2

Methodological and Statistical Techniques for the Chronometric Study of Mental Abilities

ARTHUR R. JENSEN

The study of individual differences in reaction time (RT) had its origin not in psychology, but in astronomy. The Prussian astronomer F. W. Bessel, in 1823, coined the term *personal equation* for the consistent differences among telescopic observers in recording the exact moment that the transit of a star crosses a hairline in the visual field of the telescope. The need to make corrections for the personal equation led to the invention, in 1828, of the chronograph, an instrument for the precise measurement of RT, which was later to become useful to psychologists.

But it was not until the 1860s that RT was taken up by psychologists. In that same decade, psychology was launched as an empirical science. Its founding fathers were Sir Francis Galton (1822–1911), in England, and Wilhelm Wundt (1832–1920), in Germany.

The measurement of RTs figured prominently in the laboratories of both Galton and Wundt, but their purposes were quite different and led them in separate directions. Galton's interest was mainly in the nature and measurement of individual differences. He has been claimed as the father of differential psychology, which also subsumes mental measurement, or psychometrics. Wundt is recognized as the founder of experimental psychology; he aimed to discover the general principles of mind and behavior, much as physicists had established the fundamental laws of matter and energy.

This division between the methods and aims of differential and experimental psychology has existed from psychology's very beginnings as

ARTHUR R. JENSEN • School of Education, University of California, Berkeley, California 94720.

an empirical science. Distinct lines of descent, from Galton and Wundt to the present, are discerned through the history of psychology, not just as the normal division of investigative labor, which necessarily exists in every science, but also as a difference in philosophical attitude and theoretical orientation with regard to psychology's development, both as pure science and as technology. Psychology's historical duality is often referred to today, in terms of Cronbach's (1957) well-known characterization, as "the two disciplines of scientific psychology." Cronbach deplored the theoretical and methodological separateness of the two disciplines and suggested that a proper marriage would prove fruitful, and indeed was necessary, for the advancement of psychology as a science.

This bit of history is recounted as relevant to RT research because it is exactly in this specialized domain that, finally, we are seeing the rapid development of what may well be the most promising example of the kind of marriage that Cronbach had envisaged between the two disciplines of scientific psychology.

In Galton's laboratory, RT was simply used as one among several measures of individual differences in "human faculties." In addition to RT, Galton also measured other elemental sensory–motor functions and physical traits that he judged to be significant in human evolution and believed to be more strongly influenced by heredity than by environment. He hoped that some weighted combination of such measurements would afford an objective index of an individual's largely innate general mental capacity. The practical application of this effort, as Galton (1908) stated in his autobiography, "would be to estimate the combined effect of these separately measured faculties. . . and ultimately to ascertain the degree with which the measurement of sample faculties in youth justifies a prophecy of future success in life, using the word 'success' in its most liberal meaning" (p. 267). So it was that Galton's work foreshadowed what was later to become one of the most controversial aspects of applied psychometrics—the prediction of an individual's future educational or occupational performance from current measurements of ability or aptitude.

Although Galton himself invented a novel device for measuring RT accurately (to one one-hundredth of a second), there is no evidence that he had any interest in RT as a phenomenon to be studied experimentally in its own right. He viewed RT only as one of many different means of mental measurement. Unfortunately for the history of psychometrics, Galton's overly simple method of RT measurement could only reveal a scarcely impressive relationship to other criteria of intellectual capacity, although it is noteworthy that several of Galton's laboratory tests, including RTs to auditory and visual stimuli, showed statistically significant mean differences between several occupational levels, from professional to unskilled worker. The use of only *simple* RT, and with too few trials for adequate reliability,

doomed it to failure. (The average test–retest reliability of Galton's RT measurements was only about .17.) The same mistakes, repeated by Galton's immediate disciples (most notably James McKeen Cattell), led to the premature abandonment of RT as a technique for the study of individual differences in mental ability; the technique was not to be revived for at least half a century.

In Wundt's laboratory, however, with its strong experimental emphasis, its search for general principles, and its lack of interest in individual differences (except as "error" variance), RT measurement served a very different purpose in psychological research. At that time, it proved to have a scientifically more influential purpose, so much so, in fact, that Boring, in this *History of Experimental Psychology* (1950), refers to the late nineteenth century as "the period of mental chronometry." Reaction time was used then as the principal technique for the objective analysis, or decomposition, of mental activity, identifying, and measuring in real time, such processes as perception, apperception, cognition, association, discrimination, choice, and judgment.

The essential idea for this application of RT in psychological research is credited to F. C. Donders, a Dutch physiologist, whose innovative method, first published in 1862, was taken up and developed further in Wundt's laboratory in Leipzig in the 1880s. Donders' essential methodological contribution was the *subtraction method,* which is the notion that the different speeds of reaction to experimentally varied tasks represent additive components of time for the execution of the various mental processes occasioned by the task conditions and that the differences between the reaction times to the systematically varied tasks could be used to isolate and determine the duration of each of the component processess of a complex mental act (Donders, 1868/1969). The assumption is made that the time for all such processes intervening between stimulus and reaction summate in a strictly additive fashion and therefore can be precisely decomposed by subtracting the RT to simpler tasks from the RT to more complex tasks. Although this assumption has since been seriously criticized, the basic idea was an especially important one for psychology at that time, for it demonstrated that mental events take place in real time that could be precisely measured and quantitatively analyzed, as are physical events in the natural sciences. And thus, with the advent of RT measurement, psychology took a large step on its path, from speculative philosophy to empirical science.

Then, shortly after the turn of the century, interest in RT markedly waned as experimental psychologists became increasingly engrossed in the laboratory study of conditioning and learning as pioneered by Pavlov and Thorndike. Except for an occasional study using RT (which now takes on retrospective significance), there was slight interest in RT research among

academic psychologists for almost 50 years. In the 1970s, RT was rediscovered by researchers in experimental cognitive psychology. This field has adopted RT techniques, now more broadly termed *mental chronometry,* as its most important methodology. Although a number of psychologists were instrumental in this recent revival of mental chronometry, it probably owes most to the initial work of Michael I. Posner and Saul Sternberg, who are both still active in this field.

CATEGORIES OF REACTION TIME RESEARCH

Today, RT research can be conveniently divided into three categories, although the three are not always distinct in the research literature. Each is important for our purpose.

