The Psychological Record, 2008, 58, 475–482

# EFFECTIVE ANALYSIS OF REACTION TIME DATA

Robert Whelan
*University College Dublin*

*Most analyses of reaction time (RT) data are conducted by using the statistical techniques with which psychologists are most familiar, such as analysis of variance on the sample mean. Unfortunately, these methods are usually inappropriate for RT data, because they have little power to detect genuine differences in RT between conditions. In addition, some statistical approaches can, under certain circumstances, result in findings that are artifacts of the analysis method itself. A corpus of research has shown more effective analytical methods, such as analyzing the whole RT distribution, although this research has had limited influence. The present article will summarize these advances in methods for analyzing RT data.*

Reaction time (RT; also called response time or latency), the time taken to complete a task, has been a common dependent measure in psychology for many years. Most researchers analyze RT data by conducting an analysis of variance (ANOVA) on the sample mean (Van Zandt, 2002): this type of statistical approach may not be effective, however, owing to the particular characteristics of RT data. Statistically, RTs are treated as *random variables*: that is, observed RTs, even from the same subject in the same condition, vary somewhat across trials. Reaction times collected in a particular experimental condition are assumed to represent a sample of the population of RTs from that condition. They are assumed to be identically and independently distributed (*iid*), although this is rarely the case in practice because of factors, such as fatigue and sequential effects, that are generally assumed to be of negligible impact and are therefore ignored (cf. Thornton & Gilden, 2005). Importantly, response-time distributions are not Gaussian (normal) distributions but rather rise rapidly on the left and have a long positive tail on the right (see Figure 1). Reaction-time distributions are similar to the *ex-Gaussian* distribution (Luce, 1986), which is a convolution (mixture) of a Gaussian and an exponential distribution that has been shown to fit empirical RT distributions well (e.g., Balota & Spieler, 1999). This distribution has three parameters. The mean and the standard deviation of Gaussian (the left hump) are described by mu (μ) and sigma (δ), respectively. Tau (τ) describes both the mean and the standard deviation of the exponential component (the right tail).

Robert Whelan, School of Medicine and School of Engineering, University College Dublin.
Address correspondence to Robert Whelan, Department of Psychiatry, St. Vincent's University Hospital, Elm Park, Dublin 4, Ireland. E-mail: robert.whelan@ucd.ie

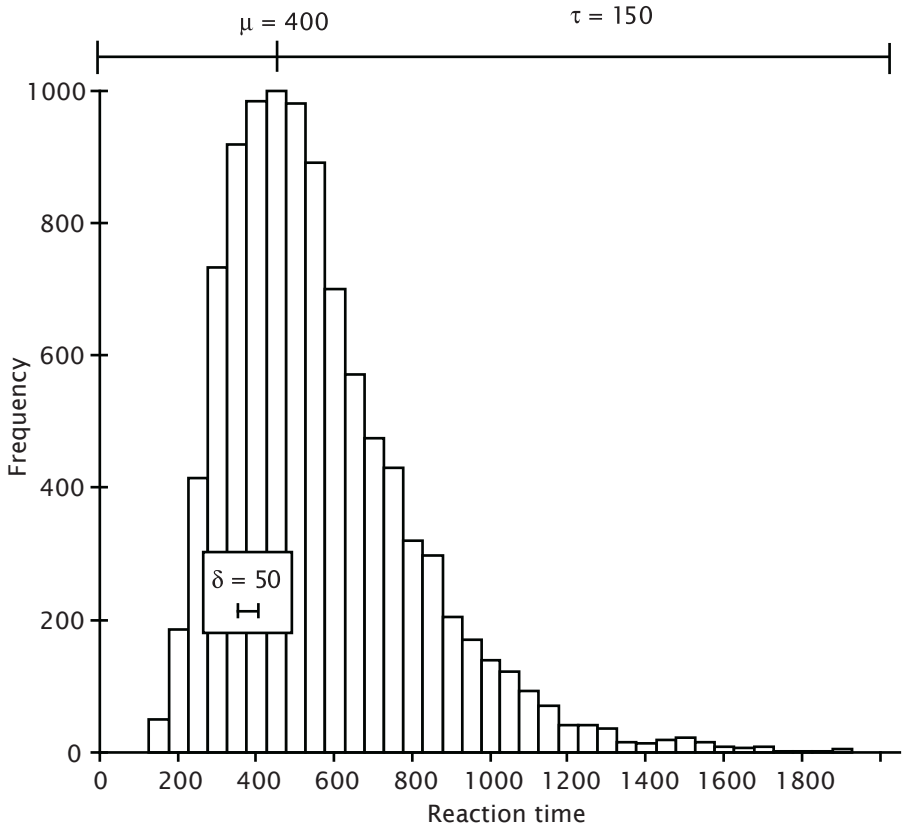


*Figure 1.* A simulated ex-Gaussian distribution showing the characteristic shape of reaction time distributions, including the parameters mu, sigma, and tau.

