Journal of Clinical and Experimental Neuropsychology
2002, Vol. 24, No. 7, pp. 858–872

1380-3395/02/2407-858$16.00

Interpreting Reaction Time Measures in Between-Group Comparisons

Timothy A. Salthouse[1] and Trey Hedden[2]
[1]Department of Psychology, University of Virginia, VA, USA, and [2]Department of Psychology, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Although reaction time measures have been used extensively in many types of between-group comparisons, the assumptions and limitations of reaction time measurement are not always recognized. In this article we discuss three issues that should be considered when designing and interpreting comparisons involving reaction time. These concern speed-accuracy tradeoffs, methods of analyzing measures postulated to reflect specific processes, and methods for distinguishing group-related influences that are shared with other variables from those that are unique to a single variable.

INTRODUCTION

This article is concerned with issues relevant to the use and interpretation of reaction time (RT) measures in comparisons of different groups. The discussion is fairly abstract because it is intended to apply to many possible groups – those defined by individual difference classifications such as age, gender, and culture, as well as by criteria such as psychopathological or neuropsychological status.

Why would researchers be interested in using RT procedures in comparing people from different groups? One major reason is that RT appears to be simple and easy to measure. The researcher merely presents a stimulus and registers a response to it, with the interval between the two events representing the RT. Furthermore, RT appears to be precise and quantitative, with properties of the highest scale of measurement (i.e., ratio scale with a true zero).

Pachella (1974) has noted that another reason for the popularity of RT measurements is that the only property of mental events that can be studied while they are occurring is their duration. This is no longer true with the development of on-line eye movement recordings, evoked potential techniques, and functional neuroimaging, but RT procedures are certainly among the simplest and least expensive methods available for on-line assessment.

However, RT measures are deceptively complex, and they may reflect much more, or much less, than what the researcher assumes. Even what appear to be simple variables have multiple determinants, and it is not always easy to identify which of the potential determinants of the observed group differences is of greatest importance. Furthermore, what might seem to be intuitively obvious methods of comparison may, upon closer examination, reveal surprisingly serious limitations. In this article we focus on three issues that we believe should be considered in the design, analysis, and interpretation of RTs in between-group comparisons. These are: speed-accuracy tradeoffs, methods of within-task comparisons, and analytical procedures for between-group

Address correspondence to: Timothy Salthouse, Department of Psychology, 102 Gilmer Hall, P.O. Box 400400, University of Virginia, Charlottesville, VA 22904-4400, USA. Tel.: +1 434 243 5034. Fax: +1 434 982 4766. E-mail: salthouse@virginia.edu
Accepted for publication: September 22, 1999.

comparisons. Other important issues concerned with the use of RT measures could obviously be identified, but we have concentrated on these because they pose special problems in connection with the interpretation of group differences in RT measures.

SPEED-ACCURACY TRADEOFFS

The typical instructions in RT tasks are to respond as rapidly and as accurately as possible. However, these are often incompatible requirements because if the responses are very fast then there are frequently many errors, and if precautions are taken to avoid any errors then the responses are often relatively slow. One way of representing these interrelations is in terms of a tradeoff function relating speed, along the abscissa, to accuracy, along the ordinate.

Consider the two different speed-accuracy tradeoff functions portrayed in Figure 1, which could be based on data from different people or groups, or from different experimental conditions within a single person or group. The function on the left, represented by points Al, A2, A3, and A4, is faster at every level of accuracy than the function on the right, represented by points B1, B2, B3, and B4. However, inspection of the figure reveals that the interpretations can become quite complicated if people can operate at different positions on their respective functions. For example, even though the points on the left function generally represent faster performance than the corresponding points on the right function, the locations of individual points vary, and comparisons would be misleading if both dimensions of performance were not considered. Even comparisons at 100% accuracy would not necessarily be meaningful because the 100% points may not be at the same distance from the hypothetical optimum RT corresponding to the fastest time at 100% accuracy. To illustrate, points A4 and B4 are both at 100% accuracy, but A4 is much closer to the optimum for the A function than B4 is to the optimum for the B function.

Unfortunately, there is no simple solution to what has come to be known as the speed-accuracy tradeoff problem. In the following paragraphs we will briefly review several strategies that have been proposed for dealing with the problem, and describe some advantages and disadvantages of each.

First, deletion of error trials ensures that detected errors are not included in the reported results, but it does not necessarily solve the speed-accuracy tradeoff problem. That is, people could still differ in the number of errors committed, and therefore the meaning of the RTs in the remaining correct trials (some of which were probably

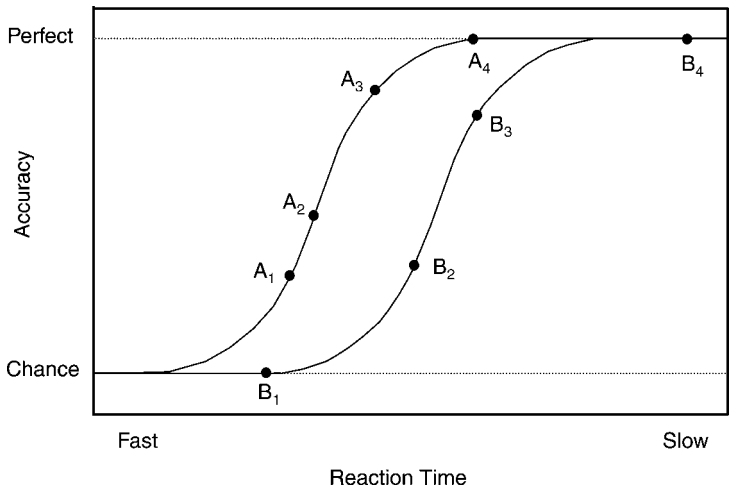

Fig. 1. Hypothetical speed-accuracy tradeoff functions comprised of 4 points (1 through 4) in each of two groups or conditions (A and B).

correct by chance) is ambiguous. Merely because the overt errors are eliminated from the analyses does not mean that the remaining RTs reflect the true duration of the relevant processing because the guesses that were correct by chance are still represented in the data.