Reaction Time per se *as a Dependent Variable.* Reaction time can be studied experimentally as a dependent variable in its own right. This research aims to comprehend all the stimulus and response conditions, and the effects of practice, on all the measurable parameters of RT performance, including the overt error rate. The main focus is on the measurement properties of RT itself. Research in this area goes beyond empirical description of functional relationships. It is now concerned with the construction and testing of theories or models of the RT process that can include all the observed variation in RT as a function of experimentally manipulable variables (Smith, 1968). A recent example of this kind of model, and the evidence brought to bear on it, is found in the recent work of Grice, Nullmeyer, and Spikes (1982). There is no generally accepted theory or model of choice for RT yet. But this type of experimental investigation of RT, along with testable hypotheses to explain the results, is valuable and necessary for the other two categories of uses of RT. The essential nature of RT cannot be ignored when RT is used as a measurement technique for the primary study of other, more complex, cognitive phenomena. Many lines of research on RT *per se,* in addition to such other lines of investigation as time perception (e.g., Pöppel, 1978), the latency of conscious awareness (Libet, 1965), and evoked brain potentials (e.g., A. E. Hendrickson, 1982; D. E. Hendrickson, 1982) must all converge on the "black box" of hypothetical cognitive or neurological processes that mediate stimulus and response in mental tasks, if this black box is ever to be scientifically fathomed. Only then we can hope to understand the basic mechanisms responsible for individual differences in performance on mental tasks. The attempt to formulate testable models of the brain, for which RT affords a promising methodology, among others, is essentially a search for simplicity. The achievement of simplicity in science

is greatly aided by sensitive and precise measurements. A scientist cannot wallow in unquantified complexity if he is to escape hopelessly vague or untestably complex causal theories. Good scientists succeed in *achieving* simplicity. Newton expressed this idea in his famous dictum "Nature is simple." In truth, however, neither simplicity nor complexity is inherent in nature. Simplicity or complexity are constructions of the scientist's effort to understand nature, an effort that is often abetted by more powerful techniques of observation and measurement. Such is the role of RT in the study of higher mental processes.

Reaction Time as an Analytic Technique in Experimental Cognitive Psychology. Research in cognitive psychology using RT techniques has been called *mental chronometry* by Posner (1978), who defines mental chronometry as "the study of the time course of information processing in the human nervous system" (p. 7). The growth of interest in RT in recent years has paralleled the growth of experimental cognitive psychology, for which the precise measurement of time has become the most frequent dependent variable used for the analysis, or decomposition, of the processes involved in *cognitive* tasks.

The emphasis here is on *cognition,* because RT measurement has too long been popularly thought of as the assessment of sensory–motor ability. It is taken for granted even by many psychologists, for example, that highly skilled athletes should outperform, say, university students in all RT tasks. Yet Mohammed Ali, perhaps the greatest boxer of all time, in his prime was found to show a very average RT (Keele, 1973). The fact is that only a small part of a person's total RT is attributable to peripheral sensory–motor functioning. The total RT sequence between stimulus onset and the initiation of response includes sense organ lag, peripheral nerve transmission time, muscle latency, and brain time. Most of the total time consists of brain activity, which is what cognitive psychologists are especially interested in evaluating. Moreover, experimental techniques permit the separation of the times required for the sensory–motor activity from brain time in a particular RT task. For example, it has been determined that only 15 to 30 milliseconds (ms) is required from sense organ to brain, whereas the fastest human RT to a single stimulus is about 150 ms. The stimulus–response (S–R) time for a spinal or subcortical reflex is more than twice the simple RT (SRT), showing that the cerebral cortex is the main source of delay in RT.

Another important fact that emphasizes the relative importance of cerebral activity in RT, as contrasted with sensory–motor mechanisms, is the finding that there is quite a large general factor in individual differences (ID) in various RT tests, which cuts across all different stimulus and response modalities—visual, auditory, tactile, left and right hands and feet,

and biting. Hence, there seems little doubt that RT is more central than a peripheral phenomenon. Even if this were not the case, it would be possible, experimentally, to determine the amount of time attributable to peripheral processes and that attributable to central processes.

An extremely simple example of how processes can be determined is the decomposition of the time required to name visually presented words. It has been observed that words that are longer, in number of syllables, take more time to name, when time is measured from the onset of visual presentation to the initiation of the spoken word, recorded by a voice-activated key. The question is, Do longer words take longer to name because they take longer to be visually encoded, that is, recognized, or because the vocal response takes longer? The experimental paradigm for answering this question is simple. In the first condition, the person sees single words projected one at a time on a screen and reads each word aloud as fast as possible. The average S–R interval, or RT, is recorded for words of different numbers of syllables. This interval comprises the amount of time it takes the person to *encode* the stimulus and to *prepare* the appropriate vocal response, an act that involves the complex coordination of breath, vocal cords, tongue, and lips. In the second condition, the person sees single words projected on a screen, but is instructed to delay the vocal response until a light flashes on, there being a brief interval between the presentation of the word and the light. The average interval between the flash of light and the initiation of the vocal response is also recorded as a function of word length. This interval represents the time taken for the first condition *minus* the time required for the *encoding* of the stimulus word. What the experiment reveals may seem surprising: the time variation in naming words, as a function of their number of syllables, is attributable to differences in encoding time and not to differences in response preparation time (Eriksen, Pollack, & Montague, 1970). This simple experiment illustrates a general assumption in mental chronometry, which is that information processing takes place in real time in a sequence of stages and that the total measured time from the initiation of a mental task can be analyzed in terms of the time required for each stage. It can also be determined whether the stages are temporally discrete, or overlap or interact for any given task.

Another possible separation of cognitive processes by mental chronometry is the important distinction between *structural* and *functional* components of information processing, which are analogous to the "hardware" and "software" components of a computer, respectively. The structural components, for example, would be less easily influenced by practice or special training than would the functional components. The functional components involve the control of processing within the structure

and would involve different responses to instruction for various strategies applied to a cognitive task. Individual differences in the structural and functional aspects of information processing would seem to correspond rather closely to Cattell's distinction between fluid and crystallized ability (Cattell, 1971). Piaget's theory of cognitive development similarly distinguishes among structure, function, and content of mental operations (Flavell, 1963, Ch. 2).

Although at present there is no general theory of information processing, practically all workers in this field view the structural, or hardware, components as consisting of such elemental processes as sensory encoding, or the mental representation of a stimulus; short-term and long-term memory storage; memory scanning and retrieval systems; and response execution. But the extent to which each of these processes can be characterized as structural or functional is still open to question and investigation, as is even the clearness of the distinction between structure and function as it applies to the brain. Analogizing from computer components to a neurological system can suggest hypothetical cognitive models, but the limitations of such a method for understanding biological systems are recognized.

