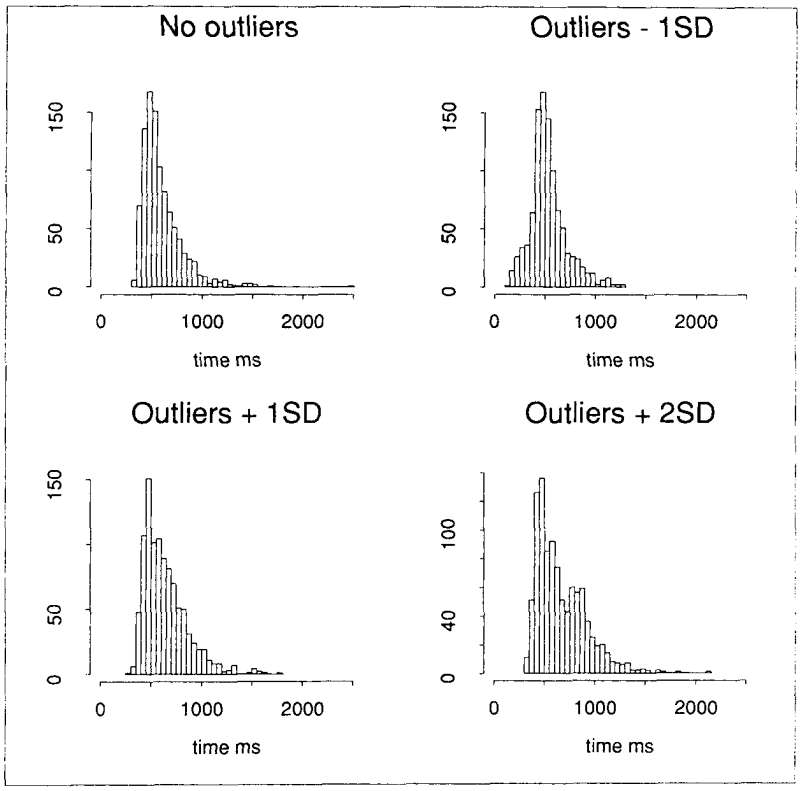

Figure 1. The mixture distribution produced by mixing a distribution with the same distribution shifted by one standard deviation to the right or left or two standard deviations to the right. (The mixture distribution has 20% of the total observations.)

Lovie, 1986; Shapiro & Wilk, 1972), suggesting that there are methods already available to address many of the questions concerning outliers discussed here. However, the specifics of the reaction time domain severely limit the applicability of these methods. The methods I have found in the literature have dealt with distributions such as the Gaussian, exponential, or gamma distributions that are not generally appropriate for reaction time distributions. It is certainly possible to develop such methods for the classes of distributions that have been applied to reaction time, but as far as I know this has been done only for the exponential and gamma distributions (see Barnett & Lewis, 1978). There are other methods for reducing the influence of outliers (e.g., M-estimators, Barnett & Lewis, 1978) that weight response times differentially as a function of distance from the center of the distribution. These methods might improve power of the data analyses by a small amount, but significant effort would be needed to determine the form of the function used to weight the data, and so far as I know, there has been no such effort in the reaction time domain.

So what should be done? A single extremely long outlier can increase the mean, inflate the standard deviation, and change measures of shape such as skewness by a very large amount (see Ratcliff, 1979). The aim is to lessen the impact of such outliers by trimming them out of the data, by using robust statistics, or by using transformations that minimize their effects. In the next section, I examine the power of ANOVA under various schemes to minimize the influence of outliers. The power was compared for reaction time distributions with and without outliers and for distributions with outliers eliminated by various rules.

Power of Analysis of Variance Under Outlier Minimization

Simulations Using the Ex-Gaussian Distribution

Analysis of variance of reaction times is used to detect whether reaction time distributions across experimental conditions are significantly different from each other. To conduct an ANOVA of simulated data requires generating reaction times from a theoretical distribution for each experimental condition and then performing an ANOVA on the reaction times. In what follows, the same distribution was assumed for each experimental condition except that the mean was changed in some of the conditions relative to others (e.g., by adding 40 ms). To examine power, 1,000 replications of this process (generating simulated data and running an ANOVA on it) were performed and the number of analyses that gave a significant difference out of 1,000 is reported.

The experiment that was simulated was chosen to mimic typical psycholinguistic and lexical decision experiments and other

experiments in which there were limited numbers of observations per condition per subject (because of limited numbers of experimental stimuli). The simulated experiment had four conditions, seven observations per condition, and 32 subjects. Response times in two of the conditions were increased to produce a main effect. The parameters for the ex-Gaussian distribution of the process of real interest were: $\mu = 400$ ms, $\sigma = 40$ ms, and $\tau = 200$ ms. The size of the main effect was 20, 30, or 40 ms and was introduced in μ for some studies or in τ for other studies. Variability across "subjects" was added by selecting a random number from a rectangular distribution with range from −50 to 50 ms (small variability across subjects) or −150 to 150 ms (large variability across subjects) and adding it to μ (a discussion of putting subject variability in τ is presented later). A response time statistic (e.g., the mean, median, or trimmed mean) was computed for each subject condition, and these statistics were compared in a $32 \times 2 \times 2$ ANOVA.

To include outliers in the experiment, reaction times were selected either from the distribution used for the real process with probability 0.9 or from an outlier distribution with probability 0.1. This means that a particular condition could have zero, one, two, or even more outliers out of seven observations. The outlier distribution was the same as the real process distribution except that a random time chosen from a rectangular distribution with range 0 ms to 2,000 ms was added to a reaction time from the real process distribution. In psychological terms, this mimics a loss of attention that would cause a subject to delay the beginning of processing. The method for choosing outliers was designed to introduce inliers as well as outliers, that is, spurious observations through almost the whole distribution. Without systematic empirical studies, this assumption seems reasonable. If other assumptions about the outlier distribution were advanced, simulations similar to those reported here could be performed to compare results.

Common methods used to eliminate outlier reaction times include using the median response time, using specific cutoff response times, and using cutoffs at some number of standard deviations above the mean response time. For comparisons of the power of different methods, the complete response time distribution (with no responses eliminated) was used as a baseline. The methods for eliminating outliers compared with this baseline were as follows: eliminating all responses longer than a cutoff value, transforming the data, trimming the mean by eliminating the longest response time in each condition for each subject, calculating medians instead of means, eliminating response times above some value determined by standard deviations, and Windsorizing (replacing observations 2 standard deviations above the mean by observations at 2 standard deviations above the mean; Barnett & Lewis, 1978). I examined five different cutoff values, two different transformations (log and inverse; each reaction time was converted to log or inverse before taking the mean for that condition), and two different standard deviation values, 1 standard deviation above a subject's mean (in which the standard deviation was calculated over all four experimental conditions) and 1.5 standard deviations above the mean. Note that in the studies that follow, absolute times were used to determine the cutoff and these times are used in the presentations of the data. But in applying these methods to real data, it would be at least as valuable to use percentages of errors eliminated by the cutoff.

There were six different sets of studies with ex-Gaussian distributions, and each study had a main effect difference between two of the four experimental conditions. Two of the six studies had the main effect in μ; one study had outliers and one did not. Two more studies had the main effect in τ; one with outliers and one without. The last two studies both had outliers, and one with the effect in μ and one with the effect in τ, but they differed from the first four studies in that there were large differences in subject mean reaction times in relation to the standard deviation of the distribution (i.e., the range could be 1.5 times the standard deviation). For each study, ANOVA provided F values for the two main effects and their interaction. The level of significance was set at 0.05. I expected that about 5% of the interactions and about 5% of the main effects for which there was no real effect would appear spuriously significant. The important question concerns the real main effect (for which 20–40 ms was added to each response time); the question is which methods of eliminating outliers led to the most uniformly high power for detecting this effect across the six studies?

Results

None of the methods for dealing with outliers affected the alpha level, and the number of significant F values out of 1,000 for the null main effect and the interaction varied randomly between about 35 and 65 across the conditions, studies, and methods of dealing with outliers.

Figure 2 shows the results from the two studies with and without outliers for which there was a 30-ms effect in μ. For the no outlier case, when a cutoff value was adopted to eliminate outliers, power increased as the cutoff was reduced from no cutoff at all to 1,000 ms. The median and trimmed mean had low power, and the standard deviation cutoffs and the two transformations had higher power. The differences in power among these methods were the smallest found across all the studies.

The bottom curve shows the effect of the introduction of outliers. The power with a 1,000-ms cutoff was about 0.6, but this dropped to around 0.2 as the cutoff was increased until there was no cutoff. The loss of power is due to the increased variability in the mean differences among conditions as a result of the presence of long spurious observations. This result shows the danger of ignoring the presence of outliers; power can easily drop to a third of what it would be under ideal conditions (no outliers). With reduced cutoff values (down to 1,000 ms), power came to within a few percent of what it would have been without outliers. For the trimmed mean, median, and log transformations, outliers reduced power. But outliers left the inverse transformation and the standard deviation cutoffs still relatively high in power.

I should stress that these results are conditional on the 30-ms effect being in μ, so that the whole distribution of reaction times shifted to the right. Cutting off the right tail reduced noise without disturbing the 30-ms effect. Figure 3 shows that the pattern of results was somewhat different when the effect was in τ, elongating the right tail of the distribution (a more common pattern of results in reaction time studies than a shift of the whole dis-

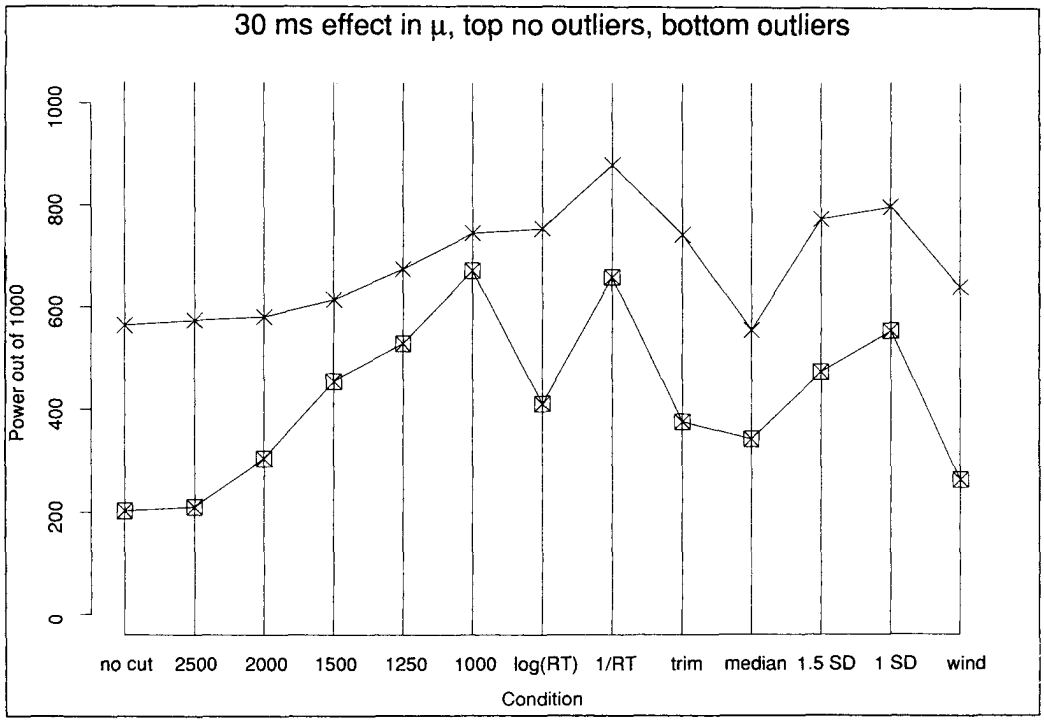

Figure 2. The power of analysis of variance for different conditions for the ex-Gaussian distribution with a 30-ms effect in μ with no outliers (crosses) and 10% outliers (boxes).

tribution). When the effect was in the tail, cutting off values reduced the effect, thereby reducing power. Introducing outliers into the right tail introduced a trade-off between reducing the effect of the outliers and eliminating real responses that were an important part of the effect. These results are discussed in detail in the next paragraphs.

