Response Time: Data and Theory

Roger Ratcliff

(The Ohio State University)

Abstract: In this chapter, I review a number of related issues that arise from the consideration of response time (RT) distributions. Practical issues include how to represent distribution shape, how to deal with outliers, how to increase the power of ANOVA, and how to average RT distributions. Theoretical issues revolve around the definition of distribution shape, what theoretical distributions can be used to summarize shape, how to fit RT distributions, and how Monte Carlo simulations can be used to quickly explore both practical and theoretical issues. One solution to dealing with RT distributions is to explicitly model the processes involved in performing the task. I present one solution using diffusion models with a description of how the models account for various empirical effects such as speed-accuracy manipulations, biases, fast and slow error responses as well as RT distributions and accuracy.

CA email: ratcliff.22@osu.edu

In the early days of research in cognitive psychology, response time (RT) was used as a measure of the duration of psychological processes and study of this variable was largely separate from the processes and representations that underlie cognition. However, in the last 10 years, much of the field of cognitive psychology has come to realize that it is necessary to model decision processes in a way that integrates RT and accuracy measures. This kind of modeling uses sequential sampling models that have been around for 50 years (LaBerge, 1962; Laming, 1968; Stone, 1960), but what has changed is that solutions to some of the most vexing problems concerning the shape of RT distributions and the relative speed of correct and error responses have been found. The sequential sampling models are now being applied in a number of areas of psychology and produce different interpretations of processing from earlier established views. The models are also being used in domains that use neurophysiological measures and so are being extended outside psychology.

In this chapter, I will begin by a review of issues in analyzing RT including a discussion of RT distributions and how to deal with RT data. Then I will review one sequential sampling model, the diffusion model, and show how it accounts for the effects on accuracy and RT distributions for a number of experimental manipulations.

RT Distributions

There are two main issues of concern when thinking about variability in distributions of response times (RTs). The first issue is theoretical: how to account for the shapes of the distributions of RTs and how the shapes may help constrain and test models. If a theory of processing is proposed, even in a relatively weak form, predictions about the distributions of finishing times of the processes can sometimes be derived and used as tests of the theories. The even stronger claim is that, even if a theory is consistent with data at the level of mean RT, it might be highly inconsistent with the behavior of RT distributions. Generally, when RTs are used as the dependent variable in testing a hypothesis, the way the RT distribution changes as a function of condition can rule out some explanations of processing. Thus, it is important to understand how RT distributions change as a function of condition when using RT measures. The second issue is practical: how might one squeeze the most out of experiments by reducing the impact of variability and outliers. A number of methods have been proposed to eliminate or reduce the impact of outliers. No one method is best

under all situations. However, there are best options if one knows how distribution shapes are changing across conditions.

There are basic questions about what happens to the shapes of RT distributions when mean RT increases. Quite different conclusions are reached if the increase in mean RT is the result of the whole distribution of RTs slowing as opposed to the longer RTs slowing more than the shorter RTs. For example, if the whole distribution shifts, then one might look for an inserted serial stage. If the distribution spreads, then one might examine a sequential sampling model in which evidence accumulation rate changes. There are a number of examples of cases in which a model might be consistent with mean RT, but inconsistent with distribution shape changes across conditions (see Hockley, 1984; Hacker, 1980; Heathcote, Popiel, & Mewhort, 1991; Ratcliff, & Murdock, 1976).

A major problem in using RT measures is the problem of outliers and contaminants. Usually a few, or sometimes more, outlier RTs are found in the tail of the distributions. Outliers are responses outside the normal spread of the RT distributions and can be random guesses (either fast or slow) or result from a delay in processing due to inattention. A more general term is contaminant that refers to spurious responses that can appear anywhere in the whole RT distributions, so outliers are one kind of contaminant. One theoretical approach to deal with contaminants is to make assumptions about their distribution and model them along with a model of the decision process, for example, modeling them as delays in processing (Ratcliff & Tuerlinckx, 2002) or as random guesses (Ratcliff & Van Dongen, 2009; Vandekerckhove & Tuerlinckx, 2007), depending on the task and experimental manipulations.

With appropriate experimental controls, short outliers can be largely eliminated by punishing an extra fast response with a long delay in the experiment. However, long outliers that appear in the long right tail of the RT distribution are probably present in most data sets. Such outliers can affect the estimate of mean RT moderately, sometimes enough to change significant effects into nonsignificant effects, and they also affect the variance more seriously. I describe approaches to dealing with outliers later in the chapter.

Density, Distribution, and Hazard Rate Functions

There are three main standard ways of representing distributions of responses: probability density functions, cumulative distribution functions, and hazard rate func-

tions. The probability density function is the normalized version of the frequency distribution. A probability density function based on data involves dividing the RT scale into intervals, finding the number of observations in each interval, (thus producing a histogram) and dividing each number by the sum of the numbers. The cumulative distribution provides the cumulative probability that a process has terminated as a function of time. A cumulative probability of 0.3 at some time means that 30% of responses have been made (or terminated) by that time. For data, the cumulative distribution function at some time t is the sum of the probabilities in the probability density function up to t. In the continuous case (for theoretical distributions), the sum is replaced by an integral:

$$F(t) = \int_0^t f(t')\,dt'$$

where $f(t)$ is the probability density and $F(t)$ is the cumulative distribution.

A relative of the density and distribution functions is the hazard rate function. The hazard rate function is defined as

$$h(t) = f(t)/(1 - F(t))$$

The hazard rate function represents the probability that a process will terminate in the next instant of time given that the process has not yet terminated. For example, the exponential function has a constant hazard rate. This means that the probability of terminating at any time does not depend on how long the process has been running. If the hazard rate decreases over time, then the longer a process has been going, the more likely it is to terminate, and if it is increasing over time, then the longer a process has been going, the longer it is likely to go on.

The hazard function can serve to discriminate between models of different classes and so be used to evaluate models that produce predictions for RT distributions (see Luce, 1986). The main limitation in the use of the hazard function in comparing predictions to data is that often the part of the function that best discriminates between models is the part of the function in the right tail of the RT distribution that corresponds to long RTs. The problem is that the tails of distributions have few observations and so are less reliably estimated and outliers appear in the right tail. Therefore, use of the hazard rate should include some examination of the effects of outliers. RT distributions can also sometimes provide strong evidence that a condition contains a mixture of processes. Two well separated peaks in the RT distribution might indicate a mixture of two separate processes. It might be, for example, that observed in-

creases in mean RT might be the result of one but not the other process slowing, or few responses in the faster process and more responses in the slower process.

The ExGaussian and Inverse Gaussian Distributions

In order to illustrate how distribution shape is evaluated, how outliers affect measures of shape, and how outliers affect applications of models to data, I use two simple explicit distributions that have shapes similar to empirically observed RT distributions, the exGaussian and the inverse Gaussian. These distributions have also been used in simulations that examine the power of statistical tests under a variety of conditions and methods of data analysis. The exGaussian distribution (Hohle, 1965; Ratcliff, 1979; Ratcliff & Murdock, 1976) has also been widely used to summarize RT distribution shape. If the model fits experimental data reasonably well, then the behavior of the model parameters can be used to interpret the distribution changes across experimental conditions. The inverse Gaussian is the distribution that is produced by a one boundary diffusion process (e. g., Burbeck & Luce, 1982; Ratcliff, 1978) and so has a more theoretical basis than the exGaussian.

For the exGaussian distribution, a RT from the distribution is the sum of a random value from a Gaussian (normal) distribution (mean μ and SD σ) and a random value from an exponential distribution (with mean τ). The density function is:

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

The mean of the distribution is $\mu+\tau$ and the SD is $\surd(\sigma^2+\tau^2)$.[①] When $\sigma=0$, the distribution is shifted exponentially (starting at μ) and the mean is $\mu+\tau$, the SD is τ, and the median is $\mu+\tau\log_e(2)$. Thus Pearson's skewness (3(mean-median)/SD, see later) is 0.912. For the exGaussian, as σ increases, the skewness decreases.