A second possible strategy is to analyze RT and errors in separate analyses, or possibly even neglect errors completely. The primary problem with this approach is that the researcher is essentially treating intrinsically bivariate data as univariate, because one dimension of performance is ignored when considering the other dimension. An illustration of how this can lead to erroneous conclusions is evident in Figure 1. Note that if the focus is only on RT then B2 is faster than A4, whereas if the focus is exclusively on accuracy then A4 is higher than B2. Quite different conclusions about the relation between A and B could therefore be reached depending on which dimension is considered, and which neglected. Treating the dimensions separately may also have the consequence of converting a ratio scale measurement into a weaker interval or ordinal scale because only crude comparisons are possible for either variable when the variables are on different relative positions on the tradeoff functions. Moreover, dismissing the possibility of a tradeoff on the grounds that the error rates were low can be misleading because the variation in RT may actually be greatest when the error rates are low, as is the case in the nonlinear functions portrayed in Figure 1. It is worth noting, however, that if the patterns are similar in analyses of RT and errors, the researcher is probably justified in concluding that the differences are not attributable to a speed-accuracy tradeoff even though precise quantitative comparisons of the magnitude of the differences still may not be meaningful.

Another strategy that has been pursued in attempting to deal with the speed-accuracy tradeoff problem involves determining the overall correlation between RT and accuracy, and then dismissing the problem if the correlation is not statistically significant, or if it is negative rather than positive. A limitation of this approach is that even small relations between RT and accuracy, in either direction, could distort the performance comparisons. Once again referring to Figure 1, if only points A2 and B3 are considered then the relation between time and accuracy is positive (i.e., B3 is slower and more accurate than A2), whereas if only points A3 and B2 are considered the relation between time and accuracy is negative (i.e., B2 is slower but less accurate than A3). However, quantitative comparisons in both cases are ambiguous because it is not obvious how much the differences in the RT axis should be adjusted to account for the differences in the accuracy axis, and vice versa. The researcher could be confident that overall performance in A3 is superior to that in B2 because it is both faster and more accurate, but the precise amount by which it is superior cannot be determined without more detailed information about the relations between RT and accuracy in each group.

Analysis of covariance is another possible solution that has been proposed for the speed-accuracy tradeoff problem, typically by using accuracy as a covariate when analyzing group differences in RT. This approach examines speed and accuracy simultaneously, thus allowing the variation in one variable to be controlled when examining effects on the other variable. However, this approach has two potentially serious limitations. First, the relation of accuracy to RT may not be linear (possibly because, as in the functions portrayed in Fig. 1, small accuracy differences may have greater impact on RT at the extremes of the function than in the middle). And second, the nature of the adjustment equation for one group may not be the same as that for the other group. Only if the complete functions are available from every individual can the equivalence of the functions be examined, and if the entire functions are available then more powerful comparisons can be employed.

Another possibility is to create a composite index by treating RT and accuracy as two indicators of a latent performance construct. For example, both variables could be converted to z scores, and then analyses conducted on the average, or the sum, of the z scores for the RT and accuracy variables. (Before computing the composite it is important to ensure that both variables are scaled in the same direction such that high scores correspond to worse, or better, performance in each variable.) This method has the

advantage of taking both speed and accuracy into consideration with a single variable, but it is based on the questionable assumption that the two aspects are equal in importance. Furthermore, because of the conversion to *z*-score units, the scale of measurement is no longer ratio, or possibly even interval, and thus one of the primary reasons for the appeal of RT measures is lost.

Still another method of dealing with the two dimensions of performance within the same task involves determining the ratio of accuracy over time as a measure of throughput rate. This method, which is used in sorne psychometric tests, has the intuitive appeal of yielding a single variable that might be postulated to reflect processing efficiency in the sense of the level of accuracy achieved per unit time. Unfortunately, this procedure is not easily adapted to traditional RT measures. Not only would some adjustment be needed for chance responding in multiple-choice situations, but information about the entire tradeoff function would be necessary to allow meaningful comparisons. For example, assume that accuracy was 80% and RT was 400 msec in one condition, and that accuracy was 96% and RT was 600 msec in another condition. The corresponding ratios would be 0.20%/msec and 0.16%/msec, respectively, implying that the former condition was superior to the latter condition. However, this conclusion and its apparent precision could be misleading because comparisons with respect to which condition has the greater overall efficiency depend on the relations between speed and accuracy in each condition. To illustrate, it is possible that these points actually fall along the same speed-accuracy function, in which case they would reflect equivalent capabilities.

What is almost certainly the most informative solution to the speed-accuracy tradeoff problem involves generating functions representing the relation between RT and accuracy across a wide range of accuracy values for each individual. If the complete functions are available, then the tradeoff problem can be eliminated by making comparisons of time at a fixed (nonchance and nonperfect) level of accuracy, or by making comparisons of accuracy at a fixed level of time.

Three basic methods have been used to generate these types of tradeoff functions (but see Meyer, Osman, Irwin, & Yantis, 1988; Wickelgren, 1977, for additional variants on these methods). One method involves the use of instructions or payoffs to induce varying emphases across different blocks of trials. For example, there could be a high reward for fast responses and a low cost for errors in one block of trials (speed emphasis), a low reward for fast responses and a high cost for errors in another set of trials (accuracy emphasis), and one or more trial blocks with intermediate (mixed emphasis) rewards and costs.