When the effect was 40 ms in the τ parameter and there were no outliers, cutoffs had the opposite effect as when the effect was in μ and there were no outliers (see Figure 3, top curve). Eliminating responses longer than a cutoff value decreased power; as the cutoff value was decreased, cutting off a larger and larger proportion of responses (to 8.5% at the 1,000-ms cutoff), the power decreased from a high of 70% to a low of less than 40% (Figure 3). The standard deviation cutoffs had low power and the two transformations and the trimmed mean had relatively high power. Again, the median had lower power than the optimal cutoff. Overall, the power was less with a 40-ms effect in the tail of the distribution (τ parameter), 0.7 maximum power, than when the effect was in a shift of the whole distribution by 30 ms (μ parameter), 0.9 maximum power (Figure 2). Windsorizing provided high power because long genuine reaction times were replaced by long genuine times at two standard deviations. But this is the only case that Windsorizing had high power.

When the effect was in the tail (40 ms in the τ parameter) and outliers were introduced, the overall power decreased from when there were no outliers (Figure 3, bottom curve). Cutoff values increased power as the values decreased from no cutoff to a cutoff of 1,500 ms, but further decreases in the cutoff values reduced power as more and more of the response times from the real distribution were eliminated. Figure 3 shows that the inverse transformation had higher power than the log, the median, the trimmed mean, and the standard deviation cutoffs.

Other simulation studies examined the effect of placing subject variability in τ compared with μ above (i.e., spreading the distribution as opposed to shifting it). This had little impact on the power of the ANOVA. There were only two main differences. The first was a decrease in overall power in a particular measure from usually a few percent to an occasional decrease of 20% or 30% when subject variability was in τ relative to μ. The second effect was for the standard deviation cutoff: when the experimental effect was in τ and subject variability was in μ, power for the 1 standard deviation cutoff was lower than for the 1.5 standard deviation cutoff (Figure 3). When the subject variability was in τ, the power for the 1.5 standard deviation cutoff was higher than for the 1 standard deviation cutoff.

Figure 4 shows results for the two studies in which the subject variability was large relative to the standard deviation for individual subjects' distributions. The motivation for these studies came from analyzing real data, which is presented in a later section. Two cases are shown, one with a 20-ms effect in μ and one with a 20-ms effect in τ, but both with a 300-ms range in subject variability. Both cases have outliers from the same distribution as above (a random number from 0 to 2,000 ms added).

Results show that power as a function of cutoff is similar to the patterns in Figures 2 and 3. However, the relative power of other measures changes. In particular, the standard deviation

cutoffs have power near the optimal cutoff, and the median is now equal to or better than the inverse transformation. The reason that the standard deviation cutoff increases in power is that outliers are trimmed relative to the base reaction time for that subject, whereas with cutoffs, outliers for a fast subject are not touched but real responses for another subject are trimmed. In the studies just discussed, subject variability was small enough that the distributions for individual subjects were hardly separated.

Simulations Using the Inverse Gaussian Distribution

To ensure that the results of these simulations were not idiosyncratic to the ex-Gaussian distribution, a similar set of studies was conducted with the inverse Gaussian distribution (Equation 2). To provide a main effect difference between two of the four experimental conditions, the drift rate in the inverse Gaussian was increased by 40 ms. This is similar to having 20% of the 40-ms effect in μ and 80% of the effect in τ in the ex-Gaussian. Subject variability in the inverse Gaussian was introduced in the drift rate and was 100 ms in range as in the first four studies. Outliers were generated from the distributions in the same way as was just discussed.

Figure 5 shows the resulting power values with outliers (bottom curve) and without outliers (top curve). In general, the profiles look similar to those produced by the ex-Gaussian with a 40-ms effect in the τ parameter. The power is higher with the inverse Gaussian but the qualitative features are the same.

Analysis of Two Experiments

To further examine the power of analysis of variance under various methods for eliminating outliers, I computed F statistics for the data from two real experiments. Two lexical decision experiments described in an article by McKoon and Ratcliff (1992, Experiments 1 and 3) were used. These experiments were selected because the effects are very small and so push the limit of experimental methods. Experiment 1 had four conditions in a 4 × 1 design and there were 52 subjects. There were 9 observations per subject per condition when the subject made no errors in the condition. Experiment 3 had two groups of 44 subjects with two conditions per group (a 2 × 2 design), and 12 observations per subject per condition when there were no errors.

All of the same methods for eliminating outliers used with the simulations just presented were applied to the data from the two experiments. The results are graphed in Figure 6 in terms of values of the F statistic. For the earlier simulations, results were graphed in terms of power, defined as the number of significant F values out of a thousand simulations. This number correlates highly with the average F value for the thousand simulations. To show this, the average value of F for four of the sets of simulations (those in Figures 3 and 4) was compared with the number of F values that were significant; the correlation was 0.989, showing that the average F value tracks power very closely.

The values of the F statistic for the different methods are

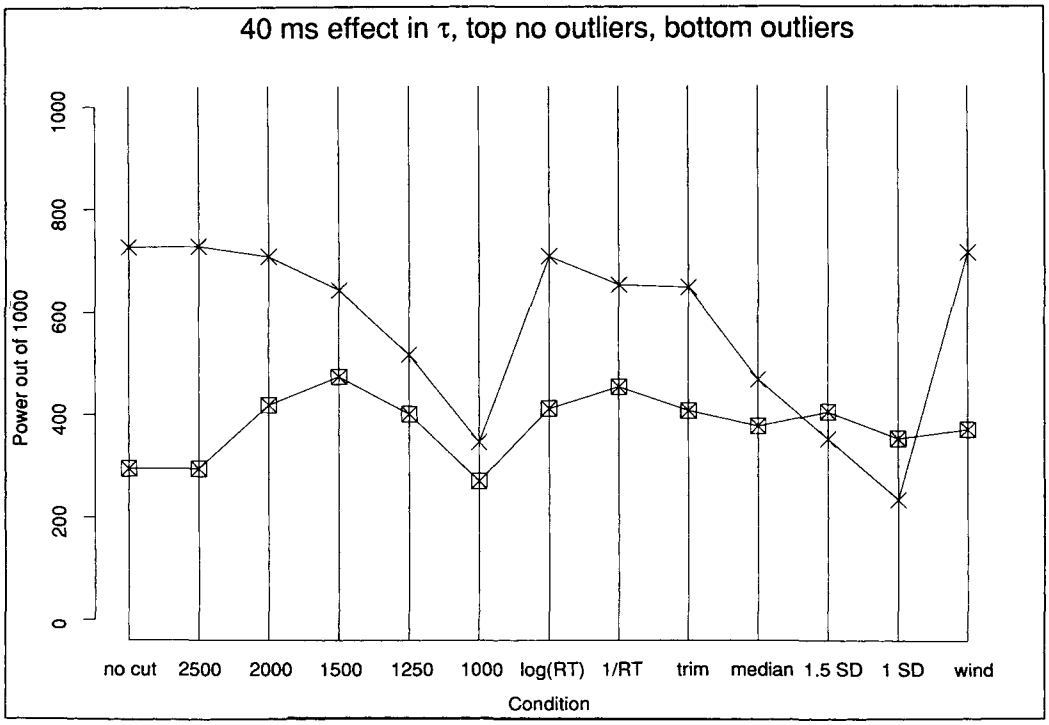

Figure 3. The power of analysis of variance for different conditions for the ex-Gaussian distribution with a 40-ms effect in τ with no outliers (crosses) and 10% outliers (boxes).

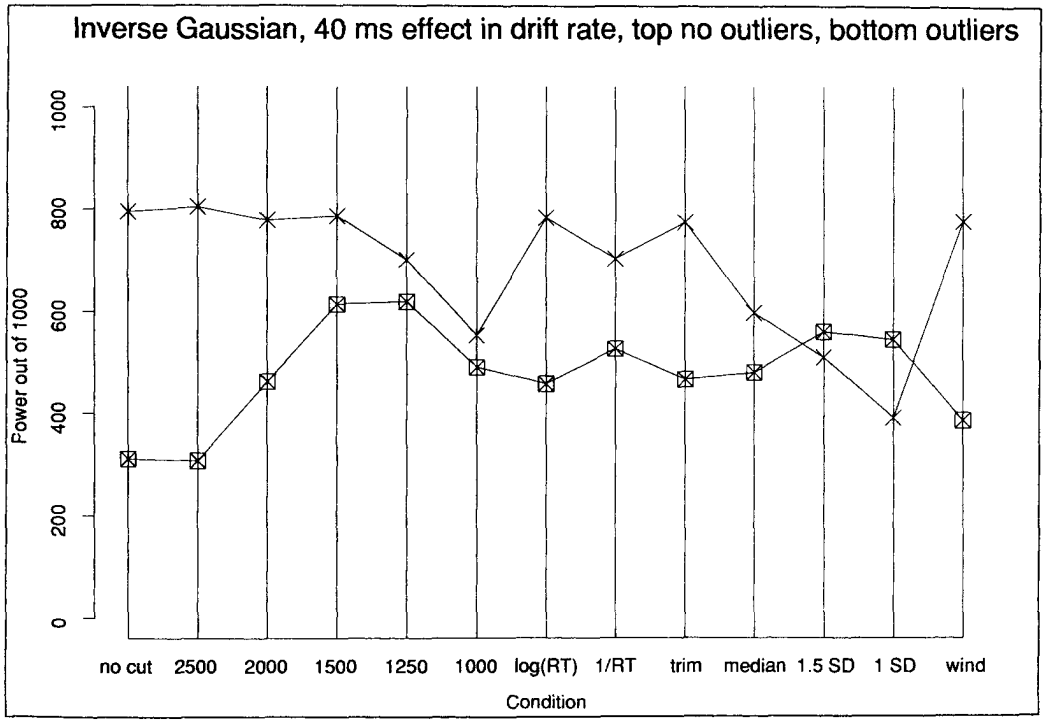

Figure 4. The power of analysis of variance for different conditions for the inverse Gaussian distribution with a 40-ms effect in drift rate with no outliers (crosses) and 10% outliers (boxes).

shown in Figure 6. A significant F value for these experiments is about 4.0, and this line is drawn horizontally on the figure. The pattern of responses for the crosses (Experiment 3) mimics the pattern found when an effect is in τ for an ex-Gaussian distribution contaminated with outliers (and also mimics the similar case for the inverse Gaussian). Note that the standard deviation cutoff is reasonably high because of high between-subjects variability (individual subjects' means ranged from 470 ms to 770 ms). For this experiment, the inverse transformation is not far from optimal.

The pattern from Experiment 1 is much like that shown in Figure 4 with ex-Gaussian distributions with large variability among subjects and small effect size. In fact, in the data, subject mean reaction time varies from around 450 ms to 750 ms. For this experiment, trimming at 1 standard deviation gives a larger F than does a low cutoff. Unlike Experiment 3, the inverse transformation is some distance from optimal. One difference between this experiment and Experiment 1 is that here, the size of the effect is smaller relative to subject variability.

Discussion

The first important general result from both the simulations and analysis of the experiments from McKoon and Ratcliff (1992) was that greatest power is obtained by eliminating response times longer than some specific cutoff value. But, the specific value of the cutoff for greatest power varied depending on whether the distribution included outliers and on how the distribution shape changed as a function of the increases in mean reaction time associated with the experimental conditions. When the difference between conditions was in μ, eliminating long reaction times increased power because faster reaction times are more stable than longer reaction times. When the difference between conditions was in τ and there were no outliers, eliminating long reaction times eliminated reaction times that were responsible for the effect, and so power decreased. When there were outliers, cutting off extreme reaction times increased power until the elimination of the real reaction times responsible for the effect had a greater impact than the elimination of outliers.

It follows from these results that a practical recommendation would be to find out how distribution shape is changing as a function of increases in the mean (e.g., Heathcoate et al., 1991; Hockley 1982, 1984; Ratcliff, 1978, 1979, 1981, 1988a; Ratcliff & Murdock, 1976). There are several ways to do this: Collect large amounts of data for a few subjects for some experimental conditions, collect smaller amounts of data from more subjects and construct distributions for the group (Ratcliff, 1979), or find appropriate results in the literature. Then a cutoff value can be chosen that is appropriate for the distribution shape. The previous simulations used absolute cutoffs (e.g., 1,000 ms or 2,500 ms). More properly, cutoffs should be selected as a function of the proportion of responses eliminated. For example, in the previous simulations, a 1,000-ms cutoff resulted in the elimination of as much as 10% of the data when there were no outliers and up to 15% of the data when there were outliers. A

reasonable range is to choose cutoffs that vary from no cutoff at all to a cutoff that eliminates 10% or 15% of the data. If the distribution is spreading, it would be reasonable to choose a cutoff value that eliminated approximately 5% of the observations, or if the distribution is shifting, a value that eliminated approximately 10% of the distribution. To confirm ANOVAs generated on data trimmed by cutoff values, analyses can be performed on data transformed by the inverse function. This transformation was only a few percentage points lower in power than specific cutoff values except when there was large variation in subject means.