The inverse Gaussian distribution is the distribution of finishing times in a one-

① In adding together RTs from two distributions (called a convolution), for example, the normal and exponential in the exGaussian, the SD of the convolution is square root of the sum of squares of the two SDs as shown here. This provides an interesting insight into variability in keyboard responses in RT experiments. In some cases, keyboard responses are collected by scanning across keys, sometimes at a slow rate, e. g., 60 ms for a complete scan. This might be seen as a major increase in inaccuracy. But, because this variability is convolved with variability in the RT distribution, the actual increase in inaccuracy is very small. For example, if the scan time is 60 ms, the SD is $60/\surd(12)$ ms for a uniform distribution, which is 17.3 ms. Thus, if the SD in the RT from the subject is as low as 100 ms, the SD for the convolution is the square root of 100 squared plus 17.3 squared which is 101.5 ms. That is, a negligible increase in the variability of the responses and almost no loss of power.

boundary diffusion process. The density function is; where the mean is $\theta + T_{er}$ and the SD is $\sqrt{(\theta^3/\lambda)}$.

$$f(t) = \sqrt{\frac{\lambda}{2\pi(t - T_{er})^3}} e^{\lambda(t-\theta-T_{er})^2/(2\theta^2(t-T_{er}))}$$

This function can be reparameterized and written with diffusion model parameters as

$$f(t) = \frac{z}{\sqrt{2\pi s^2(t - T_{er})^3}} e^{(z-v(t-T_{er}))^2/(2s^2(t-T_{er}))}$$

where the transformation between the two forms is $\theta = z/v$ and $\lambda = z^2/s^2$.

For both the exGaussian and inverse Gaussian functions, the probability density function rises quickly and then falls more slowly, like empirical RT distributions. Figure 1 illustrates the behavior of the probability density function, the cumulative density function, and the hazard rate of the exGaussian and inverse Gaussian distributions. The Gaussian mean is 600 ms, the Gaussian SD is 50 ms, and the exponential mean is 200 ms. For the inverse Gaussian, the mean is 750 ms, and $\lambda = 750$ (to give a SD of 196.4 ms). In each panel of Figure 1, there are 20 distributions, each with 500 simulated observations per distribution. The cumulative density function is computed simply by sorting the data and producing cumulative counts at each RT. The probability density function and hazard functions use kernel estimation methods (Van Zandt, 2000). These essentially smooth the function at time t by averaging over data points around t.

Variability in the random samples of simulated data produce density and distribution functions (top two panels) that differ only a little across the different samples. However, most of the hazard rates show a lot of variability in the tails. The functions rise to a peak at around the mean RT, and then the function either rises, falls, or levels off. After 1000 ms, it is difficult to see any regularity in the location or shape of the tail of the function. The theoretical exGaussian hazard function rises and then levels off with a flat tail. It is also interesting to note that there is no way of discriminating between the inverse Gaussian and exGaussian by eye in Figure 1.

If either the exGaussian or the inverse Gaussian distribution is fit to RTs from a single experimental condition for a single subject, the parameter estimates can be used to summarize RT distribution shape. Also, comparisons of parameter values across conditions can be used to summarize the changes in distributions over conditions. Visual inspection of RT distributions at the most general level shows about

three main features: the location at which the front edge of the distribution begins to rise, the rate of rise in the front edge, and the rate of fall in the tail, i. e. , about 3

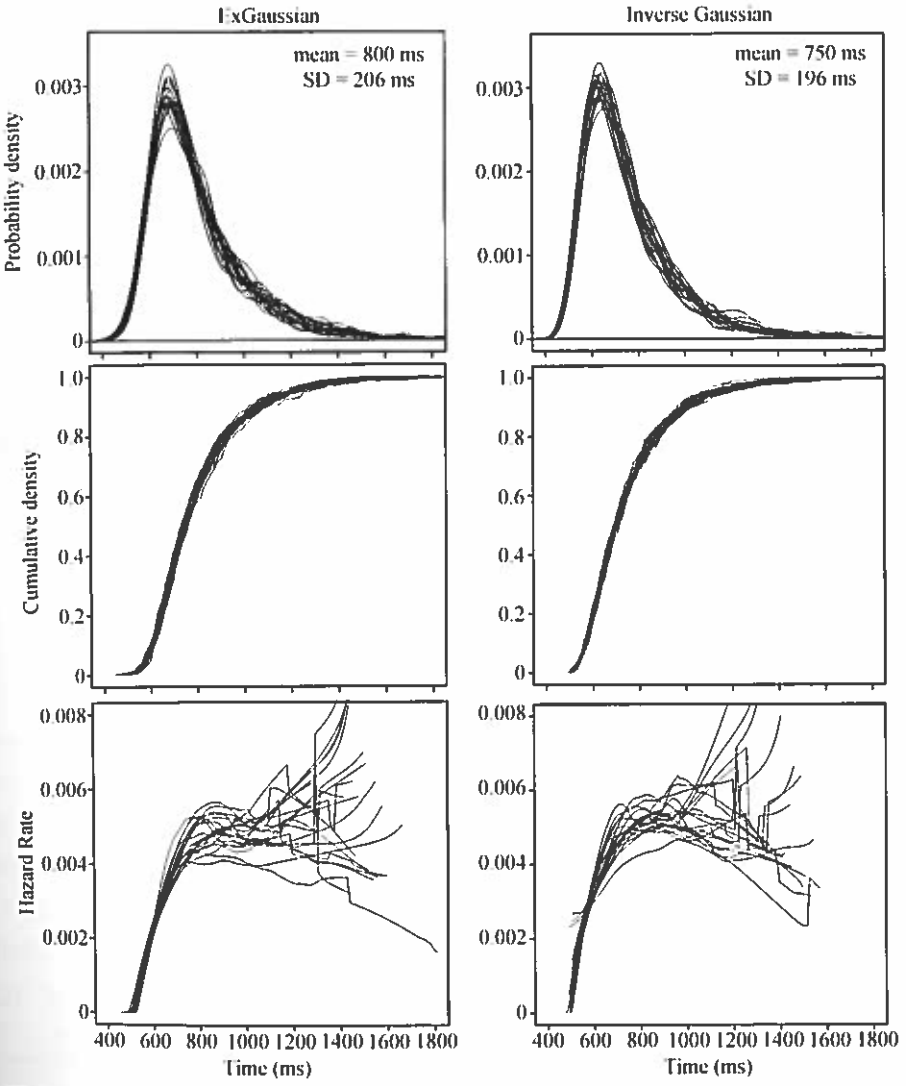

Figure 1 Probability density functions, cumulative density functions, and hazard functions for exGaussian and inverse Gaussian distributions. The left hand figures show 20 simulated exGaussian distributions with 500 observations per distribution with a Gaussian mean $\mu = 600$ ms, a Gaussian SD $\sigma = 50$ ms, and an exponential mean $\tau = 200$ ms. The right hand figures show 20 simulated inverse Gaussian distributions with 500 observations per distribution with $\theta = 350$, $\lambda = 750$, and $T_{er} = 400$.

degrees of freedom (df). Both the exGaussian and inverse Gaussian distributions have 3 parameters (3 df) with which to represent the whole distribution. The distribution that appears to be most useful is the exGaussian. Its three parameters capture the aspects of empirical distributions as noted above: the mean of the normal for the location of the fastest responses (and the mode), the SD of the normal for the rise of the front edge of the distribution, and the exponential parameter for the spread of the right tail.

However, it is hazardous to assign meaning to the parameters of these two models. For example, for the exGaussian, one might think that the Gaussian represents one process and the exponential another. Attempts to identify processes in this way have not been fruitful. The inverse Gaussian represents finishing times of a single boundary diffusion process. This has been considered a viable candidate for simple RT tasks (e. g., Burbeck & Luce, 1982; Luce, 1986; Ratcliff & Van Dongen, 2011; Smith, 1995), but would not be appropriate for two-choice tasks.