The list of substantive topics in psychology to which analysis by chronometric methods has already been applied is extensive—sensory coding; selective attention; apprehension; perceptual integration; pattern recognition; stimulus comparison, matching, and transformation; retrieval of information from short-term and long-term memory; psychological refractoriness; parallel and serial information processing; the mental representation of semantic and logical relations; inference in verbal, pictorial, and figural analogies; spatial reasoning; and the selection and execution of responses, to name only the most commonly researched processes. Chronometry has also been used to study such complex phenomena as *reading skills* (e.g., Carpenter & Just, 1975; Ehri & Wilce, 1983; Posner, Lewis, & Conrad, 1972; Spring, 1971), *dyslexia* (Spring & Capps, 1974), *mental retardation* (Baumeister & Kellas, 1968), and even *personality* (Brebner, 1980).

Reaction Time in the Analysis of Individual Differences in Mental Abilities. The elementary information processes discovered in the kind of studies described in the preceding section display a wide range of individual differences (IDs). In studies of elemental cognitive processes, even in groups with a highly restricted range of ability, such as university students, IDs constitute a much large source of variance in RT measurements than do the experimentally manipulated task conditions. For example, 85% of the total variance is ascribable to IDs and only 15% to the experimental conditions in Posner's (1978, Ch. 2) letter-matching task, in which

subjects, under one condition, must respond *same* or *different* to pairs of letters in terms of their physical characteristics (upper *versus* lower case type) or, under another condition, must respond *same* or *different* to their letter names. The latter condition has longer RTs, because in addition to sensory encoding of the stimuli, semantic encoding is required, which involves access to overlearned letter names in long-term memory. The IDs in such processes can be the main object of investigation in their own right.

If the rank order of IDs in RT measurements were found to be no more consistent than if the order were determined with a table of a random numbers, between one experimental paradigm and another, then the measurements, however reliable they may be in any one paradigm, would be so task-specific as to be too trivial for scientific study. Hence the study of IDs in elementary cognitive tasks must rely heavily on methods of correlation analysis to identify sources of IDs that are not too task-specific. The investigator seeks evidence for the generality of IDs in the chronometric measurements obtained in a particular experimental paradigm. Unfortunately, this quest is made difficult by the fact that a very substantial part of the IDs variance in fine-grained laboratory measurements is task-specific. That is to say, IDs do not remain in the exact rank order from one task to another, even when, formally, the tasks would seem to elicit the same processes. In the terminology of the analysis of variance, IDs *interact* with the specific or unique features of each task. In terms of correlation, IDs are imperfectly correlated among various tasks, even after the correlations are corrected for attenuation as a result of errors of measurement. In terms of factor analysis, the single tasks have rather large *specifics*, that is sources of reliable IDs variance that are not shared with other tasks. Hence, analyses of variance, correlations, and factor analysis are the obvious methods for determining sources of IDs in elementary cognitive tasks that have enough generality across tasks to be of theoretical or practical interest.

Two main strategies have been adopted in this pursuit. The first, but least developed, looks for correlations among RT measurements of theoretically similar cognitive processes (e.g., sensory encoding, choice, response selection) as they are hypothesized to occur in different experimental paradigms. Reliable and even fairly substantial correlations that demonstrate IDs in the hypothesized elementary processes involved in different tasks have been found. Evidence for the distinct processes depends on the finding of higher correlations among chronometric variables hypothesized to arise from the same process than among variables hypothesized to arise from different processes (e.g., Keating & Bobbitt, 1978). However, the observed correlations, even after correction for attenuation, are usually smaller than those we are accustomed to find among various tests in traditional psychometrics. The most probable reason for this is that the

individual experimental paradigms used for chronometric analysis yield scores (time measurements) that are factorially more like scores on single items in psychometric tests than total scores on tests with many different items. In psychometric tests composed of varied items, more of the total IDs variance consists of item covariance than of item specificity. The total variance in test scores comprises the sum of the item variances plus twice the sum of the item covariances, and the item covariances increase at a greater rate, by a factor of $n^2 - n$, than do the item variances, as the number of items, n, increases. It should be recalled that the single items in psychometric tests have large specifics, the average correlation among single items being usually in the .2 to .3 range. The larger the number of various items, the more is the specificity "averaged out," so to speak, accentuating whichever factors the items measure in common. To be sure, the use of many repeated trials in chronometric tasks can ensure high internal consistency reliability, but it does not diminish task specificity. Given the great homogeneity of the repeated measures in RT tasks, what is really most surprising is the finding that such homogeneous measures are as highly correlated as they are with certain other measures of ability. Generally, correlations among highly homogeneous RT parameters obtained in a single paradigm and scores on psychometric tests of ability fall between .2 and .5.

Indeed, the second method for "validating" the generality of chronometric scores is to show their correlations with psychometric tests, especially those that measure well-established factors of ability, such as general intelligence, or *g*, and verbal, quantitative, and spatial visualization abilities. The fact that these psychometric abilities, which have emerged in countless factor analyses over the past 75 years, can be very reliably measured by standardized tests, and are known to have substantial predictive validity for educational and occupational criteria, lends further interest and importance to those RT paradigms, or combinations thereof, that show the highest correlations with psychometric scores.

Analysis of the correlations between chronometric and psychometric scores is probably the "richest vein," in terms of potential, for advancing our theory of human mental ability. For some time there has been a growing consensus among differential psychologists that the traditional methodology of studying mental ability in terms of classical psychometrics, factor analysis, and external validation, over the last 75 years or so, has accumulated an impressive amount of solid empirical facts on the range, correlational structure, and practical consequences of IDs in ability, but has not contributed to the further development of theoretical explanations of the main abilities identified by factor analysis of psychometric tests. In the traditional framework, explanations of IDs have not advanced beyond statements that, to put it in the simplest form, individuals A and B differ in

performance on task X, because X is highly saturated (or loaded) with ability factor Y, and A and B differ in ability factor Y. But ability Y is a hypothetical or mathematical construct that is not invariant to the method of factor analysis used to identify it. There is unfortunately nothing in the raw psychometric data that can compel the factor theorist to explain A's and B's difference in performance on task X in terms of their differing in factor Y. Factor rotation could displace the IDs variance on factor Y and divide it between two other factors P and Q, so that then the difference between A and B would be attributed to their differing in factors P and Q. And factors P and Q would be different from factor Y, according to the usual method for psychologically describing factors in terms of the characteristics of those content-homogeneous tests that show the highest loadings on the factor. This, in essence, is the theoretical blind alley that differential psychologists find themselves in if they confine their methodology to traditional psychometric tests and factor analysis. The measurements and methods of psychometry reveal only the end products of mental activity, and, by themselves, cannot expose the processes intervening between problem presentation and a subject's response. It is in these intervening processes, at some level of analysis, that the explanation for IDs is to be sought. Mental chronometry and electroencephalography afford the chief tools for such process analysis at the interface of brain and behavior. The tools themselves do not interfere with the normal functioning of the intact brain.