A second feature of the results is the suggestion to examine the size of an effect relative to the differences in average response times across subjects. If there are large differences among subjects and the effect is small, then a cutoff that is based on individual subject standard deviations may also give high power (in addition to the cutoffs and inverse transformations just noted).

Some of the simulations were repeated with the effect as an interaction instead of a main effect. The same patterns of results were obtained as in Figures 2 and 3.

Generality. The generality of these results is an important issue because recommendations should be as widely applicable as possible. It is important to determine if the recommendations apply for manipulations such as altering the size of the effect in μ, altering the size of the effect in τ, making the size of the effect variable across subjects, altering the proportion of outliers, and altering the precise form of the distribution used in the simulations. First, altering the size of the effect in μ produces the same profiles of power for a set of simulations with a 20-ms effect in μ compared with the 30 ms effect for the simulations in Figure 2. In all conditions in which the effect was in μ and between subject variability was not too large relative to the standard deviation for any subject, a low cutoff or the inverse transformation gave the greatest power. Second, altering the size of the effect in τ does not change the profile shown in Figures 3 and 5. With variability between subjects small relative to any subject's standard deviation, the same pattern was obtained with an 80-ms effect in τ as with a 40-ms effect in τ (Figure 3). Third, a further study examined varying the size of the effects across subjects (some subjects having a large effect, some almost no effect at all). With the effect in τ varying from 0 to 80 ms randomly chosen from a uniform distribution, the pattern of results was the same as with τ set to 40 ms for all subjects. Thus differences in the size of the effect between subjects does not alter conclusions about the relative power of different methods. Fourth, changing the proportion of outliers from 0.9 to 0.8 reduced power but left the pattern across different conditions the same. Fifth, as was just pointed out, the ex-Gaussian and the inverse Gaussian produced similar results for similar changes in distribution shape, showing that the results are not tied to the precise form of the distribution assumed for these simulations. The general conclusion is that the simulations presented here capture the relative merits of different methods of dealing with outliers for a wide range of effects for distributions that are shaped like reaction time distributions. If the distributions under study are not of this form (significantly more normal, bi-

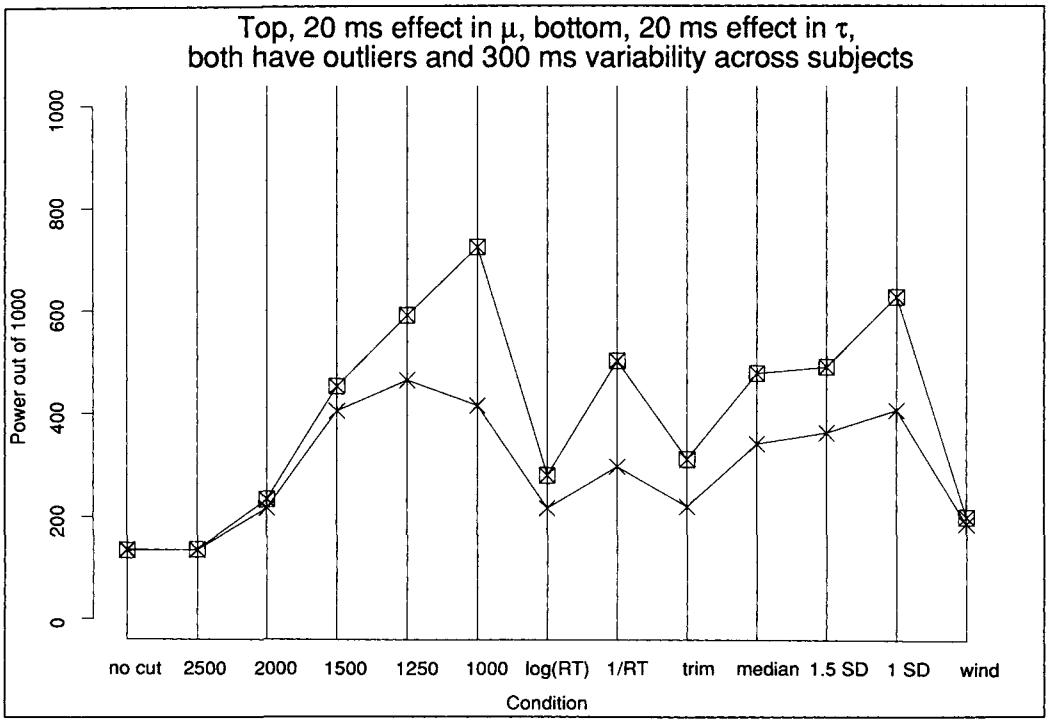

Figure 5. The power of analysis of variance for different conditions for the ex-Gaussian distribution with a 20-ms effect in τ (crosses) and a 20-ms effect in μ (boxes) with 300-ms variability across subjects.

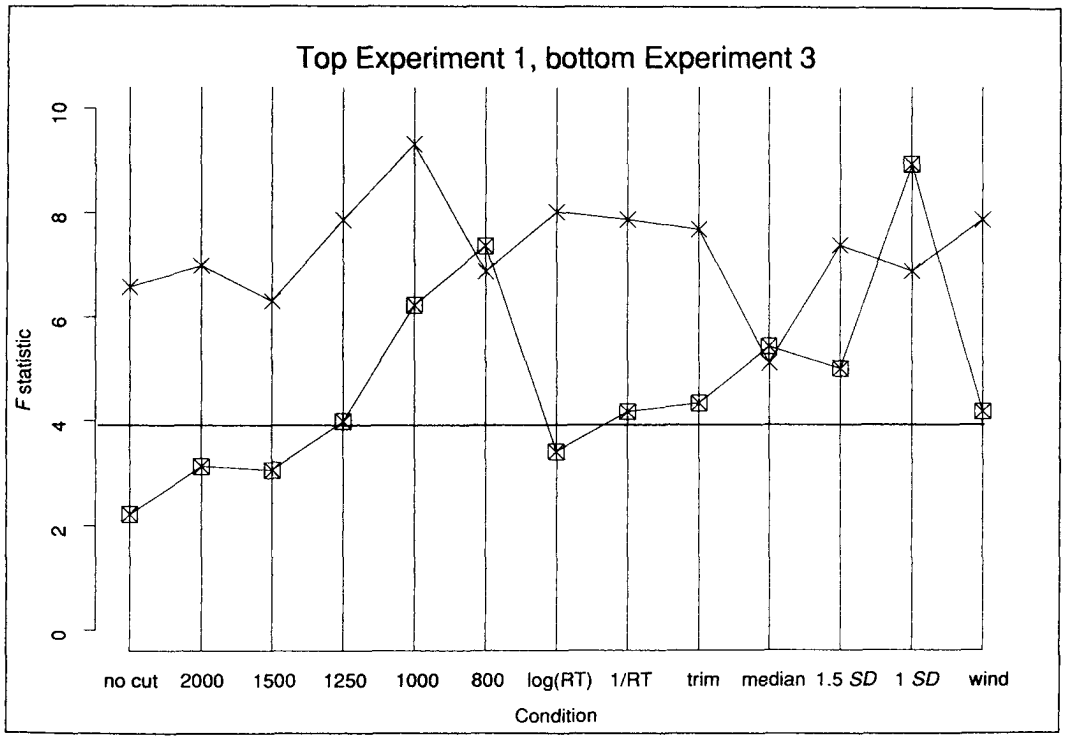

Figure 6. *F* statistics for two experiments from McKoon and Ratcliff (1992); Experiment 1 (boxes) and Experiment 3 (crosses). (The horizontal line is the 0.05 significance level.)

modal, varying greatly in spread across subjects, etc.), then studies similar to those above should be performed based on distributions that mimic the observed distributions.

Transformations. Figure 7 shows how transformations change distribution shape. The density functions illustrate two experimental conditions generated from ex-Gaussian functions, one shifted from the other by a 40-ms change in the mean by a change in μ. When the inverse transformation is used on these distributions, the distributions become more skewed to the left, the differences in the leading edges of the distributions are magnified (note that the left tail in the density function becomes the right tail in the inverse transformation), and the differences in the right tail become smaller. For the log transformation, the distributions are a little less skewed than the originals, and the difference in the tails of the two distributions is smaller in relation to the difference in the leading edges than for the original distributions.

For two ex-Gaussian distributions that differ in the τ parameter (Figure 8), one distribution has a lower peak than the other and a more elongated tail. Although the difference between the tails appears small, examination of the horizontal difference between the two curves shows that the difference is actually quite large. As before, the log transformation reduces the skewness of the distributions and the inverse minimizes the effects of long reaction times.

For two inverse Gaussian distributions that differ in mean θ because they differ in drift rate, the effects of transformations are about the same as for ex-Gaussian distributions that differ in the τ parameter. The log transformation tends to normalize the distributions and the inverse transformation leads to slightly right skewed distributions.

Figures 7 and 8 allow determination of what the effects of the transformations would be if the distributions included outliers. Both the inverse and log transformations reduce the impact of long response times in the tails of the distributions, and therefore would reduce the impact of long outliers, leading to higher power for ANOVA. The log transformation does not reduce the importance of the extreme values as much as the inverse transformation and so produces a smaller increase in power. The increase in power was shown in Figures 2–6, and Figures 7 and 8 show that the improvement in power occurs because the effect of reaction times in the tail of the distribution is de-emphasized.

Error responses. It is important to consider here how to deal with error responses. Errors can have different distributional properties from correct responses (e.g., see Ratcliff & Murdock, 1976). The empirical question is how to handle the loss of data from error responses, that is, whether errors should be eliminated and extra trials run in the conditions giving errors to replace the error data, or whether error responses should be replaced by the mean or median of the condition. Both these methods have been used; the first carries the risk that response criteria change from early conditions to those in which the extra trials are run, so that the trials with most errors would be affected most by the change in criteria. The second reduces variability for those conditions with a large number of errors. But neither of these problems will be fatal under well-controlled sit-

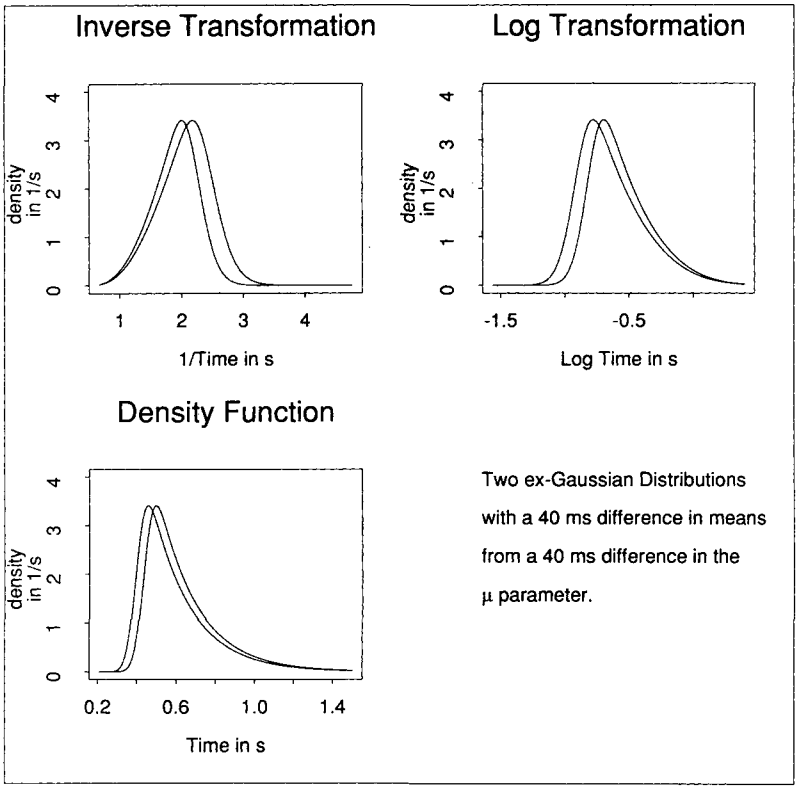

Figure 7. An example of the effect of log and inverse transformations on the ex-Gaussian distribution for distributions with a 40-ms difference in μ.

uations. However, these procedures can be finessed by using a summary statistic for the subject condition in the data analysis. In the previous analyses, there were seven observations per condition, and with trimming or cutoffs, there might be only four or five observations left to determine the statistic (median, mean, etc.). The elimination of errors work the same way; eliminating errors simply reduces the number of data points contributing to the condition (unless very few or no correct responses were left).