Fitting Reaction Time Distributions

Here, I will describe two methods for fitting the models of RT distributions to empirical data (Heathcote, Brown, & Mewhort, 2002; Ratcliff & Murdock, 1976; Ratcliff & Tuerlinckx, 2002; Van Zandt, 2000, 2002). The best method is the maximum likelihood method; the estimates are best in the sense that the variability in the estimates of the parameters is smaller than any other unbiased estimate. With this method, RT is put into the expression for the probability density to find the probability density for that RT (e. g., equations for the exGaussian and inverse Gaussian). Then the densities for all the RTs are multiplied and this product is called the likelihood. If $f(t_i)$ is the density for the ith RT, then $L = \prod_i f(t_i)$ is the likelihood. The parameter values of the model are adjusted to find those that maximize the likelihood. In practice, because the product of many likelihoods can be small, the logarithm of each likelihood is taken and minus the sum of the likelihoods ($\log(ab) = \log(a) + \log(b)$ so the logarithm of products is the sum of logarithms) is minimized as a function of model parameters. This works because the logarithm of a function is monotonically related to the likelihood and the parameter values that maximize the likelihood

are also those that minimize minus the log likelihood. Minimization can be done by any of many function minimization routines (see Press et al., 1992, Ch. 15, for a practical introduction to model fitting).

One problem with the maximum likelihood method for most distributions used to model RT is that a single short outlier RT can distort the whole fit. For example, the inverse Gaussian cannot produce a RT less than T_{er}. This means that if a RT is very short, T_{er} must be adjusted to be less than that RT in order to produce a probability density. Of course, short RTs could be trimmed, but then the value of T_{er} would be partly determined by the value of the cutoff which means it will be the choice of the person fitting the model.

One way to mitigate this problem is to use quantile RTs instead of individual RTs. If the quantile RTs are entered into the fitting program instead of the raw RTs, then for the exGaussian, if σ is much smaller than τ (as is typical in fits to empirical data), then σ will often be estimated to be zero. This is because there are fewer quantile RTs than data points and if the tail spreads out, most of the quantiles will in the tail with few representing the rise in the leading edge.

The more theoretically sound way to fit the models using quantile RTs is to use the areas between the quantile RTs to compute either a chi-square or a G-square goodness of fit statistic (see Ratcliff & Smith, 2004). If we use the .05, .15, .25, ..., .95 quantile RTs, then .05 probability mass lies outside the .05 and .95 quantiles and .1 probability mass lies between the quantiles. These probability masses are multiplied by the number of observations to give the observed frequencies. The quantile RTs can be used with the theoretical cumulative density function to compute the areas between the quantile RTs which represent the expected probability masses (which are multiplied by the number of observations to give the expected values). Then a chi-square or G-square goodness of fit measure can be computed and parameters can be adjusted to find its minimum. The observed probability masses (p_i) for the example above are .05, .1, .1, ..., .05, and if the expected probability masses are π_i, then the chi square statistic is $\chi^2 = N\sum_i (p_i - \pi_i)^2/\pi_i$ where N is the number of observations. Similarly, $G^2 = 2N\sum_i p_i \ln(p_i - \pi_i)$. I have found that minimizing chi-square and minimizing G-square produce almost the same parameter estimates. This is not surprising because they are asymptotically equivalent (Jeffreys, 1961, pp 196—197). A possible problem with the use of quantile RTs is that the

quantiles are random variables and not fixed values, as likelihood theory would require (Speckman & Rouder, 2004). However, in the cases in which fixed non-random bins and quantile bins have been used, the results have been essentially the same as each other (e.g., Fific, Little, & Nosofsky, 2010).

The use of quantiles has the major advantage that a few short RTs (e.g., less than 10% if the lowest quantile used is the .10 quantile) will not distort the fit. Along with some judicious trimming of short RTs, for example, much shorter than responses based on processing the stimulus (and not guessing), the use of quantiles largely avoids problems with short outliers. In addition, some trimming of long RTs (e.g., at say 2 or 3 seconds if the mean RT is 600 ms) along with the use of quantiles will reduce the effect of long outliers.

In order to see how well any fitting method recovers parameter values, I run Monte Carlo simulations. In these, I generate simulated data from the model and then fit the model to the simulated data. Then I repeat this, say 100 times, i.e., 100 Monte Carlo simulations, and compute the mean parameter values and the SD in the parameter values. The values of the mean provides a way of looking at bias in the parameter estimates, that is, does the fitting method produce fits that systematically differ from the values used to generate the simulated data? The values of the SDs provide estimates of variability in the parameter estimates based on the data sample size and the SDs can be used to compare the efficiency of different estimation methods. In addition, the SDs provide estimates of variability for statistical tests on parameter values across conditions and also provide estimates that allow comparison with individual differences to see if individual differences are larger than sampling errors in the parameter values. Finally, correlations or covariances among parameter values across the Monte Carlo trials can be used to examine tradeoffs across parameters. The way to interpret such tradeoffs is: if one or more data points (e.g., quantile RTs) were extra high by chance, then the model may compensate for this by moving one parameter higher and in another parameter (or more) may also move higher or lower to also compensate. These correlations can also be used to decide whether differences in parameter values might be the result of real differences or tradeoffs. A detailed discussion is presented in Ratcliff and Tuerlinckx (2002, Figure 5). Ratcliff and Murdock (1976) provide theoretical estimates of SDs in exGaussian parameter values, and Wagenmakers et al. (2004) discuss such Monte Carlo methods in detail. Properties of estimators are presented in Ratcliff and Tuerlinckx (2002, Appendix A), and

these are standard in mathematical statistics.

It is also important to note the distinction between correlations between model parameters from random sample of data from a model with fixed parameter values (these are the examples above) and correlations in model parameters from individual differences across subjects. In these two cases, the patterns of correlations can be quite different (for examples with the diffusion model, contrast Ratcliff & Tuerlinckx, 2002, Figure 5, with Ratcliff, Thapar, & McKoon, 2010, Table 6).

Representing Distribution Shape: RT Quantiles

In the prior section we saw that quantile RTs can be used in fitting a model to data, and here we show how they can be used to display distribution shape. The quantiles of a RT distribution are the times at which some proportion of the processes have terminated. In Figure 1 middle panel, the 0.2 quantile RT is obtained by drawing a horizontal line from 0.2 on the y-axis to intersect the cumulative distribution function and the quantile RT is the RT on the x-axis where the two intersect. Quantiles of a RT distribution can be used as a summary of the distribution as is shown in Figure 2A. The circles connected with the jagged line show a frequency polygon, which is a histogram but with the top of the bars replaced by the circles. On the x-axis, the arrows show RT quantiles. Because there is a probability mass of .2 between the .1, .3, .5, .7, and .9 quantiles, and .1 outside of each of two extreme values, rectangles with these areas can be constructed between and outside the quantiles. The further apart the quantiles are, the lower the height of the rectangle. The .005 and .995 quantiles are used to mark the extremes of the distribution because they provide relatively stable estimators of the fastest and slowest RTs. Because each of the rectangles has an area of .2 (with the remaining .2 shared between the two extremes), all of the information about distribution shape is carried by the spacing between the quantiles. As can be seen, the equal-area histogram captures the overall shape of the distribution, i.e., its location, spread, and skewness, as well as does the frequency polygon. This correspondence works for RT distributions with as few as 5 quantiles because RT distributions are usually regular with a fast rise and a slower fall in the right tail.