Scores on traditional tests represent a complex amalgam of causes that are not amenable to analysis in terms of elemental processes by any classical psychometric methods. Factor analysis reveals common sources of variance among various tests, but does not reveal the nature of these sources. Within this framework, we cannot answer such questions as why there are quite large correlations between tests that differ as much say, as, vocabulary, block designs, and number series, except to state that all these seemingly dissimilar tests measure a common factor, often termed *g* (for *general* ability). But that is hardly more than a tautology, not an explanation, as the emergence of the *g* factor merely reflects our original observation that scores on all the tests are positively correlated with one another. (In addition, factor analysis shows precisely the degree to which each test shares the common variances in all the tests entered into the factor analysis.) In fact, the analysis of correlations among variables, such as factor analysis, should probably be called *synthesis* rather than *analysis*, since syntheses represent a higher level of abstraction or generalization from the observed phenomena, not a decomposition of it into less complex causal elements. An important aim of chronometric analysis, in contrast to factor analysis, is to achieve a decomposition of complex abilities and to measure the IDs in the common elemental processes that effect the correlations among complex tests.

Elemental processes are hypothetical and inferential constructs, as are

factors; processes are truly *analytical* constructs, whereas factors are really principles of synthesis, or classification. Factors only signify the presence of common causal elements that remain to be identified and measured. Iron, copper, and gold, although different, have certain properties in common: They are malleable, they melt at specific temperatures, and they conduct heat and electricity. By analogy with factor analysis, we would explain these commonalities by going to a higher level of abstraction and noting that iron, copper, and gold are all metals. A process analysis, by analogy, would explain their similarities (and differences) in conductivity, in terms of the number and arrangement of their orbital electrons. When a person faces a task, such as a test item, certain things must happen, in some sequence in time, for the person to arrive at the appropriate response. The analysis of these activities in terms of the time they take is the aim of mental chronometry. The term *activity* here can refer to any level of analysis, from observed, overt behavior to inferred, hypothetical brain processes.

DIFFERENCES BETWEEN PSYCHOMETRIC AND CHRONOMETRIC DATA

Psychometric and chronometric data differ in three main ways.

Scale Properties. Scores on psychometric tests based on number of correct answers (or some transformation of the raw scores) measure ability on an arbitrary, relativistic, or norm-referenced (i.e., standardized) scale. There is no true zero point, and the interval property of the scales depends on the acceptance of certain theoretically based assumptions, however plausible, about the form of the distribution of the ability in the population. A scale for which equal intervals are claimed, based on an assumption of the true form of the population distribution of the trait, obviously cannot be used to test hypotheses about the form of the distribution. Also, without the assurance of an interval scale, one cannot meaningfully plot the form of mental growth curves. Without an absolute or ratio scale (i.e., an interval scale with a true zero point), one cannot meaningfully compare *proportions* of mental growth from one period of time to another.

Norm-referenced or standardized score scales also have the disadvantage of a questionable comparability of norm groups, across different tests and for the same test (or equivalent forms) normed at different times. Consequently, for example, Wechsler and Stanford-Binet IQs may differ because of non-comparable norm-reference groups. It is virtually impossible to determine why scores on such tests are higher (or lower) from one generation or decade to the next. Is it because of true changes in the level of ability in the population or because some of the test items are merely easier due to familiarization of the item contents or because of sampling

differences in obtaining norm-reference groups at different points in time? A "random" or "representative" sample of a national population, although a theoretically definable concept, is a mythical concept, practically speaking.

Chronometric data on IDs, in marked contrast to psychometric test scores, surmounts most of these difficulties and disadvantages because they consist of absolute measurements of real time, expressed in seconds or milliseconds, which are standard units in the universally adopted *Système Internationale* for all physical and scientific measurements.

Precision and Sensitivity. The smallest unit of measurement on psychometric tests is the scale on which single items are graded. This is usually a 2-point scale ("right" or "wrong," 1 or 0), or, as in some of the subtests of the Wechsler scales, a graded scale of several points, depending on the quality or speed of the individual's performance. In either case, performance at the item level is scored in terms of a relatively coarse scale.

The unit of measurement for RT is usually the millisecond. The obvious advantage of such a refined measurement is its extreme sensitivity. Extremely small differences in ability or performances, undetectable by tests scored right or wrong at the item level, can be detected. For example, a paper-and-pencil test of simple addition of pairs of single numbers (e.g., 5 + 2 = 7, which is answered *true* or *false*) will hardly discriminate between sixth-graders and college students. Yet such an age discrimination is very marked when true–false response latencies are measured. Other interesting phenomena, which reveal the nature of the cognitive processes involved in this simple task, are also evident from an analysis of the mean latencies for each item. For example, the mean latency is directly related to the size of the smaller of the two addends. That is, response latency increases as the smaller addend increases in size, which suggests that subjects begin with the larger addend and count up the number of the smaller addend. This strategy is also suggested by the observation of corresponding finger movement in younger children. The interesting point, however, is that the same rank order of differences in mean response latencies for different problems is observed in children *and* adults, although in adults the latencies are shorter and the relative differences between problems are less pronounced. But the chronometric data reveal that adults use the same counting strategy for simple addition as do children. Without a precise chronometric apparatus and repeated measurements, it would be virtually impossible to obtain such data. Error rates scarcely differ across various number combinations for simple addition, and the subjects have no subjective feeling that such easy problems differ at all in difficulty. The small differences in response latencies are not detectable by direct observation. But with a suitable reaction timer, even more subtle cognitive effects are revealed. For example, why should the response (*false*) to 4 + 3 = 12

have a significantly longer latency than to either 5 + 2 = 12 or 5 + 3 = 12? It is evidently because of the extra time required to discriminate between 4 + 3 = 12 (*false*) and 4 × 3 = 12 (*true*), whereas no such discrimination is called for in 5 + 2 = 12 or 5 + 3 = 12.