Recommendations

The main aim of ANOVAs in the kinds of experiments investigated here is to demonstrate the reliability and replicability of differences among experimental conditions. In our field, the most frequent way of analyzing reaction time data is to use cutoffs that are the same across conditions and that eliminate a small percentage of responses (not a fixed percentage). Although the results of the simulations show that this is not always the best way, it is probably too much to hope for the field to change radically so a reasonably conservative recommendation is as follows:

1. Try a range of cutoffs and make sure that an effect is significant over some range of nonextreme cutoffs.
2. Use the inverse transformation (or standard deviation cutoffs if subject variability is large) to confirm the cutoff analyses.
3. If the effect is novel, unexpected, or important, replicate it or partially replicate it in another experiment.
4. Most important, choose the method before analyzing the data; do not use several methods and choose only the one that is significant. Even if only one or two F values are significant out of 10 or 12 analyses using different methods for eliminating outliers, then the effect can still be real, especially if the profile of F values mirrors one of the profiles presented here. In such cases in which most of the analyses produce nonsignificant F values, if the profile is meaningful and the significant F values are from conditions in which significant F values would be expected, then the experiment should be repeated with more subjects, better control, or both.

Measures of Central Tendency, Dispersion, and Shape

The first section of this article dealt with the very practical question of how to lessen the impact of outliers and increase the power of standard ANOVA. This second section moves to the issue of the effects of outliers and methods to eliminate them in standard statistics for describing distributions. Generally speaking, this moves from the hypothesis testing questions of the prior section to questions about estimating the statistics of distributions for model testing or for describing empirical effects on the distributions (for example, is the increase in mean linear, or does variability increase as a log function).

The aim of this section is to examine the behaviors of various measures of central tendency, spread, and shape for response time distributions under conditions of outliers versus no outliers and cutoffs versus no cutoffs (see also Ratcliff 1979). There are two main questions to be asked about these measures: First, how likely are they to give a close estimate of the distribution statistic that they are designed to measure, and second, how variable are they when the data includes outliers, the data is trimmed at some cutoff value, or both? These are two different kinds of variability: The first is what is the standard deviation in the estimate of a statistic as a function of cutoffs and outliers, and the second is how much the estimate varies because of outliers or because a cutoff is used to eliminate outliers. A statistic can have a low value on one of these and a high value on the other, as is shown in the following paragraphs. For example, one statistic for which the standard deviation of its estimate for a set of data is small may have widely varying estimates (each with small standard deviation) as a function of cutoffs or outliers, whereas another statistic that has a larger standard deviation may vary little as a function of cutoff or outliers.

It is important to distinguish between the two sources of variability. For hypothesis testing, the standard deviation in a statistic should be as small as possible so that power is as high as possible: The smaller the standard deviation in measuring the statistic, the more likely a difference between two experimental conditions is to be significant. But, if the statistic is very sensitive to outliers or cutoffs, then any experiment in which conditions might be affected differentially by outliers or cutoffs could have reduced power (or an increased probability of a false-positive outcome). For example, if the difference between two experimental conditions is that one distribution of response times is shifted in relation to the other, then using the same cutoff for both conditions could eliminate enough responses from the slower distribution that power to detect a difference in some measure between the conditions would be reduced. Or, two conditions might differ only in their relative numbers of outliers, so that a measure insensitive to outliers is needed to avoid falsely finding significance for a spurious effect. In general, a measure insensitive to cutoffs reduces the likelihood of finding a significant difference between conditions when none actually exists or finding a nonsignificant difference when one really does exist. There can be theoretical reasons for the use of different statistics for testing particular predictions of models or for estimating model-based parameters. The discussion can be useful in providing a sample method for performing analyses of such model-based quantities.

To illustrate the impact of even a single outlier on the traditional measures of shape (skewness and kurtosis derived from moments), a simple simulation was performed. Two distributions of 100 reaction times each were generated from an ex-Gaussian distribution with τ = 200 ms, μ = 600 ms, and σ = 40 ms; for one of the distributions, a single response time was

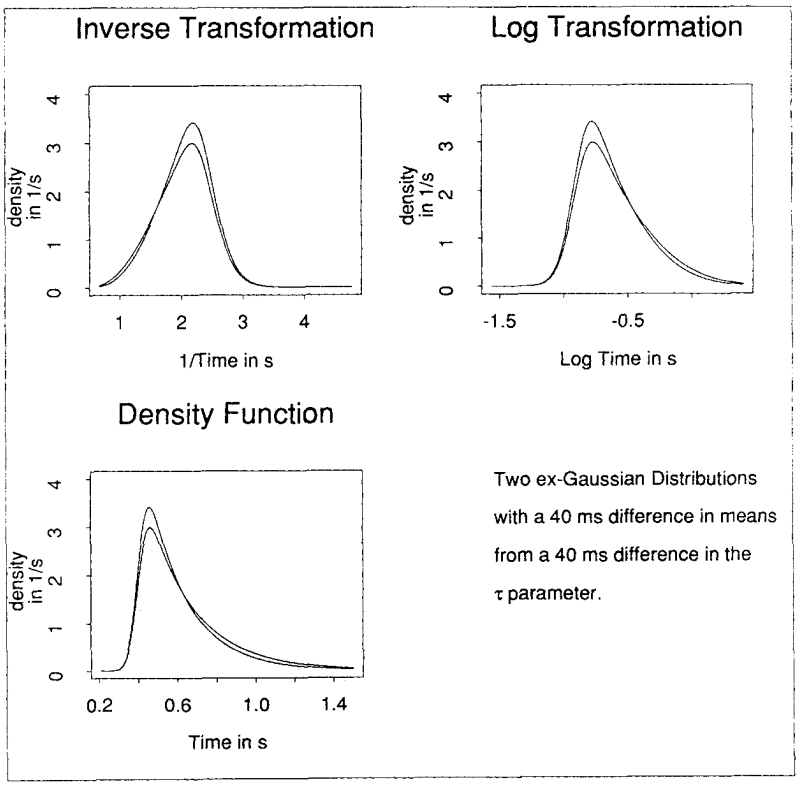

Figure 8. An example of the effect of log and inverse transformations on the ex-Gaussian distribution for distributions with a 40-ms difference in τ.

Table 1
Estimates of Parameters With and Without a Single Outlier

Outlier	*M* (s)	Variance (s^2)	3rd Moment (s^3)	4th Moment (s^4)	Skewness	Kurtosis
No outlier	802	3.95×10^4	1.37×10^7	1.01×10^{10}	1.75	6.47
Outlier at 2.5 s	821	6.75×10^4	5.88×10^7	8.86×10^{10}	3.35	19.45

replaced by an outlier response time equal to 2.5 s. Table 1 shows the mean, variance, 3rd and 4th moments, and skewness and kurtosis for the two distributions. Note that an outlier at 2.5 s is not uncommon in experiments in cognitive psychology in which mean reaction time is in the range of 600 to 900 ms (see Table 3, Ratcliff, 1979). The results in Table 1 show dramatically the effect of one outlier on the higher moments and derived measures. For example, the second, third, and fourth moments are changed by factors of two, four, and eight, respectively, and the measures of skewness and kurtosis are changed by factors of two and three, respectively. Ratcliff (1979) presented similar results: In a large experiment, the first four moments were calculated with outliers between 2 and 5 s included and with these outliers trimmed out. In these real experimental data, the changes were even more dramatic than those shown in Table 1 (for example, the fourth moment changed by as much as a factor of 30). From a practical point of view, such dependence on outliers is devastating for the use of moments in reaction time work (see Ratcliff, 1979).

Table 2 shows a number of different measures for distributions. To evaluate the effects of outliers on these measures and how the effects of outliers can be reduced, simulated distributions were generated. First, 1,000 sample reaction times from an ex-Gaussian distribution with $\mu = 500$ ms, $\sigma = 40$ ms, and $\tau = 200$ ms were generated, and the various statistics were computed over these 1,000 response times. Then, this was repeated for a total of 100 simulations. From the 100 replications, the mean and standard deviation in the mean for each statistic were calculated, and these are displayed in Table 2. These statistics were calculated for distributions with no outliers and for distributions with 10% outliers; the outliers were generated from the same ex-Gaussian as the real reaction times but with an added random delay of from 0 to 2,000 ms (from a rectangular distribution). The statistics were calculated either with no response times eliminated or with all response times longer than 1,000 ms eliminated.

Table 2 shows four different measures of central tendency (mean, median, harmonic mean, and trimmed mean with the slowest 10% of responses eliminated), three measures of spread (standard deviation, mean deviation, and quartile deviation), and four measures of shape (skewness based on the second and third moment, Pearson's second measure of skewness, quartile skewness, and kurtosis). The harmonic mean (μ_h) is defined as $1/\mu_h = \sum_i(1/t_i)$, the mean deviation is the sum of the absolute values of the differences between the mean and each reaction time; the quartile deviation is $(Q_3 - Q_1)/2$, where Q_3 is the third quartile and Q_1 is the first quartile; Pearson's second measure of skewness (Pearson 2) is $3\ (M - Mdn)/SD$, where SD is the standard deviation and quartile skewness = $(Q_3 - 2Q_2 + Q_1)/(Q_3 - Q_1)$.

Results

Central tendency. There are two issues to consider: Which measure shows the smallest standard deviation across different

Table 2
Examples of Distribution Statistics with Cutoffs and Outliers

	No outlier, no cutoff		No outlier, 1,000-ms cutoff		.1 outlier, no cutoff		.1 outlier, 1,000-ms cutoff	
Statistic	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
M	700.3	6.1	654.6	4.5	799.5	11.6	657.3	5.0
Mdn	643.2	5.8	626.9	6.3	666.3	8.0	630.0	6.5
Har *M*	656.3	4.5	631.0	4.2	694.7	6.0	633.3	4.5
Trim *M*	648.7	4.8	625.6	4.6	686.8	7.4	628.2	4.9
SD	204.4	8.4	129.7	2.5	401.8	19.2	130.8	2.8
M dev	150.3	5.4	106.1	2.3	261.7	13.5	107.4	2.7
Quart dev	111.1	4.8	91.0	3.4	140.7	7.1	92.9	4.0
Skew	1.88	0.23	0.68	0.05	2.63	0.13	0.65	0.05
Pearson 2	0.84	0.05	0.64	0.07	0.99	0.04	0.62	0.08
Quart skew	0.249	0.046	0.203	0.045	0.306	0.045	0.195	0.049
Kurtosis	8.47	2.21	2.71	0.10	10.64	1.07	2.64	0.11

Note. Har *M* = harmonic mean; trim *M* = trimmed mean; *M* dev = mean deviation; quart dev = quartile deviation; Pearson 2 = Pearson's second measure of skewness; quart skew = quartile skewness.

samples from the same underlying distribution, and which measure is least influenced by outliers and methods for eliminating outliers. Over all the conditions (with and without outliers, with and without cutoffs), the harmonic mean appears to have the smallest standard deviation. (Note that the median, harmonic mean, and trimmed mean do not estimate the mean of the distribution; they estimate the distribution's median, harmonic mean, and trimmed mean, respectively.) The median has a greater standard deviation than the harmonic mean and trimmed mean. The mean has a greater standard deviation than the other measures when there is no cutoff, and when there is a 1,000-ms cutoff (eliminating about 8% of the observations when there are no outliers and about 16% of the observations when there are outliers), the standard deviation for the mean is still higher than the standard deviation of the harmonic and trimmed means, although less than that of the median. The fact that the harmonic mean has the smallest standard deviation accords well with the results just presented that showed relatively high power for the inverse transformation.