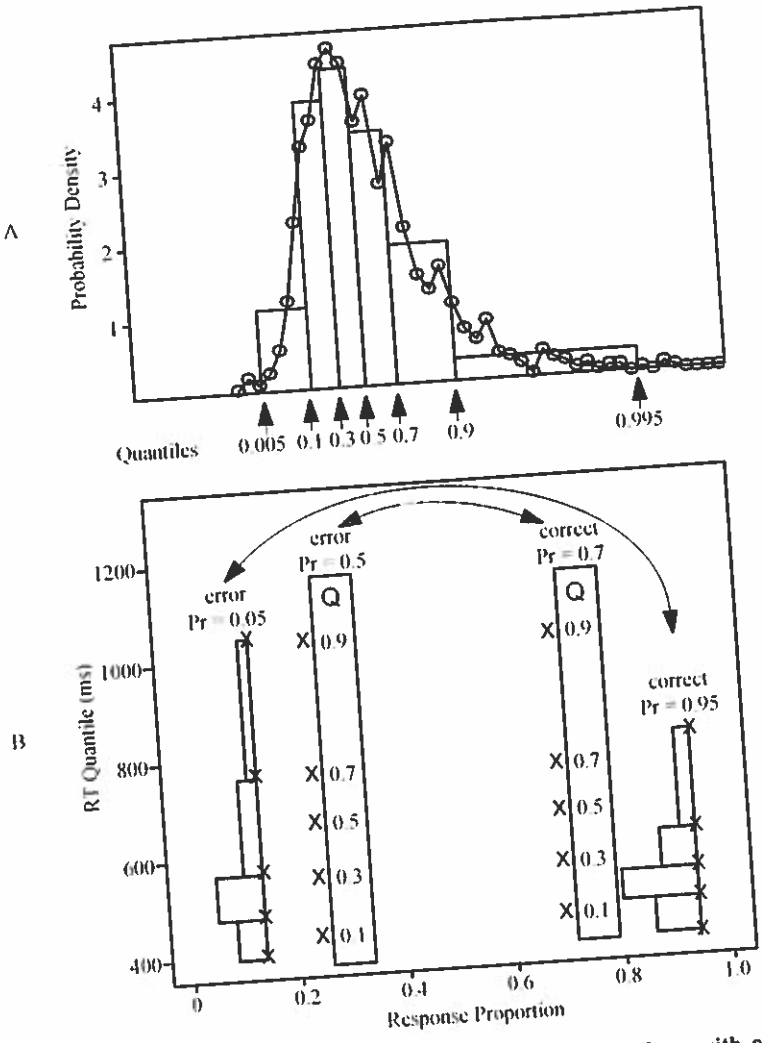

Figure 2 A shows a RT distribution as a frequency polygon, along with a quantile RT distribution with equal area rectangles drawn between the .1, .3, .5, .7, and .9 quantile RTs and rectangles with half the area outside the .1 and .9 quantile RTs. B shows a quantile probability plot with the proportion of responses for that condition on the x-axis and quantile RTs plotted as x's on the y-axis. Equal areas rectangles are drawn between two of the sets of the quantiles to illustrate how to interpret RT distribution shape in the plot (these are comparable to the distribution in A). C shows an example of a quantile probability plot (Figure 1, Ratcliff & McKoon, 2008) for data (x's) and diffusion model fits (o's and connecting lines).

C

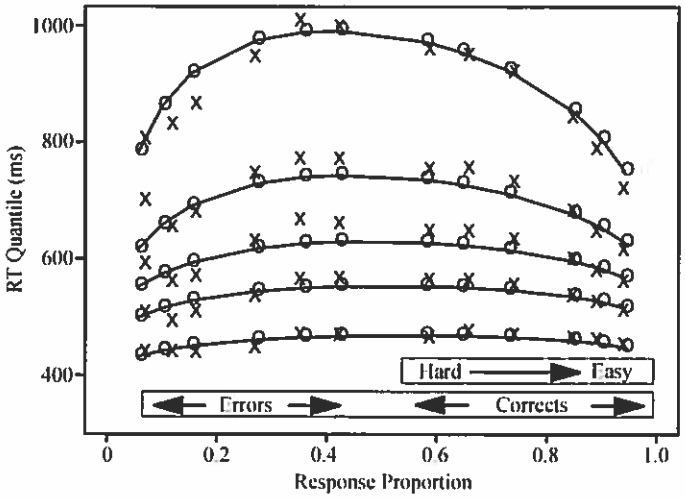

Figure 2 (continued)

Ratcliff (2001) presented plots of quantiles against accuracy as a way of showing the way response proportion and RT distributions jointly change as a function of the independent variable. Figures 2B and 2C show such plots, termed quantile probability plots. The quantiles of the RT distribution for each condition are plotted vertically with response proportion for the condition the value on the x-axis. If the probability of a correct response for a particular stimulus discriminability is p, the quantiles of the distribution of correct responses are plotted in a vertical column against p on the X-axis and the quantiles of the distribution of errors are plotted against $1-p$. In Figure 2B, this correspondence is illustrated by the double-ended arrows connecting pairs of conditions. This means that correct responses appear (usually) on the right of .5 point on the x-axis and errors appear on the left. In plots of this kind, the outermost pair of distributions in the figure are the errors and correct RTs for the easiest stimulus condition and the innermost pair are the errors and correct RTs for the most difficult stimulus condition.

When one becomes familiar with the plot, one can read off the way RT distributions change over conditions. For example, in Figure 2C (cf., Experiment 1, Ratcliff & McKoon, 2008) going from right to left, the leading edge (lower "x" .1 quantile RT) increases a little, but the tail (top "x" .9 quantile RT) increases from around 750 ms to 950 ms. This means that the RT distribution is shifting only a little, but is spreading out a lot more. Comparing the right and left quantiles, error responses have

a longer tail than correct responses, but the leading edges are about the same. These quantile probability plots can contain both the data and the predictions and hence provide a useful way of examining the joint fit of accuracy and correct and error RT distributions of a model to data (Ratcliff & Smith, 2004).

In many situations in cognitive psychology, materials are difficult to construct or a limited number are available. Also, with patient populations, relatively few observations may be able to be collected. For example, in text processing research it can be extremely difficult to construct paragraphs with the required structure while controlling potentially confounding variables, in clinical research, there may be relatively few words associated with, for example, anxiety, and in semantic memory research there may be a limited number of typical members of a category or highly associated pairs of items. In these situations it may be impossible to get more than 30 or 40 observations per subject per condition. Averaging quantiles over subjects provides a way of grouping data when the number of observations is this small. Note that just combining the RTs from the different subjects does not work. For example, if RTs were combined from 2 subjects that had narrow distributions that were well separated (e.g., with means at 500 ms and 800 ms), then the resulting distribution would be bimodal and not reflect the shape of either individual distribution.

In the process of fitting theoretical distribution functions to experimental data, it usually becomes necessary to obtain an average of the parameter values across subjects in order to make some statements about group trends. There are two main ways to do this: First to average the data across subjects in some way, then fit the model and use those parameters as the summary. Second, to fit the model to individual subject's data, then to average the individual's parameters to provide the group parameters. In fitting the diffusion model to data, we have performed this comparison a number of times. In almost all the cases, the parameters of the model fit to group data are close (within 2 standard errors) to the average of the parameter values from fits to individual subjects (Ratcliff, Thapar, & McKoon, 2001, 2003, 2004, 2010; Ratcliff, Thapar, Gomez, & McKoon, 2004; Thapar, Ratcliff, & McKoon, 2003). At this point, this is a practical result and not theoretically exact, and it has only been performed for the diffusion model and not competing models.

Measures of Distribution Shape

The question that is of interest in this section is how to obtain information about distribution shape from real RT data. The first question to be asked is what is meant by distribution shape? The notion of distribution shape can be defined in several different ways. Probably the most reasonable and least theory-bound is that given by Mosteller and Tukey (1977, Chapter 1). They define shape as what is left when location (position of the distribution) and scale (spread of the distribution) are given up, i.e., the distribution is normalized. A probability density function is not sufficient to define shape as defined by Mosteller and Tukey, for example, their Figure 4 (chapter 1) shows the family of beta density functions have the same mathematical form but differ widely in shape.