A second method attempts to obtain different combinations of speed and accuracy by the use of RT deadlines or response windows. That is, across trial blocks the value of a deadline before which the response is to be emitted, or the temporal window within which the response is to be produced, could be varied. By suitable manipulation of the RT deadlines or windows, the responses can be slow and presumably very accurate in some trials, in other trials they must be fast even if they are less accurate, and in still other trials they can be intermediate in both speed and accuracy.

The third method that has been used to generate complete speed-accuracy functions is based on post hoc analyses of RTs within a single block of trials. In this method a wide range of RTs is obtained, often by instructing the research participant to attempt to respond with a moderately high error rate. The RTs are then ordered by time, and the level of accuracy determined within each range of RT values to allow functions to be generated relating RT to accuracy.

With each of these methods the goal is to obtain functions resembling those in Figure 1 that would allow comparisons of one variable (e.g., RT) to be made at a fixed value of the other variable (e.g., accuracy). Although the analysis of complete speed-accuracy functions derived for each individual appears to be the best available method of simultaneously considering both RT and accuracy, it also has some important limitations. Among these are that the generation of complete speed-accuracy functions is very time-consuming because a large number of RTs is needed to provide precise estimates at several levels of accuracy, and if there are multiple conditions in the study then separate speed-accuracy tradeoff functions will need to be generated in each

condition. Furthermore, some individuals may be reluctant to respond with low levels of accuracy, with the consequence that their functions would be incomplete.

To summarize, researchers considering the use of RT measures need to be aware of the fact that performance variations can be manifested in RT, in accuracy, or in both RT and accuracy, and they should design and carry out analyses that take potential tradeoffs between these two aspects of performance into consideration. Ideally, complete speed-accuracy functions would be determined, but this is often not feasible in certain situations or with particular group comparisons. Where this is not possible, several analytical techniques might be applied in conjunction to overcome weaknesses of any single technique. Our recommendation is to conduct separate analyses of each variable, analyses on RT with accuracy as a covariate, and analyses of composite scores. If the results are consistent across each of these methods then the researcher can be confident in a conclusion that there was a real difference in the performance capabilities between the groups.

METHODS OF WITHIN-TASK COMPARISONS

Virtually all RT tasks can be assumed to involve a number of different processes, and when only a single measure is available it is not clear which particular processes are contributing to any observed group differences in RT. One way to be more specific about the nature of the factors responsible for overall differences in RT involves the presentation of multiple conditions within a task, followed by comparisons of the RTs in the different conditions. However, even though variants of this procedure are widely used, it is not as simple as sometimes assumed.

Techniques for fractionating RT, sometimes known as mental chronometry, have had a long history (e.g., Meyer et al., 1988). In fact, the most common technique, known as the subtraction method, was originally introduced by Donders in the 1800s (Donders, 1868/1969). This method consists of attempting to isolate the duration of a critical process by comparison of the RTs in conditions presumed to differ in the presence of a single stage or process.

The subtraction procedure has been widely used, and its applications have extended beyond RTs. To illustrate, most comparisons of the Trail Making Test (Reitan, 1992) used for neuropsychological assessment are based on the subtraction method. There are two conditions in this test, with the examinee instructed to connect targets in numerical sequence (condition A), or in alternating numerical and alphabetical sequences (condition B), as rapidly as possible. The difference in time (or errors) between the two conditions is often interpreted as a reflection of the influence of executive processes concerned with task switching, monitoring, and planning presumed to operate in the B version but not the A version (e.g., Salthouse & Fristoe, 1995).

An example of the subtraction method with RT tasks is a task introduced by Posner, Boies, Eichelman, and Taylor (1969) known as the NI-PI (Name Identity-Physical Identity) task. The research participant in this situation is instructed to classify two visually presented letters as same or different as rapidly as possible, with RT in each condition serving as the primary dependent variable. If both letters are the same case then the decision can be made simply on the basis of the physical identity of the letters. However, if one letter is in lower case and the other in upper case, then the matching decision requires access to the names of the letters. The additional time for name identity decisions compared to physical identity decisions has therefore been interpreted as an estimate of the duration required to gain access to the letter name. This is an example of the subtraction procedure because the name match decision is assumed to involve all of the processes involved in the physical match decision with the addition of name access, and the duration of the additional process is presumably what is responsible for the longer RTs in the name identity condition compared to those in the physical identity condition.

Although the subtraction method has enjoyed considerable popularity, it has at least two major limitations. First, the method requires knowledge of, or strong assumptions about, the identity and sequence of processes involved in each of the

relevant conditions of the task. That is, unless the researcher has a pretty good idea of the processes required to perform each of the conditions, and in particular of the processes that differ between conditions, then the results of the procedure may not be interpretable. Second, the method is based on an assumption that the conditions are identical except for the addition of a critical process (i.e., the assumption of ''pure insertion''). This implies that exactly the same strategy is used to perform the two versions of the task, and that the addition of the critical process does not alter the identity or the efficiency of any other processes involved in the task.

Unfortunately, these assumptions may not always be valid. Consider the case of the Trail Making Test described earlier. Because version B is always performed after version A in the traditional administration of the test, performance in version B may be susceptible to practice or fatigue effects that are not present in version A, creating a confound of condition and order. Furthermore, version B involves letters and a less familiar alphabetic sequence compared to the numbers and familiar numeric sequence used in version A. Finally, the two versions differ in the arrangement of targets on the page, and thus the direction and magnitude of the movements between successive targets are not necessarily equivalent. Any or all of these differences could be contributing to time and error differences between the two versions, in which case the poorer performance in version B may not be simply attributable to the added requirement of switching between sequences and monitoring the positions within each sequence.