Although the harmonic mean shows the smallest standard deviation, it is not the measure least influenced by outliers. It is more affected by outliers and cutoffs than the median (with the mean being most affected). So if outliers are not a problem, that is, they are consistent across conditions and a consistent cutoff value is used (either none or some fixed value), the harmonic mean is the best choice to measure central tendency for statistical analysis. However, if it is possible that the probability of outliers is different across conditions, then medians will be more stable. Questions about which measure to use can be empirically answered using Monte Carlo studies for the particular set of conditions under examination with appropriate assumptions about distribution shape and the presence or absence of outliers.

Spread. It is clear from the results in Table 2 that the quartile deviation is the most resistant to the effects of both cutoffs and outliers and that it has the smallest standard deviation. The results also show that the mean deviation and standard deviation are critically dependent on the precise location of the cutoff. This is especially problematic if the distribution means change across experimental conditions; in this case, the position of a cutoff value in relation to the means would change across conditions. For example, for two distributions with the same spread but different means, for a single fixed cutoff, estimates of the standard deviation and mean deviation might show significant differences between conditions. Thus, it would be most useful when reporting standard deviations to also report quartile deviations to confirm the trends.

Shape. Outliers and the cutoffs used to eliminate them affected quartile skewness and Pearson's 2 skewness by only a factor of about 0.5, though quartile skewness has a larger standard deviation than Pearson's measure. Thus, they are better measures of shape than those derived from moments: Skewness and kurtosis are both affected by cutoffs and outliers; in particular, introduction of a cutoff that eliminates 8% of the data reduced skewness and kurtosis by as much as a factor of four or five. Thus as in the discussion of spread, the skewness and kurtosis measures will be sensitive to the cutoff used and changes in these estimates across conditions may be due to changes in the location of the distribution in relation to the cutoff. For measuring shape, moments should not be used at all unless the sample sizes are in the tens of thousands and the extreme tails of the distributions are of interest.

Linearity under transformations. One potentially major problem in using measures such as the trimmed mean, harmonic mean, and median is that trends that are linear in the mean may not be linear in these measures. The important issue is the practical effect of these transformations on linearity. To address this issue, two simulations were conducted for an experiment with three conditions, each successive condition representing a linear increase in mean reaction time of 100 ms over the last condition.

For the first simulation, 1,000 observations were obtained from three ex-Gaussian distributions each with parameters σ = 40 ms and τ = 200 ms, one with μ = 500, one with μ = 600, and one with μ = 700 ms. The mean, median, harmonic mean, and trimmed mean (with 10% of the longest reaction times eliminated) were calculated, the process was repeated 100 times, and the averages of these measures were obtained over the 100 replications. Plotting the value of each measure for the three experimental conditions showed a roughly linear increase, with 100-ms differences between conditions. There were only slight deviations from linearity, all less than 3 ms (i.e., none of the three points lay more than 3 ms away from the straight line best fit of the three points). This result would be expected because the distribution was simply shifted by the values of μ.

In a second simulation, τ was varied from 200 ms to 300 ms to 400 ms, with μ fixed at 500 ms and σ fixed at 40 ms. The results were exactly the same as when μ was varied, with the exception of a scale change. In calculations of the mean, the differences between experimental conditions were linear, with increments of 100 ms. For the median, the differences were 68 ms with linearity; for the harmonic mean, the differences were 62 ms with linearity; and for the trimmed mean, the differences were 74 ms with linearity. The differences between conditions were all within 3 ms of linearity. Note that the estimated differences for the trimmed mean and harmonic mean are different from the estimated increases in the regular mean (100 ms) when the increase in the mean results from varying τ, that is, skewing of the tail of the distribution. Thus, when working with models, use of these alternatives to the regular mean would require predictions based on these alternate measures, not predictions based on mean reaction time (e.g., a 75-ms linear increase in the trimmed mean from a 100-ms effect in the mean of the underlying process).

What the Statistics Measure

When using measures of location, dispersion, and shape, it is important to understand precisely what aspects of the distribution are being measured. For measures of location this is clear, but for the measures of dispersion and shape, we want to know what parts of the distribution contribute to the estimates and influence them most. Skewness, kurtosis, and variance are based on the moments of a distribution. Ratcliff (1979) summarized arguments showing that estimates of moments depend on response times from the extreme tails of a distribution (see

Pearson, 1963). The extreme tails are not the most interesting aspect of a distribution from both theoretical and empirical points of view, because many observations (tens of thousands, for example) are needed to provide adequate information about the extreme tails and because the moments have high standard deviations and are sensitive to outliers (as previously discussed, see Table 1).

Two alternatives to moments for the measurement of shape were suggested by Ratcliff (1979). One was robust statistics that measure aspects of distributions that are represented by the middle 90% of the observations. The second was to fit explicit functions to the distributions (this is considered further in the next section). Robust measures were championed by Tukey (1977) and by Mosteller and Tukey (1977), and since then there have been many books written about them, and they have been incorporated into many statistical data analysis packages. The measures that have the best credentials as robust measures (though not necessarily included in the usual list of robust measures) are the median, quartile deviation, and the Pearson 2 measure of skewness. The results from Table 2 support this conclusion.

In summary, it is possible to make general recommendations both for untrimmed data and for data with responses eliminated that are slower than a cutoff value. For the untrimmed data, the harmonic mean, quartile deviation, and Pearson's skewness or quartile skewness appear to provide the best compromises in terms of insensitivity to outliers and their small standard deviations. These statistics are also likely to be relatively invariant if cutoffs are used. If the standard deviation or other moment-based estimators are to be used in conjunction with cutoffs, and the distributions of experimental conditions are shifted in relation to each other, then it is necessary to confirm any differences obtained with the moment estimators by using alternative estimators.

Results for Correlations Among Measures

Correlations among the different measures of location, dispersion, and shape provide another way to understand what they are measuring. Correlations can make apparent dependencies among the different measures, and show whether different random samples that produce different estimates of one measure also produce systematic changes in another measure. For example, a single long outlier reaction time would be expected to increase the estimates of mean, standard deviation, and skewness, leading to a positive correlation among them. In contrast, a combination of long and short outlier reaction times that increase standard deviation may not affect skewness, and so the correlation between these two measures may be low. In these examples, a number of samples are generated with the same parameter values and the different measures of location, dispersion, and shape are calculated for each sample. These measures are correlated across samples.

Tables 3 and 4 show examples of these correlations for distributions of reaction times with no outliers, both with no response times eliminated and with all response times longer than 1,000 ms eliminated. For each measure, Figure 9 shows the scatter plot of the means for that measure from each of the 100 replications of the simulated distribution (ex-Gaussian with μ = 500 ms, σ = 40 ms, and τ = 200 ms).

These results present a sampling of the kinds of observations that can be derived from examination of the correlations and scatter plots. First, the mean is highly correlated with other measures of central tendency, less with the median than the others. When a cutoff is used, the correlation rises because the measures are less dependent on variable long reaction times. The high correlations show that the three means are essentially measuring the same thing. Second, the three means are positively correlated with mean deviation and standard deviation, but the relationship is by no means strong, as Figure 9 shows. Third, the three means are slightly negatively correlated with the measurements of shape. The larger the mean, the more likely the distribution is to be less skewed, though again, the relation is small. Fourth, the median is negatively correlated with the Pearson 2 and quartile skewness measures. This is because the median enters these expressions as a negative quantity, and as with the mean, the larger the median the less likely the distribution is to be skewed.

For the measures of scale or dispersion, the standard deviation and mean deviation are highly correlated and are essentially measuring the same thing, whereas they are somewhat less correlated with quartile deviation. The shape measures correlate with each other to some extent: Pearson's 2 measure of skewness and quartile skewness measure roughly the same thing, and there is only a weak relation between each of these two measures and skewness measured from the third moment. This shows the influence of extreme values where an extreme value will have a large effect on skewness but little effect on quartile skewness and Pearson's 2 measure of skewness. The three measures correlate negatively with skewness and kurtosis when the cutoff is used because standard deviation enters the denominator of these terms (this effect is less than the effect of extreme values with no cutoff).

Recommendations

The statistic of choice for comparing distributions and testing hypotheses about differences between distributions is one for which its estimate has a small standard deviation and shows stability under conditions of outliers and cutoffs to eliminate outliers. On the other hand, if the true value of a characteristic of a distribution (e.g., central tendency or spread) is required, then the chosen statistic should be the one least sensitive to the effects of cutoffs and outliers.

The numerical examples presented in this section are not generally applicable to other domains: They are specific to distributions like reaction time distributions, which are skewed to the right with outliers to the right. In the reaction time domain, it appears that the often used median is not a particularly good measure of central tendency unless the true center (median) of the distribution is required because it has relatively high variability compared with, for example, the harmonic mean (cf. the power results in the first section). The quartile deviation is a reasonable alternative to the ubiquitous standard deviation and both are useful. For the measures of shape, the measures based on moments are too variable and measure uninteresting parts

Table 3
Correlations Between the Different Measures of Location, Spread, and Shape for the Case of No Outliers and No Cutoff

Statistic	1	2	3	4	5	6	7	8	9	10	11
1. *M*	1.00	0.60	0.93	0.92	0.70	0.78	0.57	0.00	0.24	0.18	−0.04
2. *Mdn*	0.60	1.00	0.77	0.76	0.03	0.07	0.22	−0.15	−0.58	−0.53	−0.11
3. Harm *M*	0.93	0.77	1.00	0.98	0.42	0.52	0.43	−0.09	0.05	0.03	−0.10
4. Trim *M*	0.92	0.76	0.98	1.00	0.39	0.54	0.55	−0.16	0.06	0.07	−0.14
5. *SD*	0.69	0.03	0.42	0.39	1.00	0.89	0.46	0.46	0.44	0.37	0.35
6. *M* dev	0.78	0.07	0.52	0.54	0.89	1.00	0.70	0.07	0.58	0.48	−0.00
7. Quart dev	0.57	0.22	0.43	0.54	0.45	0.69	1.00	−0.18	0.29	0.37	−0.14
8. Skew	0.00	−0.15	−0.09	−0.16	0.46	0.07	−0.18	1.00	−0.04	0.07	0.96
9. Pearson 2	0.24	−0.58	0.05	0.06	0.44	0.58	0.29	−0.04	1.00	0.89	−0.10
10. Quart skew	0.18	−0.53	0.03	0.07	0.37	0.48	0.38	0.07	0.89	1.00	0.04
11. Kurtosis	−0.04	−0.11	−0.10	−0.14	0.35	−0.00	−0.15	0.96	−0.10	0.05	1.00

Note. Har *M* = harmonic mean; trim *M* = trimmed mean; *M* dev = mean deviation; quart dev = quartile deviation; Pearson 2 = Pearson's second measure of skewness; quart skew = quartile skewness.

of the distribution (especially when outliers are possible). Thus, Pearson's 2 measure is the most useful measure of skew of the distribution.

Fitting Explicit Distributions

The first two sections of this article were concerned with issues that could be addressed without knowledge of the precise form of the reaction time distribution (although explicit distributions were used for the simulations). Even the statistics describing distribution location, spread, and shape were not focused on the form of the distribution. In the third section, the specific form of the distribution becomes important; the mean and standard deviation of an empirical distribution of response times are recovered by fitting a model of the distribution to the data and using the parameters of the model as estimates of the empirical parameters.

Methods for Curve Fitting

Empirical distributions of response times can be evaluated and compared with each other by fitting explicit theoretical functions to them. The theoretical functions also can be used to estimate statistics of the empirical distributions by deriving estimates from the parameters of the theoretical functions. The usual way of fitting a distribution to data is the maximum likelihood method. For each data point, the probability of that point occurring given the specific theoretical function is calculated (by finding $f(t)$ given a reaction time t), and the probabilities for all the data points are multiplied. The parameters of the theoretical function are then adjusted to find the maximum of this product (or, in many computer coded routines, the minimum of the product's negative value). In Mathematica, the program can be as small as three lines of code but it is very slow; for fitting one set of data this might be acceptable, but for multiple simulations, much faster FORTRAN or C code (seconds or less) is required.