Skewness

Everyone who has taken an introductory statistics course is familiar with the mean, SD, and variance of a set of scores. The mean represents the location of the distribution and the SD the spread or scale of the distribution. Introductory statistics books sometimes discuss skewness and kurtosis as measures of distribution shape. These are based on the moments of the distribution. For example, the kth moment can be written as

$$\mu_k = \int_{-\infty}^{\infty} (t - \mu_1)^k f(t)\,dt$$

Then skewness is defined as $Sk_m = \mu_3/s^4$ and kurtosis is defined as $\kappa = \mu_4/s^4$. It can be demonstrated mathematically that if all the moments of the distribution are known, then the shape of the distribution is completely determined.

However there are very serious problems with using moments as measures of distribution shape (Ratcliff, 1979). First, the contributions to the third and fourth moments come from relatively far in the tail of the distribution (see Figure 6, Ratcliff, 1979, which was reprinted from Pearson, 1963). This means that these moments are sensitive to parts of the distribution that do not correspond to what we see as distribution shape by visual inspection. Second, the higher moments have very large standard errors associated with them. This means that many more observations are needed to obtain reasonable estimates (with low standard errors) than are usually collected in

RT experiments. Third, outlier RTs can affect the size of moments from the variance on up very severely, to the extent that if there are a few extreme reaction times, the higher moments essentially reflect these long RTs. The problem becomes critical if some of the outlier RTs are from processes other than the process under consideration (e.g., a second retrieval attempt or even worse, a head scratch or moment's distraction). Then, the higher moments are measuring outlier RTs and these outliers are not of particular interest in determining distribution shape.

Better measures of distribution shape are Pearson's second skewness measure and the quartile coefficient of skewness. Pearson's second skewness measure is defined by

$$Sk_p = 3(\text{mean-median})/s$$

and quartile skewness is defined by

$$Sk_q = (Q_3 - 2Q_2 + Q_1)/(Q_3 - Q_1).$$

Ratcliff (1993) performed a set of simulation studies that included a comparison of the behavior of these different measures of skewness across different random samples of data from the same distribution. Results showed that Pearson's second skewness measure and quantile skewness correlated highly and neither correlated with skewness from the third moment. Practically, what one sees in visually examining a distribution corresponds to Pearson's second skewness measure and the quartile coefficient of skewness.

Explicit Distribution Functions

Some of the earliest attempts to model the shape of RT distributions started with an assumed distribution and attempted to work back from fits of the distribution to empirical data, and then to the processes underlying the task. For example, McGill (1963) voiced the hope that the shape of the RT distribution would serve as a signature that would help identify the underlying processes. It seems that the strong version has not worked out, rather RT distributions are critical in testing models and cannot be used to unambiguously identify processes (cf., the similar shaped distributions in Figure 1). However, RT distribution shape provides the strongest constraints on process models that attempt to account for RT distribution shape for correct and error responses as well as accuracy.

The exGaussian has used in a number of applications to summarize the shape of RT distributions (e.g., Balota & Spieler, 1999; Hockley, 1984; Hacker, 1980; Heathcote, Popiel, & Mewhort, 1991; Ratcliff, 1979; Ratcliff, & Murdock, 1976;

Yap, Balota, Tse, & Besner, 2008). The advantage of using this distribution is that the parameter τ provides an estimate of the fall in the tail relative to the rise of the distribution (which is represented by σ).

From a more theoretical perspective, Matzke and Wagenmakers (2009) used the diffusion model for two choice decisions to generate predictions and showed that if the diffusion model were correct, then the way the diffusion model parameters vary does not correspond to the way the exGaussian parameters vary. For example, a change in drift rate does not correspond to a change in one exGaussian parameter.

Outlier Reaction Times

Everyone who has used RT measures has realized that all the RTs collected do not come from the processes under consideration. One has only to observe subjects scratching themselves, adjusting an ipod, or answering a cell phone during the response period to be aware that there is likely to be bad data mixed in with the good. Even if one tests oneself then it is very apparent that sometimes a lapse of concentration produces a long RT. If outlier RTs were symmetric, so that a certain proportion of outlier RTs were long and another proportion short (with the same spread on either side), then there would be no bias in detecting differences among conditions, except for a reduction in power. The mean would remain approximately constant, but the variance and other moments would increase. However, subjects in most cognitive paradigms (if they are attempting to comply with the experimenter) will mainly produce long outlier RTs. These will affect the mean, variance, and measures of distribution shape. Furthermore, just one very long reaction time can completely change the pattern of means in an experiment. For example, suppose that 100 observations per condition are collected in an experiment and the process means are: condition A, 600 ms and condition B, 650 ms. Suppose in condition A there two outliers at 2.5 and 3.5 sec. The observed mean in condition A will then be 648 ms, thus masking the real 50 ms difference.

Fast Guesses

The other kind of outliers are fast guesses. It is possible to identify such guesses by setting an upper RT cutoff at say 150, 200, and 250 ms, and then examine accuracy. Fast guesses will produce chance or near change performance for these faster re-

sponses. As the upper cutoff is increased, the point at which accuracy starts to rise above chance will show the point at which RTs begin to come from the processes involved in the task. Specifically, accuracy might be examined within a series of RT windows, say 150—200 ms, then 200—250 ms, etc.

We find that most subjects do not fast guess, but a few do when given the opportunity. By fast guessing, uncooperative subjects would be able to leave the experiment earlier than they would if they tried to perform the experiment according to instructions. We have found that fast guessing can almost completely be eliminated by modifying the experiment by inserting a 1.5 or 2 second delay after a very fast response, e.g., responses less than 150 ms in a fast perceptual task or less than 300 ms in a recognition memory task. This long delay eliminates the motivation for fast guesses. However, sometimes fast guessing may be optimal (Bogacz et al., 2006) or they may be a domain of study in their own right (Ollman, 1966). But usually in our work, they are a nuisance that can be eliminated with this simple modification.

Slow Outliers

The following example will demonstrate the kind of results expected when long outliers are present. In order to demonstrate properties of the process of trimming long outlier RTs, 96 random numbers were generated from an exGaussian distribution with parameters, $\mu = 500$ ms, $\sigma = 200$ ms, and $\tau = 200$ ms, values like those in observed recognition memory RT distributions. To these 96 RTs, 4 outlier RTs, were added at 4, 3, 2, and 1 sec. The mean and SD in RT were calculated for all the RTs and at cut-offs of 3100, 2100, and 1100 ms. The means were 786, 754, 716, and 667 ms, and the SDs were 485, 361, 245, and 154 ms. The theoretical mean for the exGaussian distribution without outliers is 700 ms and the SD is 206 ms.

With all the RTs included, the mean is overestimated by 86 ms and the SD is over double the theoretical value (the value without the outliers). By the 1100 ms cutoff the mean is 33 ms less than the theoretical value and the SD is 44 ms less than the theoretical value. These last two values at the 1100 ms cutoff demonstrate an important problem with trimming, and that is that 7 RTs besides the outliers are trimmed out, giving serious underestimation of both the mean and SD.

The message is simple: In almost any set of RT data, there are spurious long RTs. Trimming the data will remove many of these spurious RTs, but will also remove long RTs that come from the processes under investigation. The question is can

we find a rule of thumb that will maximize removal of the spurious RTs, yet minimize removal of real data. Ratcliff (1993) examined this question and found that there was no rule of thumb that did not fail in some situations. This study is now reviewed.

Methods for Reducing the Impact of Long Outliers

The criterion that has been most often suggested involves trimming out data that falls some number of SDs outside the mean of that condition (for example 2 or 3 SDs). This process can be repeated with a recomputed mean and SD from the trimmed data. It is important to note that any trimming or otherwise removing of data should be done completely independently of the hypotheses being tested. In addition, it is also reassuring if the trimming procedure does not remove significantly more data points from one condition than other conditions.