Typically, some observed RTs are not a result of the process of interest. For example, Luce (1986) demonstrated that genuine RTs have a minimum value of at least 100 ms: time needed for physiological processes such as stimulus perception and for motor responses. Reaction times below this value could be the result of fast guesses, for example. It is easy to identify these very fast RTs, and they are normally eliminated by using a cutoff of between 100 ms and 200 ms. Response times in the middle of the distribution that are due to spurious processes are impossible to identify, because they are intermixed with genuine RTs. There is nothing that can be done—beyond tight experimental control during the task itself—to attenuate the effects of these responses. It is quite common for some RTs to be slow because the subject is inattentive, and these RTs can strongly influence the outcome of hypothesis tests. There are a number of techniques for dealing with spurious slow RTs, and these will be discussed in more detail below in the context of the different analysis methods.

## Central Tendency Approaches

The most common method of analyzing RT data is to report a central

tendency parameter (e.g., the mean) and a dispersion parameter (e.g., the standard deviation). The mean difference in RT across conditions is then often analyzed by using ANOVA. However, using hypothesis tests on data that are skewed, contain outliers, are heteroscedastic, or have a combination of these characteristics (raw RT data typically have at least the first two) reduces the power of these tests and can result in a failure to detect a real difference between conditions (Wilcox, 1998). For example, Ratcliff (1993) demonstrated that when the difference between conditions was in mu (i.e., the distribution was shifted to the right) and the data included outliers, the ability to detect the difference with an ANOVA on mean RT was severely reduced. This type of analysis is common (Van Zandt, 2002); for example, in Dibbets, Maes, and Vossen (2002) and Fields, Landon-Jimenez, Buffington, and Adams (1995) mean raw RT was reported. To obtain a better measure of central tendency, assuming that the difference between conditions is contained in the middle 85%–95% of the RT distribution, researchers can delete some proportion of the extreme trials, transform the data, or accommodate the outliers by using parameters that are less sensitive to outliers. Ratcliff (1993) investigated the effect of these methods by using Monte Carlo simulations on RT data. No method greatly affected the number of Type I errors, but the power varied considerably across methods. The most common approaches are discussed below.

Cutoffs eliminate slow RTs by excluding data longer than some absolute time, some percentage of the data, or data that are some proportion of standard deviations above the mean. According to Ratcliff (1993), when the difference between conditions was in mu, eliminating RTs above an absolute cutoff point maintained the highest power. However, when the effect was in tau and there were no outliers, the use of absolute cutoffs reduced power because real data were eliminated. When the effect was in tau and there were outliers, then absolute cutoffs had the *potential* to increase power, although if the cutoff was too large then power was decreased because genuine RTs were eliminated. The disadvantage of this method is that no reliable rule can be used to establish absolute cutoffs because they are highly dependent on the particular data that were observed. Consequently, cutoffs are often based on the standard deviation (e.g., exclude RTs greater than two standard deviations above the mean). Ratcliff found that basing cutoffs on the standard deviation could have very adverse effects on power, depending on whether the experimental factors had their effects on the fast or slow RTs (e.g., 2 conditions might differ only in the slower responses). Furthermore, Ulrich and Miller (1994) showed that cutoffs can introduce asymmetric biases into statistics such as the sample mean, median, and standard deviation and warn against using cutoffs without allowing for these effects.

Data transformations have the potential to lessen the impact of outliers or skew, or both, by reducing larger values to a greater extent than smaller values. Transforming RTs to speed (i.e., the reciprocal of latency) normalizes the distribution somewhat, reduces the effect of slow outliers, and therefore generally maintains good power (e.g., Imam, 2006; Spencer & Chase, 1996; c.f., Greenwald, Nosek, & Banaji, 2003). Ratcliff (1993) reported that this was the next most powerful approach after cutoffs for minimizing the effects of outliers. Transforming data by using the logarithm of each RT normalizes the distribution more than the inverse transformation, although the effect of long RTs is not attenuated to the same extent as the inverse, and therefore

power is reduced relative to the inverse transformation. Researchers should be careful about transforming data, however, because it is possible to eliminate significant effects by transformation. There are also issues of interpretation after transformation of a variable, because the relationship among the variables has been changed (Osborne, 2002).