The maximum likelihood method of estimation has a number of nice properties; for example, the parameter estimates it yields have the lowest asymptotic standard deviations of any parameter estimates (i.e., maximum likelihood is the most accurate estimation method) and the asymptotic distribution of estimates for a given parameter is normal, so that standard z tests

Table 4
Correlations Between the Different Measures of Location, Spread, and Shape for the Case of No Outliers and a 1,000-ms Cutoff

Statistic	1	2	3	4	5	6	7	8	9	10	11
1. *M*	1.00	0.86	0.98	0.99	0.42	0.49	0.49	−0.67	−0.32	−0.18	−0.65
2. *Mdn*	0.86	1.00	0.87	0.88	0.17	0.23	0.37	−0.76	−0.76	−0.61	−0.58
3. Harm *M*	0.98	0.87	1.00	0.99	0.25	0.33	0.40	−0.66	−0.34	−0.20	−0.59
4. Trim *M*	0.99	0.88	0.99	1.00	0.30	0.40	0.48	−0.74	−0.37	−0.21	−0.68
5. *SD*	0.42	0.17	0.25	0.31	1.00	0.95	0.58	−0.14	0.09	0.06	−0.44
6. *M* dev	0.49	0.24	0.33	0.40	0.95	1.00	0.75	−0.32	0.07	0.08	−0.64
7. Quart dev	0.50	0.38	0.40	0.48	0.58	0.75	1.00	−0.56	−0.13	−0.02	−0.78
8. Skew	−0.67	−0.76	−0.66	−0.74	−0.13	−0.32	−0.55	1.00	0.55	0.36	0.84
9. Pearson 2	−0.33	−0.76	−0.35	−0.37	0.09	0.07	−0.13	0.56	1.00	0.90	0.28
10. Quart skew	−0.19	−0.61	−0.20	−0.21	0.06	0.08	−0.02	0.37	0.90	1.00	0.13
11. Kurtosis	−0.65	−0.58	−0.59	−0.68	−0.44	−0.64	−0.77	0.84	0.27	0.13	1.00

Note. Har *M* = harmonic mean; trim *M* = trimmed mean; *M* dev = mean deviation; quart dev = quartile deviation; Pearson 2 = Pearson's second measure of skewness; quart skew = quartile skewness.

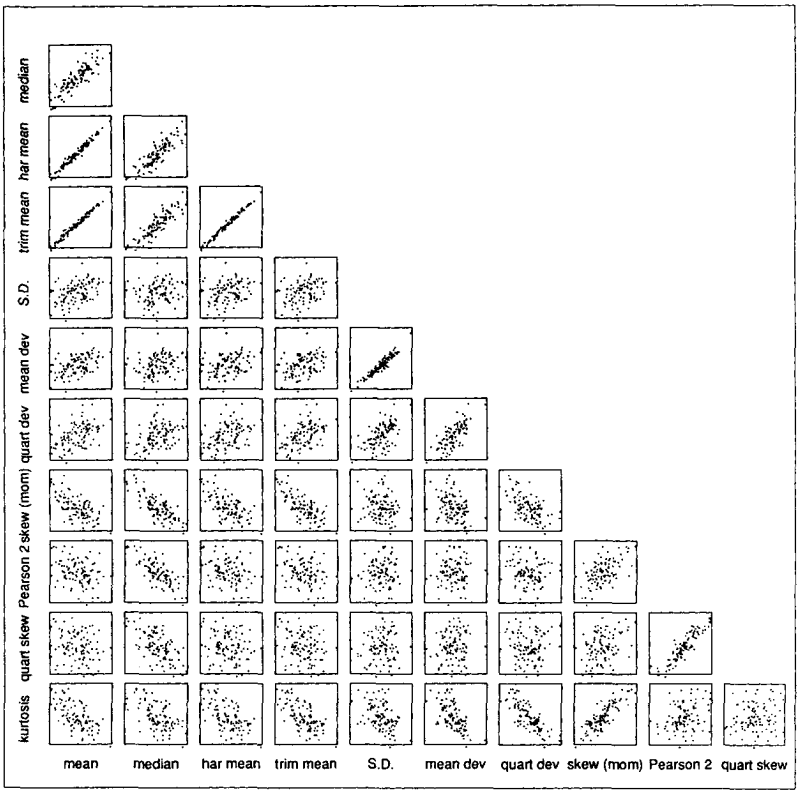

Figure 9. Scatter plots of the values for the measures of location, spread, and shape for the 100 simulated distributions. (Har mean = harmonic mean; trim mean = trimmed mean; mean dev = mean deviation; quart dev = quartile deviation; mom = moment; Pearson 2 = Pearson's second measure of skewness; quart skew = quartile skewness.)

can be used to evaluate differences among the values of a given parameter for different samples. Both of these properties hold for asymptotic or large sample sizes (see Hogg & Craig, 1965).

For a specific theoretical function and an empirical distribution of response times, the maximum likelihood procedure yields the parameters of the function that best describe the empirical distribution. However, this is not enough: The standard deviations of these parameter estimates may also be needed for comparisons among experimental conditions. There exist theoretical methods for finding these standard deviations, but they can be tedious and numerically intense (see, e.g., Ratcliff & Murdock, 1976). Fortunately, there are also computationally simple Monte Carlo methods (see, e.g., Press, Flannery, Teukolsky, & Vetterling, 1986, Chapter 14). The Monte Carlo methods use the theoretical function with its parameters set to give the best fit of the function to the data. From the function and these parameters, random numbers are generated; these numbers have the same statistical properties as the original data. Then the theoretical function is fit to the new pseudo data, yielding a new, slightly different set of parameter estimates. This process of generating pseudo data and refitting the model is repeated a number of times, each time yielding a set of parameter estimates. Then, from these sets of parameter estimates, the standard deviation in a particular parameter can be estimated. The covariances among the parameters can also be estimated, and these can be important for understanding the shape of the parameter space: For example, if one parameter has a higher than average value in a particular fit of the function to data, then another parameter might be lower than average to compensate. If the two parameters show such a pattern consistently, then the covariance or correlation between them will be negative.

As an aside, it should be mentioned that Monte Carlo methods can be used to estimate the power needed to discriminate between different candidate functions for describing a distribution of reaction times. For example, for a certain sample size, pseudo data can be generated from one of the two competing functions and both functions fitted to the data. This can be repeated many times for the first and then the second function. For 1,000 replications, it is possible to see how many times Function 1 fits better than Function 2 on Function 1's data and vice versa. This leads to a straightforward power estimate.

Recovering the Mean and Standard Deviation of a Distribution

Fitting a theoretical function to an empirical response time distribution can yield information about the functional charac-

teristics of the distribution (see Ratcliff, 1979; Ratcliff & Murdock, 1976), and it can also provide estimates of the mean and standard deviation of the empirical distribution. With simulations, the accuracy of the recovered estimates of mean and standard deviation generated from a theoretical function can be evaluated. To do this, a response time distribution can be generated from a specific, known theoretical function, so that the true mean and standard deviation of that distribution are known. Then one or more theoretical functions can be fit to the generated distribution, and the parameters of the fitted functions can be used to estimate the mean and standard deviation, and these estimates can be compared with the true values. The comparison can be done under several conditions: The response time distribution can be generated with and without outliers, and the function can be fit either to all the response times or to only those response times faster than some cutoff value.

When a theoretical distribution function is fit to real reaction time data, some of the slow reaction times are usually eliminated before fitting to eliminate outliers. For example, Ratcliff (1979) used a cutoff value to eliminate slow reaction times, turning the distribution of response times into a truncated distribution. Ratcliff (1979) fit this truncated distribution with a function that assumed a complete, untruncated distribution. However, it is easy to fit a truncated distribution to the truncated data (this method is also used by Ulrich & Miller, 1992). For example, if the distribution is $f(t)$ and has a range from zero to "cut," then the truncated distribution has a density function:

$$f(t)/\int_0^{\text{cut}} f(t')dt',$$

that is, the untruncated density function divided (normalized) by the density remaining below the cutoff. This modified density function can be used to estimate parameter values in the same way as the complete density function would be used.

The first goal of fitting theoretical functions to data is to estimate the parameters of the function that best describe the data. In attempting to accomplish this with truncated distributions, there are two issues. The first is whether fitting the truncated distribution to truncated data allows recovery of the parameter values of the untruncated distribution, and the second is what is the effect of outliers on estimation of the parameters as the truncation value is reduced. It might be that the estimates follow the parameters of the contaminated distribution or it might be that they recover the parameters of the untruncated distribution with greater accuracy as more and more of the tail (and hence outliers) is eliminated. It may also be that as the cutoff is reduced, the estimates converge on the population values (i.e., become more unbiased) but at the same time, become more and more variable.

To address these issues, simulation studies examined parameter estimates for truncated distributions, generated from both ex-Gaussian and inverse Gaussian functions, with and without outliers. The truncation point was varied and recovery of the mean and standard deviation of the original distribution was examined. For the ex-Gaussian, three experimental conditions were simulated, each one varying from the next by 50-ms increments in τ. For the inverse Gaussian, two conditions were simulated, differing in the λ parameter. The two distribution functions were used so that the effect of fitting the wrong distribution function to the data could be examined. In other words, the idea was to generate data from one of the distribution functions (ex-Gaussian or inverse Gaussian) and fit the data with both truncated distribution functions to investigate the effect of getting the exact form of the distribution wrong.

Note that this method of recovering means and standard deviations from fitting distributions assumes that the distributions fit adequately. If the theoretical distributions do not fit, then the computed means and standard deviations may be meaningless. In the cases presented here, this is not a problem because the ex-Gaussian and the inverse Gaussian do mimic each other reasonably well.

Results

Tables 5, 6, 7, and 8 show the results for distributions with and without outliers generated from the ex-Gaussian and the inverse Gaussian functions fitted by the truncated ex-Gaussian and truncated inverse Gaussian. The outlier assumption was the same as in earlier simulations: For 10% of responses, a random time between 0 and 2,000 ms was added to the time from the parent distribution. Figure 10 shows that reducing the cutoff value for outliers eliminates more and more outliers until the 1,000-ms cutoff, at which point the parent distribution begins to dominate the outlier distribution. Thus, we should expect to see estimates of the mean and standard deviation converge to their true values (the values for the distributions that were used to generate the simulated data) at the lower cutoffs. Each mean and standard deviation in the tables is based on 100 simulations, each using 1,000 simulated reaction times. In the discussion of the results that follow, I examine bias in the estimates of the means and standard deviations first and then discuss variability in those estimates.

For the distributions without outliers, the results in the tables show that the mean and standard deviation of the parent distribution are recovered quite well from the fitted functions at all cutoffs. For data generated from the ex-Gaussian, fitting the truncated ex-Gaussian recovers the mean and standard deviation of the ex-Gaussian accurately, whereas fitting the truncated inverse Gaussian recovers the mean and standard deviation of the ex-Gaussian reasonably well except at the 1,000-ms cutoff where the inverse Gaussian underestimates both the mean and standard deviation (Table 5). The effect is symmetric: The truncated ex-Gaussian and inverse Gaussian recover the mean and standard deviation accurately for data generated from the inverse Gaussian except that the ex-Gaussian overestimates the mean and standard deviation at the 1,000-ms cutoff (Table 6).

For the distributions that include 10% outliers, the recovered estimates for the mean and standard deviation at the shortest cutoff are reasonably close to those of the uncontaminated parent distribution (except for the over- and underestimation when the ex-Gaussian is fitted by the inverse Gaussian and vice versa). At the longer cutoffs, the mean and standard deviations are larger than the uncontaminated distribution values and decrease as the cutoff is reduced.