Several extensions or modifications of the subtraction technique have also been proposed for fractionating RT. One is the parametric variation method in which the independent variable is manipulated quantitatively, and then the relation between RT and the quantitative value of the manipulated factor is examined. The parametric variation method is similar to the subtraction method but with three or more conditions instead of just two. However, this method is often considered superior to the simple subtraction method because it may be more plausible to interpret the differences among the conditions in terms of the operation of a single variable if there is an orderly relation among the RTs in the multiple conditions. When the relations are linear this method yields intercept and slope values, which are roughly analogous to the initial value and the difference score in the subtraction method. However, because the slope and intercept parameters are only meaningful for a given individual if the linear regression equation used to compute them provides a good fit to the data, an index of the degree to which the equations fit each individual's data should also be examined whenever the parametric variation method is used.

Two very well-known RT tasks based on the parametric variation method have inspired a considerable amount of research. One extremely influential paradigm is Sternberg's (1966) memory scanning task. In this procedure the experimenter varies the number of items in a memory set and then presents a single probe item, with the research participant instructed to determine as rapidly as possible whether the probe item was a member of the previously presented memory set. Under these circumstances RT typically increases linearly with the number of items in the memory set. The slope parameter of the linear regression equation has been interpreted as reflecting the rate of memory search or comparison, and the intercept parameter as representing the duration of all other processes.

A second example of parametric variation is the mental rotation task of Shepard and Metzler (1971). In this procedure the experimenter visually presents a pair of items at different angular orientations relative to one another, with the research participant instructed to determine as rapidly as possible whether the two items are the same object or are mirror images of one another. RT has usually been found to increase linearly with the angular discrepancy between the items, with the slope parameter interpreted as an estimate of the rate of mental rotation, and the intercept parameter as a reflection of the duration of all other processes.

Another variant of the subtraction method is the additive factors procedure introduced by Sternberg (1969). This procedure is based on the assumption that overall RT reflects the processing of a linear sequence of independent stages. It is

further assumed that manipulations (or factors) that affect separate stages should have independent or additive effects on RT, but manipulations or factors that affect the same stage should have interactive effects. The additive-factors method has been used in between-group comparisons to attempt to localize group-related effects to particular stages by determining whether the group variable interacts with manipulations postulated to influence a particular stage.

Two of the critical assumptions of the additive factors method are that there is a linear sequence of discrete processing stages, and that there are no partial products carried from one stage to the next. These assumptions have been challenged by cascade models in which processing can occur in parallel, with partial products available before all processing from the prior stage has been completed (McClelland, 1979). It is also important to note that unlike the subtraction and parametric variation methods, the additive factors method does not yield estimates of the durations of the component processes which would be informative about the relative contribution each component makes to the overall RT differences.

The preceding review indicates that a number of techniques are available to fractionate RT and yield measures presumed to be informative about the efficiency of particular stages or processes. If the relevant conditions are administered to members of different groups, then more specific and precise inferences about the nature of the differences between groups should be possible. However, the researcher must still determine which analytical methods to employ in making within-task comparisons.

In order to make the following discussion a little more concrete, we will assume that the comparison of primary interest is a contrast between two conditions, A and B. Condition A is postulated to involve processes 1 and 2 (e.g., encoding and response), and condition B is postulated to involve processes 1, 2, and 3 (e.g., encoding, response, and transformation). (The same logic will apply if there are three or more conditions, and analyses based on a slope parameter instead of a difference score, because the slope and difference score are conceptually similar and identical when there are only two conditions.) The important question in choosing between analytical procedures is: what is the best method of estimating the duration of process 3?

The procedure implied by the subtraction method simply involves the computation of a difference score by subtracting the RT in condition A from the RT in condition B, that is, B − A. Although this is clearly an intuitively plausible method of estimating the duration of the process or processes presumed to differ between conditions, it has two important limitations. First, the difference score will often be positively correlated with the initial (or average) value for purely statistical reasons (see Chapman & Chapman, 1988; Cohen & Cohen, 1983, for mathematical proofs). In the context of the current example, this means that the estimate of the duration of process 3 derived from the difference score is not independent of the duration of processes 1 and 2. This is clearly undesirable because the difference score is typically postulated to reflect only what differs between conditions.

The second limitation of difference scores is that they often have low levels of reliability, and hence may not exhibit relations to other variables, such as group classification, because of inadequate proportions of systematic variance. The reasons for the potentially low reliability are evident when considering the formula for estimating the reliability of a difference score:

$$\text{Est. reliability } (\text{B} - \text{A}) = \{[(r_{\text{AA}} + r_{\text{BB}})/2] - r_{\text{AB}}\}/1 - r_{\text{AB}}.$$

Examination of this formula reveals that the estimated reliability of the difference score decreases as the r_{AB} correlation increases. This occurs because the correlation (r_{AB}) reflects how much of the variance in one variable (e.g., A) is accounted for by the other variable (e.g., B). Therefore, as the correlation between the two variables increases, there is a decrease in the residual or unexplained variance in either variable, and it is this residual variance that is reflected in the difference between the two scores.

Salthouse and Coon (1994) also noted that if the correlation between the two variables is very high, then there is little evidence that the two variables actually represent distinct constructs.

That is, only if the correlation between the two variables is substantially lower than the respective reliabilities would there be evidence that the variables, and the difference between them, reflect separate constructs (i.e., exhibit discriminant validity). It is important to emphasize that the relevant information in this context is not the mean values of the variables, but instead the correlation between them. That is, the variables may differ in absolute magnitude because of the addition of one or more processes, and yet they could still share all of their reliable variance if there is little or no individual difference variability in the added processes.