Range of Ability. Because psychometric test items are scored *right* or *wrong* (quantized as 1 or 0), they must be at a suitable level of difficulty for any given group if they are to detect IDs reliably. As item difficulty departs in either direction from a p value (p = proportion of a group passing the item) of .50, item variances and covariances decrease and the detection of IDs becomes less reliable. Hence, the same set of test items cannot be used for subjects with a wide ability range. For example, there is no common set of test items on which it is possible to compare, say, five-year-olds or retarded adults and college students and also reliably measure IDs *within* each group. The usual solution is to use different sets of items of the same type (e.g., vocabularly, figure analogies, matrices), but of widely differing levels of difficulty, and then show, by means of factor analysis of the tests in the overlapping ability groups, that the factor composition of the various tests is the same for all levels of ability. This procedure is often difficult to follow, practically, and is problematic, theoretically. A sameness of factor composition across widely varying ability levels does not solve the problem of comparability of scale units across the full range of ability.

Chronometric techniques have a great advantage in all these respects. Because of the great sensitivity of RT measurements, as described in the preceding section, the tasks used can be so simple that they can be performed correctly by persons who differ even as extremely in ability as severely retarded adults (with IQs below 40) and the brightest university students. The IDs are measured not in terms of "right" or "wrong," but rather in terms of response latency, or RT. Of course, some RT tasks are somewhat more limited in this respect because of their greater complexity or the knowledge or skills required. Even so, chronometric tasks are generally applicable over a much wider range of ability than is any one-and-the-same psychometric test.

Basis of Correlation between Chronometric and Psychometric Indices of Individual Differences

If IDs in chronometrically and psychometrically obtained indices correlate significantly, it can be hypothesized that they both tap the same sources of variance involving the speed or the efficiency of mental processes. The importance of a time element in mental efficiency can be understood in terms of certain well-established concepts and principles of cognitive psychology. The conscious brain acts as a single-channel, or

limited capacity, information processing system. Limited capacity also restricts the number of operations that can be performed simultaneously on the information that enters the system from external stimuli or from retrieval of information stored in short-term or long-term memory (STM or LTM). Hence, speediness of mental operations is advantageous because more operations per unit of time can be excuted without overloading the system. Also, because there is a *rapid decay* of stimulus traces and information, speediness is an advantage for any operations that must be performed on the information while it is still available. Finally, to compensate for limited capacity and rapid decay of incoming information, a person resorts to *rehearsal and storage* of the information into LTM, which has a practically unlimited capacity. But the storage process itself takes time and ties up channel capacity, so there is a trade-off between storing and processing incoming information. The more complex the information and the operations required on it, the more time is required and the greater the advantage of speediness in the elemental processes involved. Loss of information because of overload interference and the decay of traces that were inadequately encoded or rehearsed for storage or retrieval from LTM result in a failure to grasp all the essential relationships among the elements of a complex problem needed for its solution. Speediness of processing, therefore, should be increasingly related to success in dealing with cognitive tasks to the extent that their information load strains the individual's limited channel capacity. The most discriminating test items, scored in terms of right or wrong, thus would be those that bring the information processing system to the threshold of breakdown. In a series of items of graded complexity, such breakdown would occur at different points for different persons. If IDs in the speed of the elemental components of information processing can be measured in RT tasks that are so simple as to rule out breakdown failure, it should be possible to predict IDs from the point of breakdown for more complex tasks, such as the most discriminating items in psychometric tests.

Seemingly small but reliable IDs in the speed of performing certain elementary cognitive tasks, amounting to less than 100 ms, may show up on certain psychometric tests as very large differences, such as one person's vocabularly being only one-half as large as another person's. Small absolute differences in rate of information processing, involving encoding and storage, can result in large IDs in the amount of information and skills acquired over long periods of time. A good analogy would be that of two cars on the highway travelling side by side at only slightly different average rates, say, 50 and 51 mph, respectively. Within a few hours, they will be miles apart. Thus, full siblings reared together, with the same exposure to language and the same educational opportunities, may, by the time they enter high school, show large differences in vocabulary, general informa-

tion, and the intellectual skills important for success in school and in the world of work. Such IDs are found to be correlated with IDs in RT to elementary tasks for which the task requirements are easily within the capability of perhaps 98% or 99% of the school-age and adult population, with the exception of those persons who have such severe sensory or motor handicaps as to rule out the possibility of their performing most RT tasks.

The Speed–Complexity Paradox

This is the name of the observation that speed of reaction correlates most highly with scores on complex psychometric tests only when the RT task is fairly easy, but still more complex than SRT (i.e., single stimulus–single response) and when RTs fall within the range of about 200 to 1,000 ms for normal adults. The paradox is that, whereas the RT in such undemanding tasks is correlated with IQ, as measured by complex psychometric tests, the response latencies to the IQ test items themselves are not correlated with IQ. However, if IQ test items of a difficulty level appropriate for, say, second-graders were administered to university students as stimuli in a RT paradigm, their response latencies would probably be correlated with the students' IQs as obtained on an IQ test suitable for university students. The reasons for this seeming paradox are not yet fully understood, but it appears that for very complex tasks (such as highly discriminating test items), different individuals resort to different strategies, or distribute the various elemental component processes disproportionately. For example, in solving verbal and pictorial analogies, higher IQ persons tend to allot more time to stimulus encoding and less time to response selection, whereas lower IQ persons do the reverse (R. Sternberg, 1977; R. Sternberg & Rifkin, 1979). Also, when the task is highly complex, personality factors affecting persistence, impulsiveness, and involuntary rest pauses become noncognitive sources of IDs in the response latencies.

MOST RELEVANT REFERENCES IN THE REACTION TIME LITERATURE

Before reviewing the methodology of RT studies in more detail, it would seem worthwhile to provide an annotated list of the books or chapters this writer considers to be the most essential reading for anyone who expects to do empirical research in this field. All these references themselves have extensive bibliographies. They are listed here in alphabetical order, by author.

Carroll (*Individual Difference Relations in Psychometric and Experimental*

Cognitive Tasks, 1980) is a detailed, critical, integrative review of recent research in experimental cognitive psychology, most of it based on chronometric methods. An excellent, comprehensive, and critical overview of the state of the art.

Eysenck (*A Model for Intelligence,* 1982) reviews in detail the research on mental speed, RT, inspection time, evoked potentials, and componential analysis as these concepts and methods have figured in studies of general intelligence.

Pachella (*The Interpretation of Reaction Time in Information-Processing Research,* 1974) discusses the major methodological problems in RT research; emphasized are the characteristics of RT in terms of the experimental conditions that affect it as a dependent variable. It contains probably the best available introduction to the subtraction method of Donders and the *additive factors* method of S. Sternberg, and detailed criticisms of these methods. It also thoroughly considers the *speed–accuracy* (error rate) problem in RT research.