For the ex-Gaussian simulations, three different parent distributions were used, differing from each other in steps of 50 ms

Table 5
Means and Standard Deviations Recovered From the Ex-Gaussian

Generating distribution ExG			Fitting distribution	Theoretical value		Cutoff							
						2,500 ms		2,000 ms		1,500 ms		1,000 ms	
μ	σ	τ		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
400	50	200	ExG	600	206	600	200	597	206	600	207	601	207
400	50	250	ExG	650	255	650	248	647	254	649	255	650	255
400	50	300	ExG	700	304	699	297	697	304	699	305	705	310
400	50	200	IG	600	205	601	199	600	199	599	196	586	179
400	50	250	IG	650	255	651	247	650	247	647	243	630	219
400	50	300	IG	700	304	701	299	700	299	697	293	667	254

Note. All times are in ms. ExG = ex-Gaussian; IG = inverse Gaussian. The number of observations eliminated at the 2,500-ms cutoff was less than 0.1%, the percentage eliminated at the 1,000-ms cutoff was 5% for τ = 200, 10% for τ = 250, and 15% for τ = 300. The standard error (across replications) in the mean calculated from the distribution parameters was about twice the standard error in the untrimmed mean at the 1,000-ms cutoff (the two were about the same at the 2,500-ms cutoff). The values ranged from 6 to 10 for τ = 200, from 7 to 14 for τ = 250, and from 10 to 25 for τ = 300. For the standard error in the standard deviation, the values ranged from 7 to 12 for τ = 200, from 8 to 19 for τ = 250, and from 10 to 27 for τ = 300.

in τ. The results from the fit of the truncated ex-Gaussian show that these differences in means were captured reasonably accurately. The differences down the columns in Table 7 for the fitted distributions with a 1,000-cutoff are 56 ms and 50 ms, corresponding well to the theoretical difference of 50 ms in τ.

As an aside, the mean recovered from fitting the distribution tends to the mean of the parent distribution, and the mean computed from the raw response times underestimates the true mean by as much as 100 ms at the 1,000-ms cutoff (approximately 30 ms, 60 ms, and 90 ms for τ = 200 ms, 250 ms, 300 ms, respectively). Thus, fitting the truncated distributions does more than just eliminate the long reaction times, it allows the mean of the parent distribution to be more accurately estimated.

The variability in the estimates of the mean and standard deviation can be compared with the expected standard deviation in the mean and standard deviation of the untrimmed distribution with no outliers. The variability in the estimates would have to be at least as large as this expected standard deviation. The expected standard deviation for both is approximately $\tau/\sqrt{N}$, where N is the number of observations in the distribution (this is because the influence of variability in the normal distribution is small because the standard deviation for τ = 200 ms and σ = 50 ms is $\sqrt{(\sigma^2 + \tau^2)}$ = 206.2). For the simulations presented in Tables 5 through 8, where N = 1,000, the theoretical values of the standard deviation in the mean are 6.5 ms, 8.1 ms, and 9.7 ms for τ values of 200 ms, 250 ms, and 300 ms, respectively, and these can be considered lower bounds on the estimates of the standard deviations in the mean and in the standard deviation of the sample distribution. Standard deviations in the means and standard deviations of the distributions were calculated from the simulations (over the 100 replications) and are shown in the footnotes to the tables. The estimates are roughly the same as the lower bound theoretical values for the 2,500-ms cutoff but rise to 2 to 3 times the theoretical values at the 1,000-ms cutoff. This shows that the lower the cutoff, the more observations are excluded and the more variable are the estimates of the mean and standard deviation. So, even though the estimates are approximately unbiased, that is, the average of the 100 replications closely matches the mean and standard

Table 6
Means and Standard Deviations Recovered From the Inverse Gaussian

Generating distribution IG			Fitting distribution	Theoretical value		Cutoff							
						2,500 ms		2,000 ms		1,500 ms		1,000 ms	
θ	λ	T_{er}		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
400	1,000	200	ExG	600	253	571	261	601	259	602	260	618	280
400	2,000	200	ExG	600	179	599	189	591	199	601	191	618	212
400	1,000	200	IG	600	253	601	252	601	254	602	256	605	261
400	2,000	200	IG	600	179	599	179	600	178	600	179	600	179

Note. All times are in ms. ExG = ex-Gaussian; IG = inverse Gaussian. At the 2,500-ms cutoff, about 3% of data are cut out and at the 1,000-ms cutoff, about 7% of the data are removed. Standard errors in the means rise as a function of cutoff: For λ = 1,000, the standard error ranges from 8 to 23 starting with slightly higher values for the fit of the ExG and for λ = 2,000, the range is 6 to 12. For the standard deviation, the patterns are the same but the ranges are from 11 to 31 and 6 to 17, respectively.

Table 7
Means and Standard Deviations Recovered From the Ex-Gaussian With 10% Noise

Generating distribution ExG			Fitting distribution	Theoretical value no outliers		Cutoff 2,500 ms		2,000 ms		1,500 ms		1,000 ms	
μ	σ	τ		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
400	50	200	ExG	600	206	689	304	636	271	623	238	609	217
400	50	250	ExG	650	255	736	345	690	316	675	285	665	271
400	50	300	ExG	700	304	784	390	745	363	727	333	715	318
400	50	200	IG	600	205	693	343	653	275	624	229	595	188
400	50	250	IG	650	255	745	390	707	327	675	277	633	220
400	50	300	IG	700	304	800	452	762	387	728	335	681	270

Note. All times are in ms. ExG = ex-Gaussian; IG = inverse Gaussian. The number of observations eliminated at the 2,500-ms cutoff was about 1%, the percentage eliminated at the 1,000-ms cutoff was 13% for τ = 200, 17% for τ = 250, and 20% for τ = 300. The standard error (across replications) in the mean calculated from the distribution parameters was about twice the standard error in the untrimmed mean at the 1,000-ms cutoff (the two were about the same at the 2,500-ms cutoff). The values ranged from 8 to 17 for τ = 200, from 12 to 20 for τ = 250, and from 15 to 29 for τ = 300. For the standard error in the standard deviation, the values ranged from 13 to 23 for τ = 200, from 12 to 34 for τ = 250, and from 15 to 36 for τ = 300. The standard errors showed U-shaped functions with the long and short cutoff having larger standard errors.

deviation of the parent distribution at low cutoffs (or approaches them if the distribution includes outliers), the estimates increase in variability at low cutoffs.

This variability in the recovered mean and standard deviation comes from two sources: One is the different means and standard deviations that sample distributions generated from the same parent distribution will have and the other is the variability in the estimates that results from fitting a function to the sample distribution. From the simulations it is possible to determine how accurately fitting a truncated function recovers the mean of that particular sample. For the results in Tables 5–8, 100 replications for each set of parameter values were performed, and for each of these 100 samples, the mean can be calculated. Using these means and the means estimated from the parameters of the best fitting functions, it is possible to estimate the standard deviation in the difference between the two values and the correlation between the two. So, for example, in Table 5 for the ex-Gaussian fitting the ex-Gaussian with parameters 400, 50, and 200 with no outliers, for the 2,500-ms cutoff, the standard deviation in the difference between the mean recovered from fitting the function and the sample mean is 0.5 ms (compared with a standard deviation of 6 ms in the recovered mean) and the correlation is 0.995. This means that the mean of the sample and the recovered mean are in close relative agreement (when one is higher than the population value, so is the other). For the 1,000-ms cutoff, the standard error in the difference between the sample mean and recovered mean is 6 ms (compared with a standard error of 10 ms in the recovered mean) and the correlation between the sample and the recovered mean across replications is 0.91. For the parameter values 400, 50, and 300 with a 0.1 probability of outliers (Table 7), for the 2,500-ms cutoff, the standard error in the difference between sample and recovered means is 2 ms and the correlation is 0.996 (compared with a standard error of 15 ms in the recovered mean). For the 1,000-ms cutoff, the standard error in the difference between sample and recovered means is 26 ms (compared with a standard error of 29 ms in the recovered mean) and a correlation of 0.84. This change to a low correlation is sharp because at the 1,500-ms cutoff, the standard error in the difference is 8 ms (compared with 16 ms in the recovered mean) and the correlation is 0.96.

Table 8
Means and Standard Deviations Recovered From the Inverse Gaussian With 10% Outliers

Generating distribution IG			Fitting distribution	Theoretical value		Cutoff 2,500 ms		2,000 ms		1,500 ms		1,000 ms	
θ	λ	T_{er}		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
400	1,000	200	ExG	600	253	662	353	656	320	632	295	633	297
400	2,000	200	ExG	600	179	690	295	650	255	626	221	626	220
400	1,000	200	IG	600	253	701	422	660	345	634	301	620	280
400	2,000	200	IG	600	179	695	328	655	260	625	210	612	190

Note. All times are in ms. ExG = ex-Gaussian, IG = inverse Gaussian. At the 2,500-ms cutoff, about 6% of data are cut out and at the 1,000-ms cutoff, about 13% of the data are removed. Standard errors in the means fall and rise: For λ = 1,000, the standard error ranges from 10 to 23 starting at 15 and for λ = 2,000, the range is 9 to 15 starting at 13. For the standard deviation, the patterns are the same but the ranges are from 10 to 38 and 9 to 17, respectively.

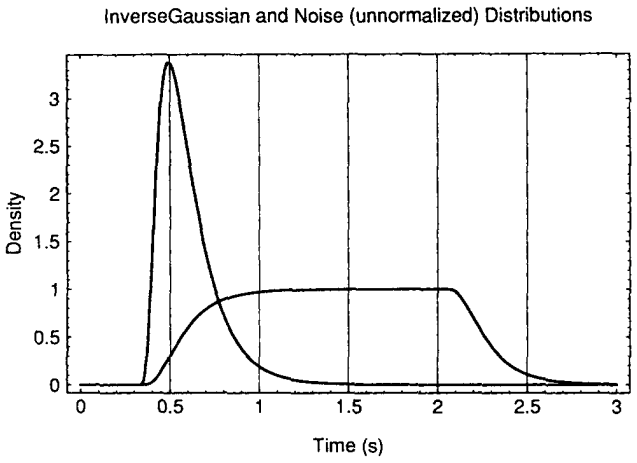

Figure 10. An inverse-Gaussian distribution with the noise distribution derived by shifting the inverse-Gaussian by a random time between 0 and 2,000 ms.

Thus, when the distributions did not include outliers, fitting the truncated distribution recovered the mean and standard deviation of the sample quite accurately (i.e., unbiased), to within a millisecond or two in these simulations, but at the shorter cutoffs, the standard deviation in the estimates increased and the standard deviation in the difference between the sample mean and the recovered mean approached the overall standard deviation in the mean. The same outcomes apply when the distributions include outliers.

These conclusions indicate that recovering means and standard deviations by fitting explicit theoretical functions will be useful when there are either few outliers or the outliers reside in the extreme tail. With either few outliers or most of the outliers in the extreme tail, a relatively long cutoff will not leave so many outlier response times in the truncated distribution, so that estimates of the mean and standard deviation will be relatively accurate. In addition, the standard deviation in the estimates will be low. If there are relatively few outliers, the means and standard deviations will be near the population values as a function of cutoff as shown in Tables 5 and 6, and if the outliers are mainly in the extreme tails, the means and standard deviations will converge to the population values as shown in Tables 7 and 8. Of course, recovery of the mean and standard deviation is not the only product of such fitting: The fitted distribution provides an estimate of distribution shape through the distribution parameters (see Ratcliff, 1979; Ratcliff & Murdock, 1976).

Note the scope of what I have examined here. I have been primarily concerned with recovery of the mean and standard deviation of the distribution. If measures of distribution shape are also required, then fitting a distribution such as the ex-Gaussian or the inverse Gaussian provides information that may be more valuable than that provided by other measures (e.g., skewness). Fitting a distribution provides estimates of the parameters of the model and the residuals (the differences between the reaction time distribution and the fitted distribution as a function of time). This allows the use of the distribution parameters as descriptions of distribution shape, and it also allows deviations from the theoretical distribution to be used if they are substantial or needed theoretically (e.g., a model may be fitted and its deviations checked against the deviations from the fit of the summary distribution). Note that Ulrich and Miller (1992) have performed similar studies and their conclusions should also be examined.

Recommendation

The mean and standard deviation of a distribution contaminated with outliers can be recovered from fitting a theoretical function to the data if there are not too many outliers, if most of the outliers are in the extreme right tail of the distribution, or both. This method can only be used with any accuracy when there are several hundred or more observations and should be used when the shape of the distribution is also under evaluation. By reducing the cutoff of the distribution and by examining the estimates of the mean and standard deviation as a function of cutoff, the parameters can be examined to see whether they converge on single values; convergence would support those values as estimates of the population mean and standard deviation. If the distribution is not a perfectly accurate model of the empirical distribution, systematic biases can occur, but Monte Carlo studies can be used to identify what these are likely to be. If the fitted distribution cannot fit the data adequately, then it should not be used.