Cohen and Cohen (1983, pp. 414–421) recommended an alternative to difference scores based on analyses of residuals. Their recommended procedure involves two steps. First, the contribution of processes 1 and 2 to variable B is estimated by predicting B from A, that is, $B' = a + b(A)$. And second, the residual $(B - B')$ is derived by subtracting the predicted B value (B', reflecting processes 1 and 2) from the observed B value (B, reflecting processes 1, 2, and 3). Because all of the linearly related effects of A are removed by the regression equation, the residual method eliminates the problem of the dependence of the derived score on the initial value.

Unfortunately, the residual method does not necessarily solve the problem of potentially low reliability of the derived score, or the concern that the constructs represented by the observed variables may not be distinct. Consider the consequences of increasing the r_{AB} correlation, which in the case of a difference score serves to reduce its effective variance, and to reduce the expected reliability. Increases in the correlation between variables A and B means that they share a greater proportion of variance, and thus the residual variance in B that can be attributed to process 3 (along with measurement error) will tend to decrease. This decreased variance could in turn reduce the estimated reliability because of a restriction-of-range phenomenon, and also lead to questions about the distinctiveness of the construct represented by the residual score if it has a small amount of unique variance.

To summarize, a variety of methods have been used to attempt to fractionate RT, but there are still questions about the best methods of obtaining measures of the durations of the isolated processes. The use of difference scores is intuitively appealing, but it is problematic because of the relation of the difference score to the baseline score, and potentially low reliability. The method based on estimating residuals has certain advantages, but it may also suffer from low reliability if the correlation between the initial values is high. Fortunately, some of the concerns about reliability of derived scores might be resolved if direct estimates of reliability (see below) are found to be acceptably high.

ANALYTICAL PROCEDURES FOR BETWEEN-GROUP COMPARISONS

A very important, but all too frequently neglected, requirement for meaningful between-group comparisons is that the scores being compared are reliable. Reliability corresponds to the proportion of variance in the variable that is systematic, and thus when this proportion is low the magnitude of the relation the variable can have with other variables, such as group membership, is severely restricted. More precisely, because reliability corresponds to the proportion of systematic variance in a variable, the square root of the reliability coefficient represents the largest correlation that can be expected with another variable if all of the reliable variance in the target variable is shared with the other variable. These considerations imply that, whenever possible, direct estimates of the reliability should be obtained for all variables in which comparisons are to be made. Perhaps the simplest way of assessing reliability of RT measures consists of dividing the set of trials into two, and basing the estimate of reliability on the correlation between the corresponding variables in the two sets. (Because the researcher is typically interested in estimating the reliability of the variables based on all trials, the correlation between the two sets of trials, which is equivalent to a test-retest reliability coefficient for only one set of trials, should be boosted by the Spearman–Brown formula.)

One of the primary determinants of reliability is the number of trials in each relevant condition.

There is no magical number of trials that will yield high reliability and ensure meaningful comparisons, but with the proviso that there are diminishing returns, larger is usually better. As a general rule, 5 trials is almost certainly too few, and 500 trials in the same condition will often be excessive unless speed-accuracy tradeoff functions are to be generated. A reasonable compromise might consist of between 50 and 200 trials in each condition.

Another important question to consider in between-groups comparisons is which RT variables should be analyzed? Group differences can be manifested in different ways, and it is often possible to examine several of them with the same RT data. Virtually all published comparisons report some type of central tendency measure, with the arithmetic average or mean the most common. However, the median is sometimes preferred because it is less sensitive to occasional very slow RTs that can result in severe distortions of the mean. Properties of the distribution of RTs such as variability (range), skewness (asymmetry), and kurtosis (peakedness) can also be examined to provide additional information about the nature of the group differences. Recently theory-based decomposition of RT distributions have also been explored in between-group comparisons. As an example, ex-Gaussian analyses are based on the assumption that the overall RT distribution is composed of a convolution of exponential and Gaussian distributions. Researchers willing to accept these assumptions have used mathematical algorithms to attempt to separate the distributions, and obtain parameters reflecting properties of each distribution that have then been compared across groups (e.g., Spieler, Balota, & Faust, 1996).

A somewhat related procedure with fewer theoretical assumptions simply consists of comparing the groups across different percentiles of each individual's RT distribution. That is, the RT values at successive percentiles (e.g., 10th, 25th, 50th, 75th, 90th, etc.) of each individual's RT distribution are determined, and then the averages compared across groups. Comparisons of this type are especially appropriate for investigating particular hypotheses, such as the proposal that lapses of attention are a major contributor to group differences, because this would imply that the largest group differences would be expected among each individual's slowest RTs. Results of these types of percentile analyses suggest that this is apparently not the case in comparisons of adults of varying ages because the group differences were similar across all percentiles of the RT distribution (Salthouse, 1993).

Given that the variables to be analyzed have been identified, what analytical procedures should be used in between-group comparisons? The most common analytical method is some version of a group-by-variable analysis of variance, which in its simplest form might involve two groups (1 and 2) and two variables (A and B). The interaction in this analysis is often of greatest interest, because when it is significant the researcher is likely to interpret it as evidence of a specific group-related effect on the processes involved in variable B but not also involved in variable A.

However, questions arise concerning the interpretation of interactions when the groups being compared also differ in their average or baseline RTs. A familiar principle in statistics is that the presence of an interaction qualifies the interpretation of main effects, but in some circumstances the reverse may also be true in that the presence of main effects could qualify the interpretation of interactions. That is, because it is frequently the case that the poorer performing group (either in terms of RT or accuracy) has a greater absolute difference between conditions than the better performing groups, this may not always signify a specific deficit.