Posner (*Chronometric Exploration of Mind,* 1978) is already a classic. Probably no other single reference shows the many ways that chronometric techniques can be used in psychological research. However, relatively very little of the book deals with IDs or with psychometric abilitities *per se.*

Welford (*Reaction Times,* 1980) is the most advanced and comprehensive work on RT *per se,* dealing largely with theoretical formulations of RT phenomena. It is also a mine of information on empirical research on RT.

Woodworth (1938) and Woodworth and Schlosberg (1954) (*Reaction Time*) are chapters in the classic textbook of experimental psychology. For their relatively short length, they are the most thoroughly informative and lucid introductions to RT research, and certainly the best places to begin one's reading in this field. These chapters, of course, antedate the modern revival of chronometry in experimental cognitive psychology, but the material they cover is basic and essential. Although there is considerable overlap in contents between the original (1938) and revised (1954) editions, both are well worth reading. In some respects, the earlier version is better for our purpose in that it gives more consideration to IDs in RT and to the correlation of RT with psychometric intelligence.

TYPES AND TERMINOLOGY OF REACTION TIME

Definitions of Reaction Time

Reaction time has been defined in a number of ways. Warren's (1934) *Dictionary of Psychology* defines RT as "the interval of time between the

onset of a stimulus and the beginning of the observer's overt intentional response . . . . The term *reaction time* is historically established; *intentional response time* is a more accurate term" (pp. 223–224). The qualification of *intentional* is now ambiguous, since we know that many RTs are faster than the speed of conscious awareness of a peripheral stimulus, which is about 500 ms (Libet, 1965). Another definition of RT is that it is the minimum amount of time needed for the observer to produce a *correct* response. This definition expresses the important fact that *false* responses can occur in an RT experiment and that the RT for false responses cannot be treated in the same manner as that for correct responses. But the qualification *minimal* amount of time makes the definition theoretical rather than operational because we cannot reliably measure the *minimal* RT of a given subject without some operational specification of what we mean by *minimal* RT (e.g., the mean of the shortest 5% of the subject's RTs in n number of trials). Actually all that is important for a definition of RT is that it be made explicitly operational in terms of the details of the experimental paradigm that is being used to measure RT.

Classical Reaction Time Paradigms

The classic paradigms and their terminology originated with the work of Donders and Wundt. These can be described most easily by means of the five schemata shown in Figure 1.

1. *Simple reaction time* (*SRT*), which Donders called the *A*-reaction, describes a single response (*R*) to a single stimulus, the *reaction stimulus* (*RS*). The single-stimulus–single-response condition for SRT distinguishes it from all the other paradigms (2–5) in Figure 1, which are examples of

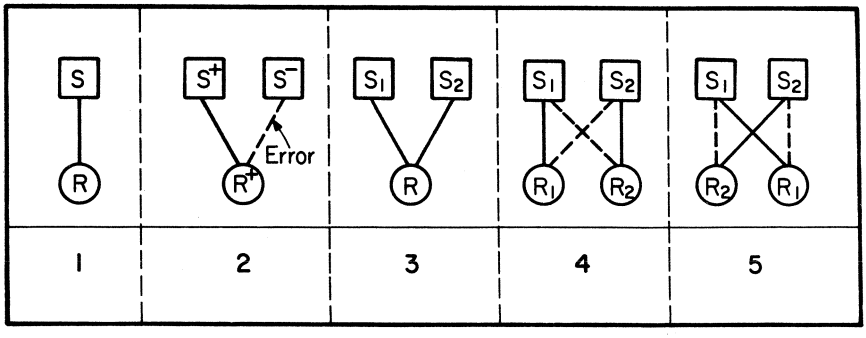

FIGURE 1. Schemata for classical RT paradigms. S, stimulus; R, response; solid lines, correct response; dashed lines, error response. (See the text for the name and an explanation of each schema.)

what Wundt originally called *compound* (or *complex*) reactions. These compound RT paradigms generally result in a longer RT than does the SRT paradigm. This is especially true of paradigms 4 and 5, which are examples of what is called *choice reaction time* (*CRT*). In SRT, the subject is instructed to respond (e.g., by either releasing or pressing a Morse key) as quickly as possible on the occurrence of the RS (e.g., a sound, or a light going on). Typically, a *preparatory stimulus* (*PS*) precedes the RS, usually by a random interval of from one to several seconds. The PS, which is often in a different sensory modality from the RS (e.g., PS, auditory signal; RS, visual signal), focuses the subject's attention on the RS and determines his readiness to respond. The duration of the *warning interval* (*WI*) is usually randomized from trial to trial to prevent the subject's learning to anticipate the occurrence of the RS precisely. Figure 2 shows the typical RT procedure. In this example, the beginning of the subject's response (R) terminates the RS. In another procedure, the RS has a set duration independent of the subject's response.

2. *Discriminative reaction time* (*DRT*), or Donders' *C-reaction*, requires that the subject discriminate between a *positive* and a *negative stimulus* (S^+ and S^-), but allows only one response. The subject should respond only to the S^+, and inhibit response on the occurrence of S^-. The task of discriminating between S^+ and S^- can be made easy or difficult, depending on the experimenter's purpose. It should be understood that S^+ and S^- (and all other alternative Ss in Figure 1) may appear in the same place or in different places. The DRT affords the possibility of false responses or errors (i.e., responding to S^-). To minimize the error rate, subjects may be

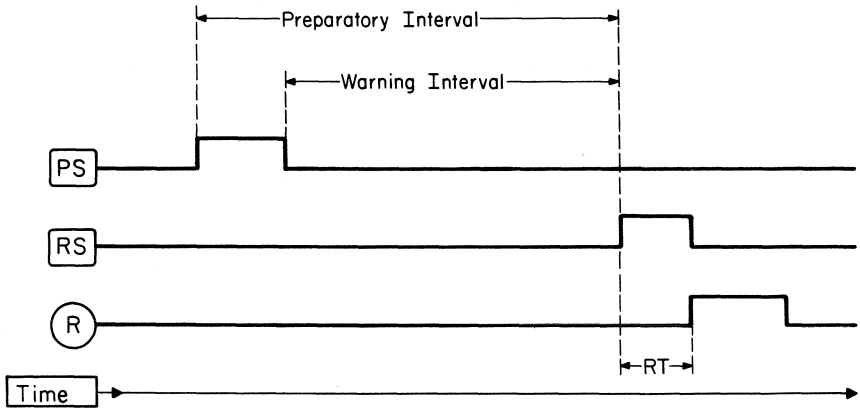

FIGURE 2. Paradigm for simple reaction time (SRT), and for all the other RT paradigms shown in Figure 1. PS, preparatory stimulus; RS, reaction stimulus; R response.

instructed to respond only as fast as they can without making errors. Even then, some errors will be made, and they are recorded as an essential part of the RT data.