Neither the mean nor standard deviation are said to be *robust* measures. That is, the mean is not reflective of the typical response if the distribution is skewed, because the mean is distorted in the direction of the skew. The standard deviation can be greatly increased by a relatively low number of slow RTs. Therefore, many researchers report the median RT as a central tendency parameter, because it is less susceptible to departures from normality (i.e., robust). The interquartile range (the range between the third and first quartiles) is a robust method of estimating the dispersion. Ratcliff (1993) reported that using the median generally resulted in lower power than using cutoffs or transformations. This was the case when the effect was in mu or tau, and both with and without outliers. However, when there was large variability among participants the median had more power than some other methods, such as the inverse transformation.

A difficulty with using the median is that unlike the sample mean, it is a *biased* estimator of the population median when the population is skewed (a *biased* estimator does not, on average, equal the value of the parameter or function that it estimates): the true population median will, on average, be underestimated. This is not a problem when comparing conditions with the same number of trials, because the bias is approximately equal across conditions. Crucially, the bias becomes more extreme as the sample size becomes smaller (Miller, 1988) and thus the median is more likely to be overestimated in the condition with fewer trials (e.g., Bentall, Jones, & Dickins, 1998, Experiment 1). The median should never be used on RT data to compare conditions with different numbers of trials.

## Examining the Whole RT Distributions

Although analysis of the central tendency is the most popular method of analyzing RT, there are drawbacks to this approach. The mean RT can be the same even if two RT populations are genuinely different. For example, the modal part of the distribution (the left hump) can decrease in value while the number of data in the tails increases, thus producing a null effect of condition. In addition, examining the mean alone could obscure interesting details, such as the behavior of fast and slow responses across the conditions of an experiment. An increasingly popular approach is to analyze the whole distribution itself, thereby discovering effects that would otherwise have been missed.

It is perhaps easiest to describe the benefits of analyzing the whole distribution by describing some actual results from a recent study. Hervey and colleagues (2006) employed a Go/No go task to measure differences in neuropsychological performance between children with attention deficit hyperactivity disorder (ADHD) and control subjects. Reaction times faster than 100 ms were eliminated. The traditional RT measures—sample mean and standard deviation—showed that children with ADHD were significantly slower and more variable in responding than children in the control group. However, when the ex-Gaussian measures of RT were employed, a different

pattern of results emerged. Children with ADHD demonstrated *faster* reaction times on the mean of the normal component of the ex-Gaussian RT curve compared with normal controls (296 ms vs. 319 ms, respectively). However, the difference between groups was largest on the exponential part of the curve, tau, indicating that children with ADHD had a greater number of RTs that were well beyond their mean performance than did the control group (229 ms vs. 144 ms, respectively). The authors concluded that children with ADHD were not generally slower than controls but rather were prone to attentional lapses on some trials. Balota and Spieler (1999) analyzed data from a lexical decision task using this approach.

One drawback of analyzing the whole distribution is that many data points per participant and condition are required, which may be difficult to obtain in practice. Recently, Rouder, Sun, Speckman, Lu, and Zhou (2003; see Rouder, Lu, Speckman, Sun, and Jiang, 2005 for a more accessible account) have described a distribution that adequately describes RT data, even if there are relatively few data points per participant. Distributional analyses are also subject to the influence of outliers. Ratcliff (1993) reported that absolute cutoffs work well when the outliers are extreme. If, however, the outliers are intermixed with genuine RTs then the parameters will be overpredicted when using cutoffs, although trends in the data will be preserved.

## Aggregating RT Data

The most common method to aggregate RT data is parameter averaging (e.g., averaging mu, sigma, and tau across subjects). However, the situation is complicated because individual differences across measures of central tendency, dispersion, or distributional shape can be considerable (Luce, 1986). A method called *Vincentizing* is often used to address this problem. With this approach, estimates of the quantiles (the percentage of points below the given value; 0.4 quantile means that 40% of the data are below and 60% are above that value) of individual observers' distributions are aligned to produce an estimate of the aggregate distribution (see Ratcliff, 1979, for a detailed discussion of this issue). This approach is not without its limitations, however, and Rouder and Speckman (2004) have demonstrated that parameter averaging outperforms Vincentizing as sample size increases.

## Implementation of These Methods

It is easy to employ many of these analytic approaches. For instance, if the raw data are stored in Microsoft Excel format, then Excel's built-in formulas can be used to efficiently conduct simple data analyses, such as transformations (see the appendix for a list of usable formulas). However, Excel should not be used to conduct inferential statistical and some types of descriptive calculations (e.g., quartiles; McCullough & Wilson, 2002). Programs such as SPSS, SAS, and the MatLab Statistics Toolbox (http://www.mathworks.com) offer the scope to accurately and efficiently calculate the central tendency and spread of data. For example, analyzing data with the *Explore* option in SPSS will automatically produce the median, interquartile range, and 5% trimmed mean of the data. A number of software programs may be used to estimate parameters of various distributions (e.g., the MatLab Distribution Fitting Toolbox; see also Van Zandt, 2000). For example, a researcher can use

distribution-fitting software to estimate the parameters of the ex-Gaussian distribution (mu, sigma, and tau) for each subject and condition. These methods typically require some programming knowledge, however.