One method that has been proposed to separate specific effects from effects associated with a different baseline is similar to the residuals method described above. In the version of this method described by Chapman, Chapman, Curran, and Miller (1994), a regression equation is used to determine the relation of the difference score to the overall latency in the normal or control group, that is,

$$(\mathrm{B} - \mathrm{A})' = b(\mathrm{B} + \mathrm{A}) + a.$$

Next the parameters of that regression equation are used to compute the residual difference in each group, that is,

$$(\mathrm{B} - \mathrm{A}) - (\mathrm{B} - \mathrm{A})' = (1 - b)\mathrm{B} - (1 + b)\mathrm{A} - a.$$

Because the regression procedure removes the relation of the overall RT (i.e., B + A), the adjusted difference score indicates the extent to which the difference score deviates from what would be expected on the basis of the overall latency.

Effective use of this method requires moderately large samples to obtain stable estimates of the regression parameters, and a wide range of baseline RTs in the normal or control group in order to have sufficient overlap of RTs with the other group so that the regression parameters can be applied with minimal extrapolation. The major limitation of this method is that the adjustment procedure is based only on the relation between the difference and the overall RT in one group, and it ignores any relations that might exist among sets of RT variables across the two groups. However, if this is the only relation of interest for a given hypothesis then this method can provide interpretable results.

Another analytical procedure that is sometimes employed in between-group comparisons consists of converting the observed RTs into ratios or log-transformed scores prior to carrying out the group comparisons. (Ratios and log-transformed scores are very similar because equal ratios in the original units correspond to equal absolute differences in the $\log_{10}$ scale.) Transformations such as these are based on the assumption that the slower group is slowed by the same relative amount for all processes, and that all differences between groups are multiplicative and not additive. These assumptions are represented by the following equations,

$$B2 = xB1 + 0,$$

and

$$A2 = xA1 + 0,$$

where A and B are different variables, the numbers refer to different groups, x is a global slowing factor, and the 0 indicates that there are no additive effects. These equations therefore imply that the individuals in Group 2 are slower than the individuals in Group 1 by factor x on each variable. To the extent that these assumptions are valid, and all group differences are attributable to a single global influence, then the ratios B2/B1 and A2/A1 should be equal to the same value, namely x. Given these assumptions, any group-by-variable interactions evident with log-transformed scores, or significant differences between variables evident with ratio scores, could presumably be interpreted as reflections of specific, or at least nongeneral, group-related effects.

The plausibility of the preceding assumptions can be examined by determining if, and how, RTs for a range of variables in two groups are systematically related to one another. Figure 2 portrays a hypothetical, but plausible, relation among four variables obtained from each of two groups. Each point in the figure corresponds to the RT in a

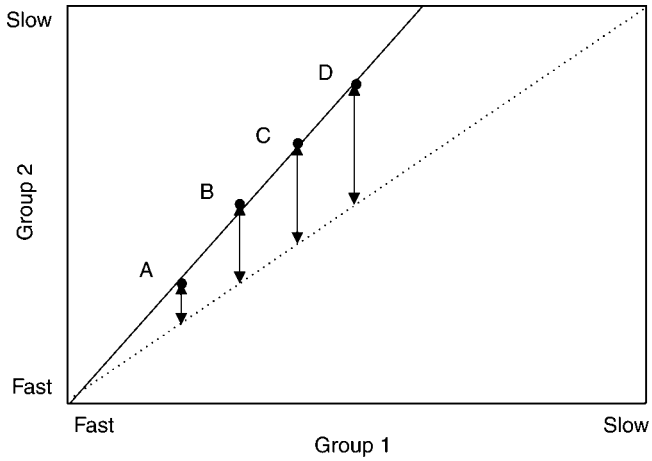

Fig. 2. Illustration of possible systematic relation between the RTs in different tasks (A through D) from two groups (1 and 2).

particular task, with the abscissa representing the time scale for Group 1, and the ordinate representing the time scale for Group 2. The solid line is the regression line relating the values in the two groups. Note that an interaction will likely be detected if a researcher is primarily interested in a contrast between variables A and B because the length of the vertical arrow, which represents the absolute RT difference in the variable between the groups, is longer for variable B than for variable A. However, this apparent interaction does not necessarily represent a specific group-related difference on the processes involved in variable B and not in variable A. That is, because there is a systematic relation between RTs in the two groups such that the differences between the RTs increase with increases in average RT, it is conceivable that almost any manipulation that increases overall RT will lead to increases in the differences between the groups, regardless of the type of processing involved. To the extent that the RTs from different types of tasks fall along the same systematic function, therefore, it may be more plausible to infer that the groups differ in some type of global respect rather than in terms of processes specific to particular tasks.

There are now a large number of studies in which systematic relations such as the one portrayed in Figure 2 have been examined with groups consisting of children or adults of different ages (e.g., Cerella, 1985; Kail, 1991; Salthouse, 1985), normals and patients with Alzheimer's Disease (e.g., Nebes & Brady, 1992), normals and patient with Multiple Sclerosis (Kail, 1998), individuals under sober and intoxicated conditions (Maylor & Rabbitt, 1993), and so forth. Most of the studies have found the empirical functions to be linear, although they are sometimes best described by more complex functions (e.g., power or exponential), and the intercepts are frequently negative rather than zero. These properties are noteworthy because the use of ratios or log-transformed RTs will not provide the appropriate correction for general influences when, as has often been found to be the case, the intercepts are not zero and the functions are not linear.

However, if a systematic relation does exist between the RTs in different groups, there are at least two additional ways in which general and specific group-related effects might be distinguished. For example, if the regression equation describing the systematic relation is assumed to represent the general differences between the two groups, statistical tests (e.g., standardized residuals, Cook's D) could be conducted to detect deviations or outliers from that regression equation. If the deviations from the predicted values are statistically significant, the researcher could then conclude that the relevant variables have effects that are larger (or smaller) than those expected on the basis of the overall relation between the two groups.