3. *Nondiscriminative reaction time (NDRT)* (or *conjunctive reaction time*) was introduced by Wundt, who called it the *D-reaction,* in keeping with Donders' terminology of A, B, and C reactions. The NDRT differs from the DRT only in that the subject makes the only possible response to whichever stimulus occurs. It is a rarely used paradigm because the results are often only slightly different from those for the SRT. (Separate spatial locations of S_1 and S_2 tend to increase the RT.) Merely the *uncertainty* of the occurrence of S_1 or S_2 on each trial causes the RT in this paradigm to be slighly longer than in the SRT paradigm. As in SRT, there is virtually no chance for errors.

4. *Discriminative-choice reaction time,* or *CRT,* also known as Donders' *B-reaction* or *disjunctive reaction time* requires that different responses be made to different stimuli. Hence it involves *discrimination* between stimuli and *choice* of the appropriate response from among a number of alternatives. Paradigm 4 represents CRT with a high degree of S–R *compatibility,* that is, there is a close spatial correspondence (or some other form of close correspondence) between the S and R alternatives.

5. *Discriminative-choice reaction time* is here shown with a *low* degree of S–R compatibility. Both RT and error rate generally increase, the lower the S–R compatibility.

ANALYSIS OF ELEMENTARY COGNITIVE PROCESSES

The basic RT paradigms shown in Figure 1 can be used to illustrate simply how elementary cognitive processes can be distinguished and measured in terms of the time taken by each component process.

Two main methods have dominated the field: Donders' (1868/1969) *subtraction method* and S. Sternberg's (1969) *additive factor method.* (Because there are two noted psychologists in this field with the name *Sternberg,* Saul Sternberg of Bell Laboratories and Robert J. Sternberg—no relation—of Yale University, it is less confusing to affix their first initials whenever we refer to him.)

The Subtraction Method

In this method, it is assumed that information processing proceeds in a sequence of discrete mental events, each taking a certain amount of time. If the processing requirements of two tasks differ only in the presence or

absence of one of these mental events, or processing stages, the difference between the total time taken for each task is the time required for the one mental event in which the task requirements differ. Conversely, if the time taken by each of two tasks differs, it is presumed that the tasks differ either in the duration or the number of different processes required, or both.

To illustrate this in terms of Donders' classic paradigms, consider the processing requirements for SRT, DRT, and CRT (paradigms 1, 2, and 4 in Figure 1).

The SRT involves (1) sense organ lag; (2) afferent neural conduction time, from sense organ to brain; (3) apprehension of the S; (4) efferent neural conduction time from brain to muscle; and (5) muscle lag.

The DRT involves everything involved in SRT, *plus* (6) time for *discrimination* between S^+ and S^-.

The CRT involves everything involved in DRT, plus (7) time to *choose* between R_1 and R_2.

Hence, subtracting SRT from DRT (i.e., DRT − SRT) yields the *discrimination* time. And CRT − DRT yields the time taken to choose the correct response.

Interestingly, the first measurement of the speed of afferent nerve conduction in humans, by Helmholtz, in 1850, was based on the subtraction method, using only SRT. He applied the RS to the person's toe and to the thigh, and noted the difference in the RT. With this information, the speed of the sensory nerve impulse was calculated to be between 50 and 100 meters per second—less than one-third the speed of sound.

Donders' subtraction method has met with a number of criticisms. One is that its application presupposes that the investigator already has a rather clear concept of the discrete processes or stages involved in each of the tasks compared by the subtraction method. Thus, it begs the questions it is intended to answer. Another class of criticisms centers on the fact that the method does not allow the components of a task to interact: It is assumed that additional processing requirements can be inserted into a given task, or deleted, without in any way affecting the other processes involved in the task. The subtraction method, by itself, affords no means of objectively testing the validity of its assumption of "pure insertion" or complete additivity of the time required by each of the information processing elements involved in the task. Consider three tasks, A, B, C, which have reaction times $t_A < t_B < t_C$, respectively. We hypothesize that $t_B > t_A$ because task B involves all the processes involved in task A, *plus* process x, which is not involved in task A; and $t_C > t_B$ because task C involves process y (in addition to all the processes involved in task B). By subraction, then, $t_B - t_A = t_x$ and $t_C - t_B = t_y$. Now, if the processing stages x and y are

purely additive, as we have assumed in order to obtain their time values, then $t_C - t_A$ should be exactly equal to $t_x - t_y$. But it is obvious that this is a mere tautology, since, if $t_x = t_B - t_A$, and $t_y = t_C - t_B$, then $t_x + t_y$ *must* be equal to $t_C - t_A$ (i.e., $t_B - t_A + t_C - t_B = t_C - t_A$). It is therefore not an independent proof of the additivity of x and y. What is required is some way to determine whether stages x and y act in an additive or an interactive manner.

The Additive Factor Method

Introduced by S. Sternberg (1969) as an improvement over Donders' more limited subtraction method, the additive factor method also begins with the assumption that information processing proceeds in a sequence of stages, each involving different processes. Although it is assumed that the times for each of the processing stages are additive, the question of which inserted or deleted task requirements, or *factors* (in the analysis of variance sense), act additively or interactively is left open to empirical investigation. The finding of pure additivity of the factors, as shown by the absence of significance interactions in an analysis of variance, identifies the factors with different and separate processing stages, whereas the finding of an interaction between factors is interpreted as signifying that of the two (or more) factors, each affects some one-and-the same stage of processing. By means of a converging series of ingeniously planned factorial experiments, it is possible to infer a processing model for a given type of cognitive task in which the experimentally manipulable factors in the task and their interactions are assignable to different processing stages. S. Sternberg has applied the method to a number of RT tasks, including what is referred to later in this chapter as the S. Sternberg short-term memory scanning paradigm. More detailed expositions of the additive factor method are to be found in S. Sternberg (1969), Pachella (1974), and Welford (1980, see index).