A diagnostic signal that outliers may be present in a condition of an experiment is the SD for the condition relative to those for other conditions. Suppose there are three conditions that are expected to have an increasing RT from first to last (e. g. 500, 600 and 700 ms). Suppose that the means are 520, 750, and 680 ms. Then it may be thought that the results disconfirm the prediction. However, an examination SDs in each condition may show that the conclusion is premature: If the SDs are 250, 500, 285 ms respectively, then it may be that outliers are responsible for the long mean RT in the middle condition. Trimming outliers may then produce the following set of results: 505, 602, and 678 ms respectively with SDs, 220, 249, 277 ms respectively. This pattern is more satisfactory because it has monotonically increasing means and SDs.

Power of ANOVA and Outliers

There are several alternative methods for dealing with long outlier RTs in common use (beside trimming described above). First, the data can be Winsorized: Long RTs, instead of being trimmed out can be replaced by RTs at some predetermined ceiling, e. g. 3 SDs above the mean. This method makes the strong assumption that if there was no inattention (that produced the long outlier response), the process would have still produced a long RT. In general, I believe that this assumption is not justified in the RT domain. A second method involves use of medians. Typically the median RT is computed for each subject in each condition and then these medians are used in an analysis of variance (for example). The advantage of

medians is that they are insensitive to a few outliers. This raises the question: If medians avoid the problem of outliers, then why is the median not used routinely instead of mean RT? The reason is that medians have higher variability than means. Simulations using medians showed that they rarely produce as much power as other methods. A third method involves transforming the RT data, by a log or inverse transformation. Both of these transformations reduce the impact of long RTs on the means.

Ratcliff (1993) performed a number of simulations to mimic the use of these different methods of dealing with outliers on the power of ANOVA. It turns out that the methods differ in their ability to increase power as a function of how the RT distribution changes across conditions. Thus the simulations were carried out with assumptions that, first, the RT distributions were spreading with an increase in mean RT (as occurs with changes in memory strength or perceptual strength) or that the distributions were shifting with increase in mean RT (as occurs with visual search when the order of search is controlled, Hockley, 1984). Simulations also examined the effects with and without outliers.

Results showed that no one method was optimal. If the difference in means between two conditions occurs because the RT distribution spreads with no change in the leading edge, and there are no outliers, then trimming reduces power as the cutoff is reduced. If there are outliers, then trimming increases power to a maximum as the cutoff is reduced, and then power decreases as more and more genuine RTs are eliminated (i.e., the data that are responsible for the difference in means RTs). On the other hand, if the difference between two conditions is due to the distribution shifting with no change in the spread of the tail, then trimming increases power as more and more long RTs are eliminated (the long RTs are more variable and hence reduce power). This occurs both with and without outliers. Results were also reported for log and inverse transformations, using medians, trimming the longest RT, trimming at some number of SDs above the mean, and Winsorizing. One method that seemed to give high power for studies with and without outliers and with the distributions both spreading and shifting was the inverse transformation.

The prescription we follow in dealing with outlier RTs in analyzing data is as follows: In a new experimental paradigm, we analyze the data in several different ways. We look at the results without any trimming, we trim at several cutoffs and we calculate medians. If all these measures tell us the same thing then we proceed to other experiments with the measure (e.g., mean) derived from trimming at some reasonable

point (if the mean is 700 ms and SD is 300 ms, then a cutoff of between 1500 and 2000 ms will probably work very well) that is determined independently of the hypothesis being tested. What we really want to see (if the means calculated from the raw data are noisy) is order coming out of variability as the RT cutoff is reduced. We also want to see that the median tells us the same thing as the trimmed mean. If the measures do not agree, then there may be problems. It may be that there are no real trends in the data; statistical tests will usually confirm this by not producing significance at any cutoff or with medians. The results collected from the different cutoffs, transformations, and medians may then point the way towards a better design. But in all of these analyses, one should not experiment with different methods with marginal data to try and find one that produces a significant effect. Such a significant effect may be spurious. The most important outcome is to have reported a result that replicates.

Diffusion Models

Diffusion models are models of the decision process typically in two-choice tasks for which mean RTs are under a second or second and a half. They can be also seen as replacements for signal detection theory, one of the most-used methods of analyzing accuracy data. In addition to accuracy, it is important to realize that all behavioral cognitive tasks provide RT measures for both correct and error responses. Sometimes, examination of one variable may suggest one effect while examination of the other variable might suggest something different. For example, in several tasks in which the effects of aging on processing are examined, accuracy shows no decrement with age while RT shows a large decrement (e. g. , Ratcliff et al. , 2010). To understand processing, explicit diffusion sequential sampling models can be fit to data and the parameters, which reflect components of processing, can be used to interpret the differences between accuracy and RT.

There are several different kinds of diffusion decision models that account for distribution shape in two-choice tasks. These diffusion models (Ratcliff, 1978; Ratcliff & Smith, 2004; Ratcliff & McKoon, 2008; Usher & McClelland, 2001; Wagenmakers, 2009) provide an account of how accuracy and the shapes of RT distributions change across experimental conditions that manipulate difficulty and speed/accuracy criterion settings. Here I will focus on the single diffusion process model (for a review, see Ratcliff & McKoon, 2008).

The diffusion model assumes that decisions are made by a noisy process that accumulates information over time from a starting point toward one of two response criteria or boundaries, as shown in the top panel of Figure 1. The starting point is labeled z and the boundaries are labeled a and 0. When one of the boundaries is reached, a response is initiated. The rate of accumulation of information is called the drift rate (v), and it is determined by the quality of the information extracted from the stimulus in a perceptual task or the match between an item and memory in a memory task. In an experiment, the value of drift rate, v, would be different for each stimulus condition that differed in difficulty. The zero point of drift rate (the drift criterion, Ratcliff, 1985; Ratcliff et al., 1999) divides drift rates into those that have positive values, that is mean drift rate toward the A response boundary in Figure 3A, and negative values, mean drift rate toward the B boundary. There is noise ("within trial" variability) in the accumulation of information so that processes with the same mean drift rate (v) do not always terminate at the same time (producing RT distributions) and do not always terminate at the same boundary (producing errors), as shown by the three processes, all with the same drift rate, in Figure 3A. Within-trial variability

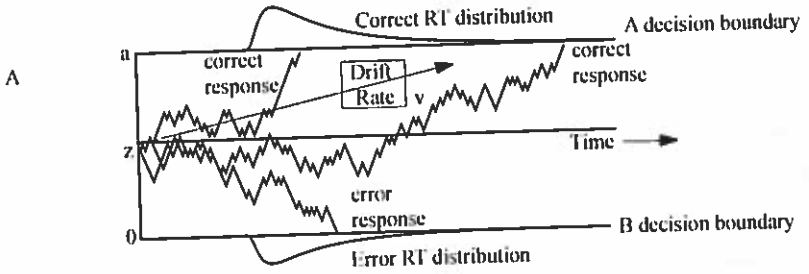

Figure 3 A shows an illustration of the diffusion model with drift rate v, boundary separation a, and starting point z. B shows how changes in drift rate affect RT distributions. If there is a reduction in drift rate of X, then the fastest processes are slowed by Y and the slowest by Z. C shows the way the model accounts for speed-accuracy manipulations and changes in evidence from the stimulus. D and E show explanations for the effects of response probability manipulations. D shows starting point changing with probability and E shows changes in the drift criterion (zero point of drift). When the probability of response A is higher, the drift rates are v_a and v_b with the zero point close to v_b. When the probability of response B is higher, the drift rates are v_c and v_d and the zero point is closer to v_c. Note that this second alternative is exactly equivalent to how the criterion would change in signal detection theory from psychophysics.

in drift rate (s) is a scaling parameter for the diffusion process (i. e., if it were doubled, other parameters could be multiplied or divided by two to produce exactly the same fits of the model to data). Stimulus encoding, memory access, response preparation and output are all represented in the model by a distribution of nondecision processes with mean T_{er} and range s_t.