A second way in which the parameters of the systematic relation could be employed is to use the regression equation parameters to simulate the performance of one group of participants, and then rely on statistical significance tests to assess the accuracy of the resulting predictions (Madden, Pierce, & Allen, 1992). For example, if the regression equation relates the performance of older adults to that of young adults, the RTs of individual young adults could each be transformed by the regression equation parameters to create RT values for a simulated sample of older adults. If a statistically significant discrepancy is detected between the actual and simulated values for a variable, the researcher might then conclude that the variable has effects that are larger (or smaller) than those expected on the basis of the overall relation between the two groups. In other words, the group difference in that variable can be inferred to reflect something different than the general effects that are operating on the other variables.

As with virtually any analytical procedure, methods based on the existence of systematic relations have a number of limitations. Perhaps the most serious limitation is the requirement for a range of RT tasks (or variables), involving a variety of different types of processes, that are all administered to the same individuals. The number of tasks and the number of trials within each task are both important because the precision of the estimate for each individual variable is directly related to the size of the sample, and the precision of the estimates of the parameters of the regression equation is directly related to the number of different variables included in the analyses. The combination of the need for many trials, in many

tasks, with large samples in each group, may make these procedures impractical in most cases. Unfortunately, at the current time there do not appear to be good alternative methods for separating general and specific group-related influences on RT variables.

Analytical methods involving the same type of variable, for example, all RT variables, address the question of the specificity of the group-related effects. That is, are the group-related effects on different RT variables specific, or are they merely manifestations of a more general slowing phenomenon? Another question that can be asked if a variety of different types of variables are included in the analyses concerns the uniqueness of the group-related influences. That is, to what extent are the group-related influences on RT variables independent of, and distinct from, the group-related influences on other types of variables? The answer to this question is potentially important because it has implications for the kinds of explanations that might eventually be necessary. That is, if nearly all of the group-related effects on RT variables were independent of the group-related effects on other types of variables, then explanations for the group differences in RT could be fairly specific. In contrast, much broader explanatory mechanisms would presumably be needed if many of the group-related effects on the RT variables were found to be shared with those on other types of variables.

For purposes of illustration, assume that each member of two groups is assessed on two RT variables, A and B, and on three other variables (V_1, V_2, and V_3) that could represent memory, reasoning, or virtually any other aspect of performance. The interesting question in the present context is to what extent are the group-related effects on all these variables independent? The top panel of Figure 3 portrays the possibility that

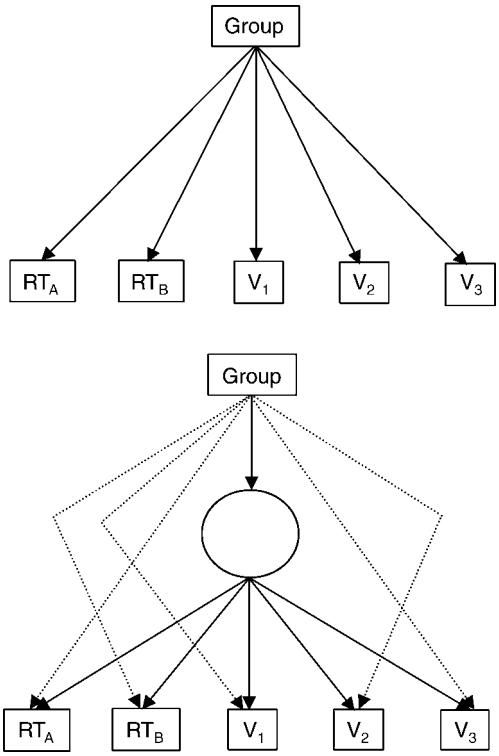

Fig. 3. Schematic illustration of completely independent group-related effects on a set of variables (top panel) and of both shared and independent group-related effects on the variables (bottom panel). The dotted lines in the bottom figure represent group-related effects on individual variables that are independent of the effects shared with other variables.

the group-related effects on the five variables are completely independent of one another because each variable has a direct arrow from the box representing group membership. In contrast, the bottom panel of Figure 3 portrays a situation in which some of the group-related effects on the variables are shared, as represented by the effects mediated through the circle, whereas other group-related effects are independent, as represented by the dotted arrows directly from the box signifying the group classification to the individual variables.

One method of analysis that allows these possibilities to be distinguished is known as shared influence analysis. A desirable sequence of steps with this method is as follows. First, the interrelations of the variables are determined both before and after controlling for group membership, or if the samples are large enough, separately within each group. If the pattern of interrelations is similar across these comparisons, then the researcher will have some confidence that the inferred structure among the variables is not an artifact of the relation of each variable to the group classification. Next the researcher determines which variables have independent effects related to group membership after controlling for the shared or general effects. This is conceptually analogous to using the first principal component in a principal components analysis to represent what all variables have in common, and then controlling that component with hierarchical regression before examining any effects related to group membership on the individual variables. A more elegant method of obtaining quantitative estimates of the relative proportions of unique and shared group-related influences relies on methods based on structural equation modeling (e.g., Salthouse, Hambrick, & McGuthry, 1998) in which all values are estimated simultaneously.

If the direct relations associated with group membership are only a small proportion of the total group-related effects on the RT variables, then the researcher would likely infer that many of the group influences on the RT variables are shared with the influences on other types of variables. Although this shared-influences method is not necessarily informative about the nature or identity of any shared group-related influences that might exist, it does provide a means of distinguishing shared and unique effects, and of estimating the relative contributions of the two types of influences. This analytical method is still quite new, but it has been used in comparisons of adults of different ages (Salthouse et al., 1998), and in contrasting normals with Alzheimer patients (Salthouse & Becker, 1998), and normals with HIV patients (Becker & Salthouse, 1999).