## Conclusion and Recommendations

A key point of the present article is that some statistical approaches, such as ANOVA on sample means, are unsuitable for RT data, because power to detect differences can be poor. Although some protocols require specific methods of analyzing RTs (e.g., Greenwald et al., 2003), some general recommendations for analyzing RT data may be made. For example, methods to allow for the effects of outliers should be applied carefully. Applying an inverse transformation can maintain high power under many situations. An alternative, and perhaps superior, approach to analyzing central tendencies is to examine the whole distribution, because this method can yield valuable information about the differences between conditions. In view of the time and effort required to design an experiment and collect data, the use of more advanced analysis methods can maximize the return from the obtained data.

## 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.

BENTALL, R. P., JONES R. M., & DICKINS, D. W. (1999). Errors and response latencies as a function of nodal distance in 5-member equivalence classes. *The Psychological Record, 49*, 93–115.

DIBBETS, P., MAES, J. H. R., & VOSSEN, J. M. H. (2002). Contextual dependencies in a stimulus equivalence paradigm. *Quarterly Journal of Experimental Psychology, 55B*, 97–119.

FIELDS, L., LANDON-JIMENEZ, D. V., BUFFINGTON, D. M., & ADAMS, B.J. (1995). Maintained nodal distance effects after equivalence class formation. *Journal of the Experimental Analysis of Behavior, 64*, 129–146.

GREENWALD, A. G., NOSEK, B. A., & BANAJI, M. R. (2003). Understanding and using the Implicit Association Test I: An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.

HERVEY, A. S., EPSTEIN, J. N., CURRY, J. F., TONEV, S., ARNOLD, L. E., CONNERS, C. K., HINSHAW, S. P., SWANSON, J. M., HECHTMAN, L. (2006). Reaction time distribution analysis of neuropsychological performance in an ADHD sample. *Child Neuropsychology, 12*, 125–140.

IMAM, A. A. (2006). Experimental control of nodality via equal presentations of conditional discriminations in different equivalence protocols under speed and no-speed conditions. *Journal of the Experimental Analysis of Behavior, 85*, 107–124.

LUCE, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

MCCULLOUGH, B. D., & WILSON, B. (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics & Data Analysis, 40*, 713–721.

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

OSBORNE, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, 8(6). Retrieved April 23, 2007, from http://PAREonline.net/getvn.asp?v=8&n=6

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

RATCLIFF, R. (1993). Methods of dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–532.

ROUDER, J. N., LU, J., SPECKMAN, P., SUN, D., & JIANG, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 199–223.

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

ROUDER, J. N., SUN, D., SPECKMAN, P. L., LU, J., & ZHOU, D. (2003). A hierarchical Bayesian statistical framework for skewed variables with an application to response time distributions. *Psychometrika, 68*, 589–606.

SPENCER, T. J., & CHASE, P. N. (1996). Speed analyses of stimulus equivalence. *Journal of the Experimental Analysis of Behavior, 65*, 643–659.

THORNTON, T. L., & GILDEN, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review, 12*, 409–441.

ULRICH, R., & MILLER, J. (1994). Effects of truncation of reaction time analysis. *Journal of Experimental Psychology: General, 123*, 34–80.

VAN ZANDT, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review, 7*, 424–465.

VAN ZANDT, T. (2002). Analysis of response time distributions. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology* (3rd ed., pp. 461–516). San Diego, CA: Academic Press.

WILCOX, R.R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*, 300–314.

## Appendix: Formulas and Results for Analyzing RT Data with Microsoft Excel

Examples of transformations and cutoffs are given for cell A1. Data that are cut off are replaced with text, thereby excluding the contents of that cell from subsequent analyses. The central tendency and dispersion calculations are given for cells A1:A10. The formulas should be entered into cells other than those in which the raw RTs are stored (e.g., enter the formulas in column B).

| Desired Result | Formula |
|---|---|
| Mean | = AVERAGE(A1:A10) |
| Standard Deviation | = STDEV(A1:A10) |
| Cutoff RTs < 200 ms | = IF(A1<200,"FAST",A1) |
| Cutoff RTs > 2500 ms | = IF(A1>2500,"SLOW",A1) |
| Cutoff RTs > 3 standard deviations | = IF(A1>(STDEV(A1:A10)*3),"SLOW",A1) |
| Speed (1/RT) | = 1/A1 |
| Log (base 10) | = LOG10(A1) |
| Median | = MEDIAN(A1:A10) |
| Interquartile range | Do not use Excel |