Other Considerations

Because this article has been concerned with practical issues about the use of reaction time data, it is worth commenting on some other potential problems, misinterpretations, and misunderstandings that are sometimes found in reports of reaction time data. None of the points that follow are original, but some of them have not been explicitly discussed in the reaction time literature (although others are presented in a classic article by Pachella, 1974).

The first issue concerns lack of information about error rates. It is often suggested that "error rates can be ignored because they were so low." This means that error rates are so low that the experimenter has no information about them and so does not know whether they are changing in any significant way across conditions. A small change in error rates near ceiling or floor can mean a change in sensitivity (d') just as great as for larger changes in error rates in the middle of the accuracy range (see also Luce, 1986, p. 240; Pachella, 1974). For example, consider what a change in positive and negative error rates from 1% to 2% would correspond to in terms of d'. The d' value for a 0.99 hit rate and 0.1 false alarm rate is 3.61; the d' value for a 0.98 hit rate is 3.35, a difference of 0.26. To see how large this difference is lower in the d' scale, consider a hit rate of 0.80, much nearer the middle of the accuracy range (with a false alarm rate of 0.2). For this hit rate, d' is 1.68, and to lower d' by 0.26, the hit rate has to fall to 0.72. Thus to argue that differences in error rates under 5% (or 2% in this example) are so small that they can be neglected is equivalent to saying that differences in the 5 to 10% range (or even more) in the middle

of the accuracy range can be ignored. The lesson here is that whenever possible, conditions should be chosen so that error rates are enough above floor that error rate differences are detectable and measurable.

The second issue is scaling effects (see Townsend, 1992) that result when baseline reaction times change as a function of experimental conditions. Reaction time has a minimum but no maximum value. Suppose that baseline reaction times are different for different experimental conditions, but all are above the minimum possible reaction time value, and that each experimental condition slows responses over its baseline. Then requiring subjects to speed up (or allowing them to speed up on their own) will decrease the size of the differences between the experimental conditions and their baselines, as those baselines come closer to the minimum possible reaction time. In fact, when the minimum value possible is reached, there will be no differences among the conditions. Thus, at a baseline of 500 ms, a 40 ms difference between two conditions might be the equivalent of a 70 ms difference at a baseline of 700 ms. A related point concerns the way accuracy grows over time. Plotting a time–accuracy curve with data from a deadline or response signal procedure often shows that the curves for different conditions begin to rise above chance at the same point in time and then later diverge from each other. This means that any reaction time criterion placed across those curves will produce larger differences between them as the criterion is moved toward slower reaction times (see Reed, 1973, 1976; also Ratcliff, 1978). For both of these reasons, comparisons of the differences in reaction times across conditions having different baseline reaction times must take account of scale differences and without examination of scale differences, any conclusions should be suspect.

Third, models that use the difference between reaction times for positive responses and reaction times for negative responses as a measure of the time required to execute a stage of processing are subject to the argument that the reaction time difference is simply a consequence of criteria setting. This issue was debated in the literature, and the conclusion was that the existence of a processing stage could not be demonstrated in the absence of a specific processing model (Proctor, 1986; Proctor & Rao, 1983; Ratcliff, 1985, 1987; Ratcliff & Hacker, 1981).

Fourth, mean response times are not sufficiently constraining to test one model of search processes against another. It sometimes seems that the folklore of cognitive psychology holds that straight line reaction time functions necessarily indicate an underlying serial search mechanism, and that parallel slopes for positive and negative responses indicate an exhaustive (as opposed to self-terminating) search. Townsend (see Townsend, 1990; Townsend & Ashby, 1983) has shown that there is considerable mimicking among models that deal only with mean reaction time. Even with distributions of reaction times, it is possible to develop adequate serial and parallel models that mimic each other. However, some of these models may be bizarre and implausible. The solution to this problem is mainly theoretical: Models are developed to provide a coherent account for a range of different measures (including error rates and time course of processing) across a range of experimental paradigms, or alternative experimental methods are designed in conjunction with the models.

Fifth, it has sometimes been argued that subjects can use differences in processing time to discriminate two conditions. But if the differences are small, then discrimination is not possible. A simple simulation illustrates this, with three ex-Gaussian distributions with μ, σ, and τ, respectively: for Distribution 1, 400, 40, 200; for Distribution 2, 440, 40, 200; and for Distribution 3, 400, 40, 240. The standard deviations of the three distributions are a little larger than 200 ms, and there is a 40-ms difference between the Distribution 1 mean and the means of Distributions 2 and 3. To test for discrimination on the basis of processing time, that is on the basis of some specific response time value, 1,000 observations from each distribution were generated, and the number of times a reaction time from each distribution was greater than a specific value (600 ms) was counted. Results showed that the probability of a value from Distribution 1 greater than 600 ms was 0.366; for Distribution 2, 0.454; and for Distribution 3, 0.426. The largest of these differences would correspond to a difference in d' of less than 0.25, a difference unlikely to be useful in discriminating between conditions. Even if the difference could be discriminated, inspection of the reaction time distributions would show truncated distributions for the two responses, one with no responses lower than the cutoff, and one with no responses higher. There is a caveat to this argument and that is the subjects may be making their decision on the basis of internal timing that does not have some of the variable components of, for example, motor output time. In this case, it would be necessary to present a model of the processing stages and make predictions about the distributions of finishing times for the two decisions and test these against data as well as making the model plausible by not assuming too much accuracy for the internal variables (i.e., the variability has to come from somewhere).

Sixth, sometimes experiments generating what has been called a *micro speed–accuracy trade-off function* have been used as an alternative to experiments that generate a regular, macro speed–accuracy trade-off function, and it has been assumed that the two procedures provide the same information. The micro-trade-off function is obtained by partitioning the reaction times from a single experimental condition into ranges (e.g., from 400 to 500 ms and from 500 to 600 ms) and then examining accuracy within these ranges. Macro-trade-off functions are obtained by varying speed–accuracy instructions (or by using deadline or response signal procedures). Pachella (1974) discussed these two trade-off functions and correctly argued that the two measures are independent: The micro-trade-off provides information about the relative positions of the error and correct reaction time distributions and the macro-trade-off provides information about the growth of accuracy over time.

General Discussion

The aim of the simulations presented here has been to examine the effects of outlier response times on hypothesis testing and on recovering parameters of response time distributions such as location, dispersion, and shape. The transition between the three sections can be viewed as a move from empirical con-

siderations of ANOVA to more detailed empirical plus some theoretical issues about distribution shape to theoretical issues concerning fitting specific distributions to data.

The first section examined the effects of different ways of dealing with outliers on power of ANOVA. Six different methods were examined including transformations, trimming a certain percentage of the responses, trimming according to standard deviations, trimming at cutoff values, and using medians. The conclusions were that there is always an optimal cutoff, but the location of the optimal cutoff depends on the way the distribution shape changes as a function of changes in average reaction time (spreading or shifting). When variability among subject means was low relative to the standard deviations in the distributions, the inverse transformation (1/RT) was always close to the optimal cutoff. When variability among subject means is high, then standard deviation cutoffs have higher power. I recommend that either the inverse transformation or standard deviation cutoffs (depending on variability in subject means) be used to confirm more traditional analyses. It is important to keep in mind that the main aim of such analyses is to provide evidence for replicability of the results, and replication or partial replication is important for theoretically important results.

The second section dealt with the behavior of different statistics that describe reaction time distributions. The main results showed the sensitivity of the different measures to outliers and cutoffs and how related the various measures were to each other. To represent location of the distribution, the harmonic mean and trimmed mean show the smallest deviation across samples from the same underlying distribution, and the median is least influenced by outliers and cutoffs. So if an accurate estimate of the location of the distribution is required and the data might contain outliers or cutoffs are used, then the median is the best choice (note the warning by Miller, 1988: If the number of observations is small and different across conditions, and means across medians are used in data analysis, then biases may lead to significant F or t values where there is no real difference). But for hypothesis testing, variability in the harmonic mean is lower and so it gives more power than the median, confirming the results from the first section. The quartile deviation as a measure of spread and Pearson 2 skewness as a measure of skewness were the most resistant to outliers and changes in cutoff values. The measures based on moments (standard deviation, skewness, and kurtosis) were very sensitive to outliers and cutoffs. Correlations among the statistics showed that the three means (ordinary, trimmed, and harmonic) were highly correlated, the mean and standard deviation were correlated, the skewness and kurtosis were correlated, and the quartile deviation and Pearson 2 skewness were correlated.

The third section examined the use of fitting explicit distributions to recover the parameters of a response time distribution. If the theoretical distribution is an accurate representation of the reaction time distribution, then a range of cutoff values allow recovery of the parameters of the distribution (with varying accuracy) when there are no outliers. However, when there are outliers, recovery is more problematical. If the model is reasonably accurate and most of the outliers are extreme, the mean and standard deviation will converge on the theoretical values as the cutoff is reduced, but if the outliers overlap response times from the real distribution, then overprediction of the means and standard deviations can result. However, trends in the data will be preserved. For this technique to work, however, several hundred or more observations are needed. The use of this method to recover means and standard deviations is of most value when the aim is to test the fit of a specific distribution from a specific model to the data and probably should only be used under these circumstances.

Although it has been assumed here that many or all long reaction times are outliers, not from the process under study, this is by no means certain in all situations. In most experimental studies with subjects tested for one session, most extra long reaction times are probably outliers. In relatively uncontrolled multisession single-subject experiments, many long reaction times are also probably outliers. But for some models, the shape of the tail of the distribution might be critical (see Luce, 1986), and in this circumstance it might be possible to run experiments with motivated subjects using self-reporting to identify spurious long reaction times. Then genuine long reaction times might be used in model evaluation.

Finally, I recommend that if the parameters used in this article are not close to those of the data the reader is studying, then simulations similar to those reported here should be performed with values near those in the data.

References

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data.* New York: Wiley.

Chhikara, R. S., & Folks, J. L. (1989). *The inverse Gaussian distribution.* New York: Dekker.

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

Hockley, W. E. (1982). Retrieval processes in continuous recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 497–512.

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

Hogg, R. V., & Craig, A. T. (1965). *Introduction to mathematical statistics.* New York: Macmillan.

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

Lovie, P. (1986). Identifying outliers. In A. D. Lovie (Ed.), *New developments in statistics for psychology and the social sciences* (pp. 44–69). British Psychological Society: London.

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

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1155–1172.

Miller, J. (1988). A warning about median reaction times. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 539–543.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41–82). New York: Halstead Press.

Pearson, E. S. (1963). Some problems arising in approximating to probability distributions, using moments. *Biometrika, 50,* 95–112.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes.* Cambridge, England: Cambridge University Press.

Proctor, R. W. (1986). Response bias, criteria settings, and the fast-"same" phenomenon: A reply to Ratcliff. *Psychological Review, 93,* 473–477.

Proctor, R. W., & Rao, K. V. (1983). Evidence that the *same–different* disparity is not attributable to response bias. *Perception and Psychophysics, 34,* 72–76.

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

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

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88,* 552–572.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92,* 212–225.

Ratcliff, R. (1987). More on the speed and accuracy of positive and negative responses. *Psychological Review, 94,* 277–280.

Ratcliff, R. (1988a). A note on the mimicking of additive reaction time models. *Journal of Mathematical Psychology, 32,* 192–204.

Ratcliff, R. (1988b). Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review, 95,* 238–255.

Ratcliff, R., & Hacker, M. J. (1981). Speed and accuracy of same and different responses in perceptual matching. *Perception and Psychophysics, 30,* 303–307.

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

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 and Cognition, 4,* 16–30.

Shapiro, S. S., & Wilk, M. B. (1972). An analysis of variance test for the exponential distribution (complete samples). *Technometrics,* 355–370.

Swensson, R. G. (1972). The elusive trade-off: Speed versus accuracy in visual discrimination tasks. *Perception and Psychophysics, 12,* 16–32.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1,* 46–54.

Townsend, J. T. (1992). On the proper scales for reaction time. In H-G. Geisler, S. W. Link, & J. T. Townsend (Eds.), *Cognition, information processing, and psychophysics: Basic issues* (pp. 469–489). Hillsdale, NJ: Erlbaum.

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

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Ulrich, R., & Miller, J. (1992). *Effects of outlier exclusion of reaction time analysis.* Unpublished manuscript.

Received June 16, 1992
Revision received March 22, 1993
Accepted April 5, 1993 ■