To summarize, many researchers seem to assume that the group-related differences they observe on a particular variable are attributable to effects on processes specific to that variable, but the plausibility of this assumption cannot be examined unless other variables are considered at the same time. A variety of multivariate procedures could be used to distinguish between shared and unique group-related influences. Regardless of the particular analytical method used, whenever possible, researchers should obtain a variety of different types of variables from their research participants to allow the possibility of broader influences to be investigated.

SUMMARY

Reaction time tasks are well-suited to the study of cognitive processes that could not be otherwise behaviorally observed. As with any method, RT measures do have limitations, and it is important to be aware of these when using and interpreting such measures. There are also special difficulties involved when RT measures are used in between-group comparisons. Nevertheless, RTs are often critical to the investigation of particular hypotheses, and RT variables have been used productively in the study of attention, memory, and processing efficiency. Although we hope to have convinced the reader that the use and interpretation of RT measures is considerably more complicated than is typically assumed, we also hope that the reader will be able to use this knowledge to improve the design and analysis of future studies involving RT measures. The near inevitability of speed-accuracy tradeoffs complicates most interpretations of RT, and there is no simple solution to this problem. The best a researcher can probably do at the present time is rely on a combination of analytical procedures involving

different assumptions, and look for a consistent pattern. If the results of the procedures are not consistent, then there may be no alternative to generating complete speed-accuracy tradeoff functions for every individual in every condition. Many possible processes are hypothesized to contribute to RT, and therefore researchers often attempt to isolate specific processes with difference scores. However, difference scores are limited by an expected correlation with the initial or baseline value, and decreasing reliability as the correlation between the two scores increases. The use of regression-based residuals solves some, but not all, of these problems. Between-person comparisons of RT variables are often conducted with group-by-variable analyses of variance, but they have the disadvantage of not allowing general and specific effects to be distinguished. A variety of methods can be used to attempt to adjust for group differences in baseline RT, but each requires careful thought for proper application and interpretation. Finally, it is desirable to examine group effects on RT in the context of group effects on other types of variables to determine the uniqueness of the influences associated with group membership.

REFERENCES

Becker, J.T., & Salthouse, T.A. (1999). Neuropsychological test performance in the Acquired Immunodeficiency Syndrome: Independent effects of diagnostic group on functioning. *Journal of the International Neuropsychological Society, 5,* 1–7.

Cerella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin, 98,* 67–83.

Chapman, L.J., & Chapman, J.P. (1988). Artifactual and genuine relationships of lateral difference scores to overall accuracy in studies of laterality. *Psychological Bulletin, 104,* 127–136.

Chapman, L.J., Chapman, J.P., Curran, T.E., & Miller, M.B. (1994). Do children and the elderly show heightened semantic priming? How to answer the question. *Developmental Review, 14,* 159–185.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Donders, F.C. (1969). On the speed of mental processes. In W.G. Koster (Ed. and Trans.), *Attention and performance II* (pp. 412–431). Amsterdam: North Holland. (Original work published 1868)

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Developmental Psychology, 109,* 490–501.

Kail, R. (1998). Speed of information processing in patients with multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology, 20,* 98–106.

Madden, D.A., Pierce, T.W., & Allen, P.A. (1992). Adult age differences in attentional allocation during memory search. *Psychology and Aging, 7,* 594–601.

Maylor, E.A., & Rabbitt, P.M. (1993). Alcohol, reaction time and memory: A meta-analysis. *British Journal of Psychology, 84,* 301–317.

Meyer, D.E., Osman, A.M., Irwin, D.E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology, 26,* 3–67.

McClelland, J.L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

Nebes, R.D., & Brady, C.B. (1992). Generalized cognitive slowing and severity of dementia in Alzheimer's Disease: Implications for the interpretation of response-time data. *Journal of Clinical and Experimental Neuropsychology, 14,* 317–326.

Pachella, R.G. (1974). The interpretation of reaction time in information-processing research. In B.H. Kantowitz (Ed.), *Human Information Processing: Tutorials in performance and cognition* (pp. 41–82). Hillsdale, NJ: Lawrence Erlbaum.

Posner, M.l., Boies, S.J., Eichelman, W.H., & Taylor, R.L. (1969). Retention of visual and name codes of single letters. *Journal of Experimental Psychology Monographs, 79,* 1–16.

Reitan, R.M. (1992). *Trail Making Test: Manual for administration and scoring.* Tucson, AZ: Reitan Neuropsychology Laboratory.

Salthouse, T.A. (1985). Speed of behavior and its implications for cognition. In J.E. Birren & K.W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed.). New York: Van Nostrand Reinhold.

Salthouse, T.A. (1993). Attentional blocks are not responsible for age-related slowing. *Journal of Gerontology: Psychological Sciences, 48,* P263–P270.

Salthouse, TA., & Becker, J.T. (1998). Independent effects of Alzheimer's Disease on neuropsychological functioning. *Neuropsychology, 12,* 1–11.

Salthouse, T.A., & Coon, V.E. (1994). Interpretation of differential deficits: The case of aging and mental arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1172–1182.

Salthouse, T.A., & Fristoe, N.M. (1995). Process analysis of adult age effects on a computer-administered trail making test. *Neuropsychology, 9,* 518–528.

Salthouse, T.A., Hambrick, D.Z., & McGuthry, K.E. (1998). Shared age-related influences on cognitive and noncognitive variables. *Psychology and Aging, 13,* 486–500.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171,* 701–703.

Spieler, D.H., Balota, D.A., & Faust, M.E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of ExperimentaI Psychology: Human Percepion and Performance, 22,* 461–479.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica, 30,* 276–315.

Wickelgren, WA. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85.