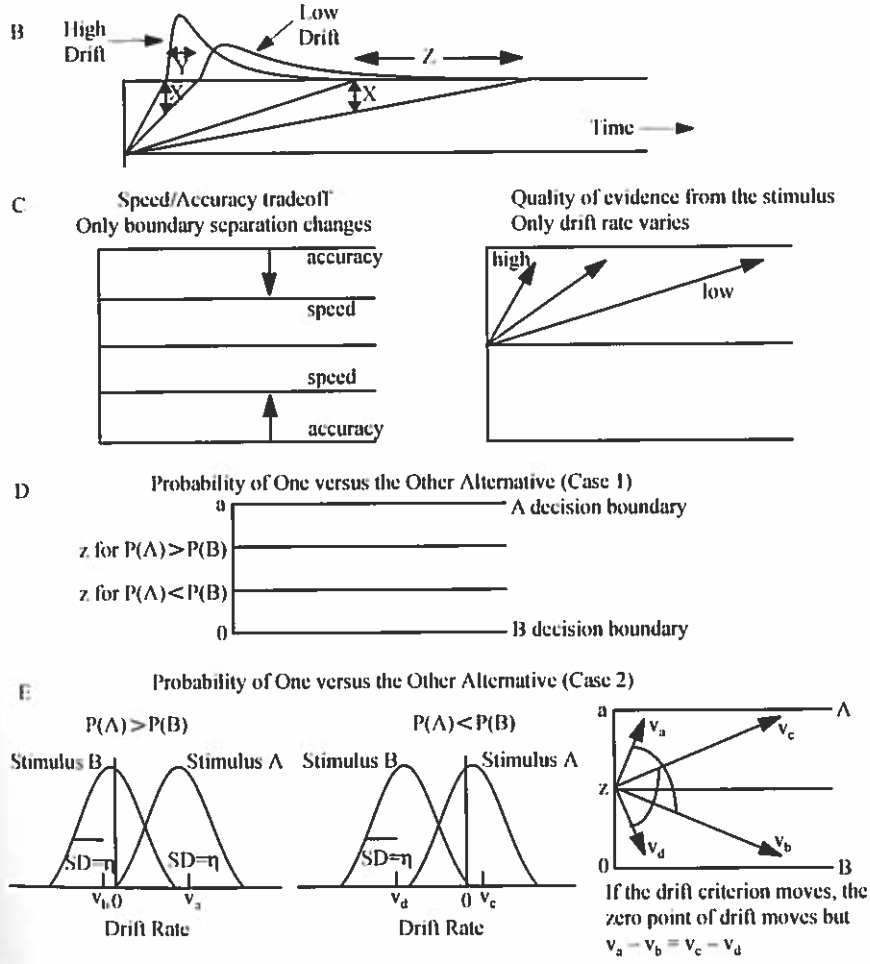

Figure 3 (continued)

Empirical RT distributions are positively skewed and in the diffusion model, this is naturally predicted by simple geometry. This is illustrated in the Figure 3B, distributions of fast processes from a high drift rate and slower responses from a lower drift rate are shown. If the higher and lower values of drift rate are reduced by the same

amount ("X" in the figure), then the fastest processes are slowed by an amount "Y" and the slowest by a much larger amount, "Z".

Figure 3C illustrates how boundary separation and drift rate change with speed-accuracy instructions and with stimulus quality. All the dependent variables, RTs for correct and error responses and accuracy, change with each of these two manipulations. But in fits to data, each manipulation is modeled by changes in one model parameter. The effect of speed-accuracy instructions is modeled by changes in boundary separation, and the effect of stimulus strength is modeled by changes in drift rate (e.g., Ratcliff & McKoon, 2008; Ratcliff et al., 2001, 2003, 2004). The effects of altering these two model parameters are identifiable: changes in boundary separation affect both the leading edge of the RT distribution and the tail (with a modest change in accuracy, typically 4%—8%, accompanied by a large change in RT, typically 200 ms or more) while changes in drift rate mainly affect the tail of the RT distribution as well as accuracy. More general issues of model falsifiability are discussed in Ratcliff (2002) in which fake data sets were generated and it was shown the diffusion model could not fit them. Note that signal detection theory would produce different values of d' for each different speed-accuracy condition in an experiment. In contrast, in all the cases in which speed-accuracy instructions have been used with diffusion model fits, invariant values of drift rate across conditions have been found.

A standard manipulation in two-choice experiments in psychophysics and human performance research is to vary the relative proportions of the two responses (e.g., Swets, 1961). This can be accomplished by changing the proportions of the stimuli: Stimuli for which one response is correct are presented on a larger proportion of trials than stimuli for which the other response is correct. Response proportions can also be manipulated without changing the proportions of stimuli: Subjects can be asked to be more careful about one response than the other, or subjects can be rewarded to a greater degree for one response than the other.

In the diffusion model, there are two ways of modeling the effects of these proportion manipulations. For one (Figure 3D), the starting point moves closer to the more likely response. When the starting point is far from the boundary at which a response is correct, the whole distribution of correct responses is shifted to longer RTs, with the slowest responses (e.g., .9 quantiles) slowing much more than the fastest responses (.1 quantiles). The second way of modeling response proportion manipulations is to adjust the zero point of drift rate. The Figure 3E illustrates the distributions

of drift rates for stimuli for which A is the correct response and stimuli for which B is the correct response. The distributions arise from across-trial variability in drift rate. Values of drift rate above the zero point are positive, that is, with drift towards the A boundary, and values below the zero point are negative, with drift towards the B boundary. When the probability of A being the correct response is higher (left panel), the zero point shifts toward the B distribution, and when the probability of B being the correct response is higher (right panel), the zero point shifts toward the A distribution. The differences between the means of the distributions do not change ($v_a - v_b = v_c - v_d$), only the zero point. The consequences for accuracy and distribution shape are the same as those for changing drift rate.

The components of processing are assumed to be variable across trials. For example, all words studied at list position 6 in different lists in a recognition memory task would not have exactly the same drift rate. The across-trial variability in drift rate is assumed to be normally distributed with standard deviation η and the starting point is assumed to be uniformly distributed with range s_z. One might also expect that the decision criteria would be variable from trial to trial. However, if the variability is not extremely large, the effects closely approximate the effects of starting point variability, and, computationally, only one integration over a distribution of starting points is needed instead of two separate integrations over distributions of the two criteria.

Error responses are typically slower than correct responses when accuracy is stressed in instructions or in experiments where accuracy is low and errors are usually faster than correct responses when speed is stressed in instructions or when accuracy is high (Luce, 1986; Swensson, 1972). Figure 4 shows how the across-trial variabilities work to produce the relative speeds of correct and error RTs. To illustrate how across trial variability in drift rate affects RTs (Figure 4A), the full (normal) distribution of drift rates around the mean v is abbreviated to just two values, one (v_1) a larger value of drift rate and the other (v_2) a smaller value. For each of these single drift rates, for equidistant boundaries from the starting point, correct and error responses have the same values (and distributions). This is a property of this diffusion process and random walk processes (see Feller, 1968). Accuracy is lower and both correct and error RTs are longer for the v_1 drift rate than the v_2 drift rate. When the two processes are combined, errors are slower than correct responses because the slow error responses (RT 600 ms) from v_2 have a greater probability of occurrence (proba-

bility .20) than the fast error responses (RT 400 ms) from v_1 (probability .05). This logic applies to a continuous distribution of drift rates. In the Figure 4B, the distribution in starting point due to across-trial variability is abbreviated to two values, one closer to the A boundary (at $z = a + .5s_z$), and one farther from the A boundary (at $z = a - .5sz$). Processes starting near the incorrect boundary have a greater probability of reaching that boundary (probability .20) and are faster than those starting farther

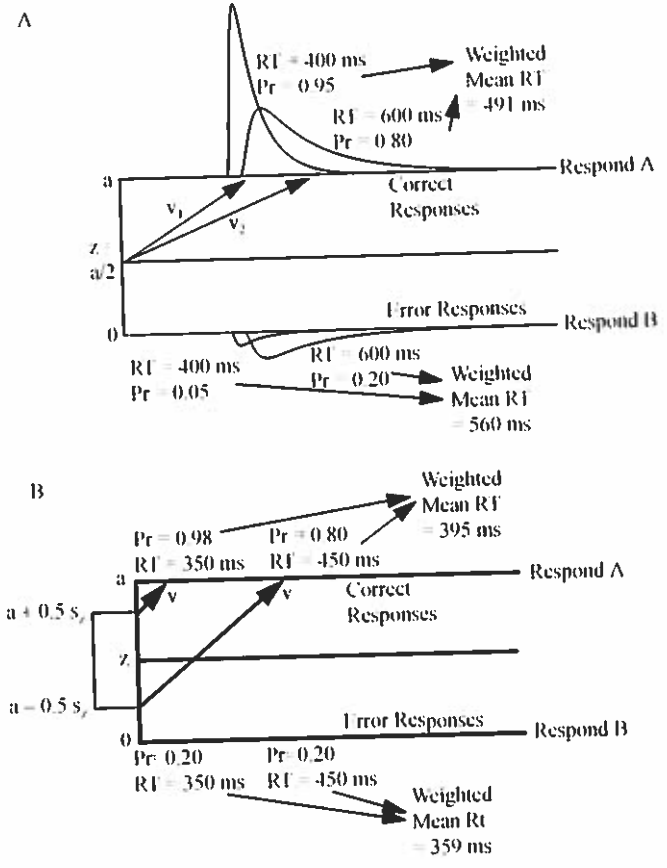

Figure 4 Variability in drift rate and starting point and the effects on speed and accuracy. A shows two process with drift rates v_1 and v_2 and the starting point halfway between the boundaries with correct and error RTs of 400 ms for v_1 and 600 ms for v_2. The weighted average yields error responses slower than correct responses. The bottom panel shows processes with two starting points and drift rate v. Averaging processes with starting point $a + .5sz$ (high accuracy and short RTs) and starting point $a - .5sz$ (lower accuracy and short RTs) yields error responses faster than correct responses.

away (probability .02), so their combination leads to errors faster than correct responses.

There are a number of empirical generalizations about how RT distributions change with experimental variables (some described above). First, RT distributions spread out to the right as the mean increases as a function of difficulty. There is only a small shift in the distribution. This is a strong prediction of diffusion models. Second, as speed/accuracy settings are altered (by, for example, instructions), RT distributions shift and spread. Again, this is a strong prediction of the models. Third, RT distribution shape is approximately invariant under all these manipulations, for example, plotting quantile RTs against quantile RTs across a wide range of conditions produces approximately linear functions (Ratcliff & McKoon, 2008; Ratcliff & Smith, 2010). Diffusion models also produce this kind of invariance. Fourth, diffusion models produce increasing hazard functions or functions that increase to a peak and then decrease a little (Ratcliff et al., 1999) to a flat tail, and these are the patterns of hazard functions that are observed in most two-choice data. In sum, most observed behaviors of RT distributions are captured almost automatically by diffusion models of the decision process.

Methods of fitting the model to data can be found in Ratcliff and Tuerlinckx (2002) and computer packages to fit the model are described in Vandekerckhove and Tuerlinckx (2007) and Voss and Voss (2007). These methods deal with outlier and contaminant RTs by explicitly modeling them and extracting estimates of the proportions of these responses.

The diffusion model has been applied in a number of domains including applications to aging and speed of processing, IQ, sleep deprivation, aphasia, hypoglycemia, anxiety, and depression. In all of these applications, the model is fit to the data from each subject individually with about 45 minutes of data collection per subject. The model has also been applied to behavioral data collected in studies along with data from on single cell recording in monkeys in perceptual decision making tasks (Ratcliff et al., 2007), and in perceptual decision making tasks with EEG (Philiastides et al., 2006; Ratcliff et al., 2009). In each case, relationships were found between diffusion model parameters and these neurophysiological measures (see Ratcliff & McKoon, 2008 for a brief summary of these topics).

Conclusions

This chapter discusses a number of issues, both practical and theoretical, that revolve around knowing the shape of RT distributions and how distributions change across conditions. From a theoretical perspective I described methods of fitting distributions, methods of summarizing distributions with quantiles or explicit distributions, methods of averaging distributions over subjects, and covariances among model parameters. A practical consideration in most research using RT measures is how to identify and deal with outlier RTs. It is clear that methods are available to deal with outliers that can increase the power of an experiment substantially. I then reviewed a theoretically based approach using an explicit model of the decision process to fit the accuracy and correct and error RT data. This approach produces estimates of the duration of processes other than the decision process, the amount of evidence needed to produce a decision and the quality of evidence from the stimulus or the match to memory. This approach produces an estimate of drift rate that is invariant under speed-accuracy manipulations and it can be the meeting point of the decision model and models of perceptual processing or memory retrieval (see Ratcliff, 1981; Smith & Ratcliff, 2009 for examples of integrating encoding and decision processes).

Probably the most important lesson from an empirical perspective that can be taken from the discussion presented here is that someone using RT measures needs to know how their RTs are distributed and how the RT distributions change over conditions. From a theoretical perspective, knowing about RT distribution can lead to insights about processing that are not available from mean RT alone.

References

Balota, D. A. & Spieler, D. H. (1999). Word Frequency, Repetition, and Lexicality Effects in Word Recognition Tasks: Beyond Measures of Central Tendency. *Journal of Experimental Psychology: General*, *128*, 32—55.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review*, *113*, 700—765.

Burbeck, S. L. ,&Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception and Psychophysics*, *32*, 117—133.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* New York: John Wiley & Sons.

Fific, M., Little, T. D., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117*, 309—348.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6*, 651—675.

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

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

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

Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford: Oxford University Press.

LaBerge, D. A. (1962). A recruitment theory of simple behavior. *Psychometrika, 27*, 375—396.

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

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin and Review, 16*, 798— 817.

McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology* (pp. 309—360). New York: John Wiley and Sons, Inc.

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

Ollman, R. T. (1966). Fast guesses in choice reaction time. *Psychonomic Science, 6*, 155—156.

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

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision-making during perceptual categorization: a timing diagram. *Journal of*

Neuroscience, *26*, 8965—8975.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical Recipes*. Cambridge, England: Cambridge University Press.

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 speed and accuracy of positive and negative responses. *Psychological Review*, *92*, 212—225.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510—532.

Ratcliff, R. (2001). Putting noise into neurophysiological models of simple decision making (letter). *Nature Neuroscience*, *4*, 336.

Ratcliff, R. (2002). A diffusion model account of reaction time and accuracy in a two choice brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review*, *9*, 278—291.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, Y. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightnessdiscrimination task. *Journal of Neurophysiology*, *97*, 1756—1774.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873—922.

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

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences*, *106*, 6539—6544.

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

Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: the temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General*, *139*, 70—94.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). Adiffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*, 278—289.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*, 323—341.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception and Psychophysics*, *65*, 523—535.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*, 408—424.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 127—157.

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

Ratcliff, R., & Van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin and Review*, *16*, 742—751.

Ratcliff, R., & Van Dongen, H. P. A. (2011). A diffusion model for one-choice reaction time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*, 11285—11290.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261—300.

Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review*, *102*, 567—591.

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*, 283—317.

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

Stone, M. (1960). Models for choice reaction time. *Psychometrika*, *25*, 251—260.

Swets, J. A. (1961). Is there a sensory threshold? *Science*, *134*, 168—177.

Swensson, R. G. (1972). The elusive tradeoff: speed versus accuracy in visual discrimination tasks. *Perception and Psychophysics*, *12*, 16—32.

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*, 415—429.

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

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

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011—1026.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. Journal of *Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582—600.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.), *Stevens' Handbook of Experimental Psychology (3rd Edition), Volume 4: Methodology in Experimental Psychology* (pp. 461—516). New York: Wiley Press.

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

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology, 21*, 641— 671.

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

Yap, M. J., Balota, D. A., Tse, C-S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 495—513.