As a simple example of how the additive factor method works, consider the RT paradigms 4 and 5 in Figure 1. The factors here are presumed to be (1) stimulus discrimination (S_1 vs. S_2) and (2) response choice (R_1 vs. R_2). Each of these factors can be experimentally varied. For simplicity, say we have two *levels* of factor 1—high *versus* low stimulus similarity (discriminability)—and two levels of factor 2—high *versus* low S–R compatibility (e.g., paradigm 4 *vs.* paradigm 5 in Figure 1). The RT tasks with every possible combination of the 2 factors × 2 levels—four tasks in all—would be administered to four independent randomized samples of a pool of subjects. The analysis of variance of all the RT data (i.e., the *mean*

RTs of each subject as the unit of analysis) would have four terms:

	Source of variance	*df*
Main effects	Between factors	1
	Between levels	1
Interaction	Factors × Levels	1
Residual	Subjects within groups	$N - 4$

If the main effect of factors is significant and substantial and the interaction term is nonsignificant (Figure 3A), we would conclude that the two factors (stimulus discrimination and response choice) occur in two separate stages of information processing. A significantly large Factors × Levels interaction (Figure 3B), however, would mean that some stages of processing involve *both* factors, perahaps to the exclusion of separate stages involving one factor each. It should be noted that for the interaction term to be cogently interpreted, the data used for the analysis must consist of the RT measurements *per se* (or the *arithmetic means* of these measurements), which are expressed in units of real time, rather than any scale transformation of the measurements. Also, *median* RTs are ruled out for this type of analysis, as medians are not necessarily additive, whereas arithmetic means are always additive.

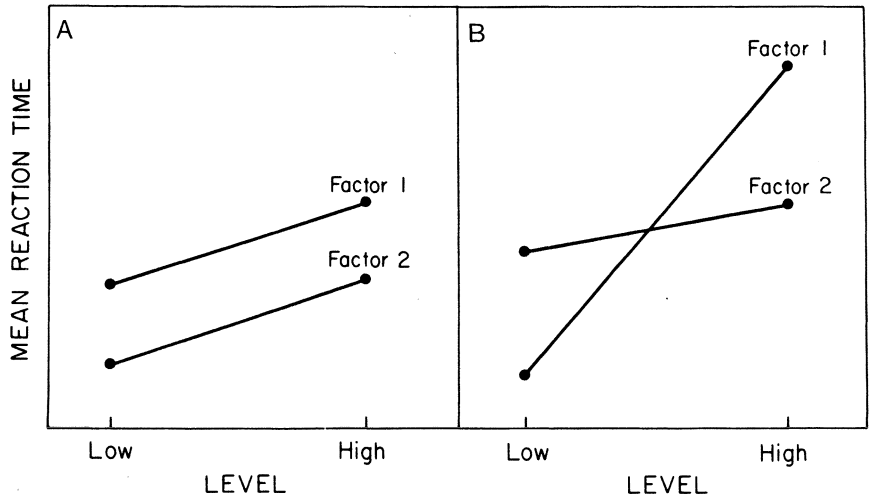

FIGURE 3. Graphical representation of the analysis of variance of the factorial experiment described in the text in which (A) there is a main effect for both factors and levels, but no interaction between them, and (B) a main effect for both factors and levels and an interaction between factors and levels.

Inferential ambiguities are now recognized in both the subtraction method and the additive factor method, and most present-day investigators hold that each method has valid uses under the special conditions for which it is appropriate and that both methods can be used in a complementary fashion. Carroll (1980) expresses a consensus that "the method of 'converging operations' (i.e., the accumulation of evidence from a variety of related studies) can be expected eventually to produce scientifically valid results and interpretations" (p. 63).

Elaborations of Classical Reaction Time Paradigms

The classical RT paradigms just described are the prototypes for almost countless modifications and elaborations designed to study a variety of cognitive phenomena. There are many variations in the number and character of the stimulus and response alternatives. In addition to presenting some stimulus situation (S) and having some mode of overt response (R) to it, the only common feature of all paradigms is the precise measurement, usually in milliseconds, of the amount of time that elapses between S and R.

The main experimental variations most commonly encountered in the literature are the following.

Experimenter-Paced versus Subject-Paced Presentation of the Stimulus. Reaction time experiments always use repeated trials in order to minimize measurement error. Each trial can be conceived of as a cycle comprising the sequence of events as depicted in Figure 2. The cycle may be initiated by either the experimenter or the subject. We have found that self-pacing of trials by the subject is especially desirable when the task requirements are fairly complex and repeated trials may incur fatigue.

Presence or Absence of a Preparatory Stimulus (PS). Experimenters now rarely omit the PS (see Figure 2) because its use not only focuses attention on and shortens the RT, but it also decreases its trial-to-trial variability, thereby yielding a more reliable measure when the RTs are averaged over n trials.

Single Location versus Separate Locations of Multiple Stimuli. In discriminative RT or choice RT, the two (or more) stimuli may be randomly presented (sequentially), either in the same location (e.g., different-colored lights appearing in a single aperture) or in spatially separate locations. When the different stimuli to be discriminated appear at the same visual fixation point on every trial, variance due to eye movements, or visual scanning, is minimized, as compared with random presentation of the stimuli in separate locations.

Sequential versus Simultaneous Presentation of Stimuli. When multiple stimuli are presented, such as a string of numbers, letters, words, or

symbols, they may be displayed either sequentially or simultaneously. An example of a paradigm that has used both sequential and simultaneous displays in different studies is S. Sternberg's short-term memory scan paradigm. A set of digits (varying in number from one to seven) is displayed, either simultaneously or sequentially (at a rate of, say, two digits per second), and is followed immediately by a single "probe" digit, which serves as the RS. The subject responds *yes* or *no* (usually by pressing buttons labeled *yes* and *no*), depending on whether the probe digit was or was not a member of the previously displayed set of digits. The RT is the interval between the onset of the probe digit and the subject's response.

Variations of Response Mode. The RT for verbal responses can be measured by a voice-activated key. Except when the experimenter is studying the speed of word associations or features of vocalization *per se,* the use of a voice key has certain disadvantages, for example, time variations in the initiation of pronouncing different words.

When only a small number of alternative responses is presented, finger-activated response keys are usually preferable. In the classical CRT experiment, the subject poises the index fingers of the left and right hand lightly on two Morse keys, ready to make the appropriate response. The response to the RS may consist of *releasing* one of the keys (when both keys are initially depressed) or of *pressing* one of the keys.

When more than two response alternatives are required, any number of keys, up to ten, can be used, each one activated by a different finger. At the beginning of each cycle, the subject's fingers are poised lightly on the keys (for a press response by one finger) or they depress all the keys (for a release response). Multiple response keys used this way have the distinct disadvantage of unwanted variance because the muscular capabilities of the right and left hands, and of the different fingers of each hand, differ. As every pianist knows, the ring finger of each hand is comparatively weak and inept.

To overcome this problem, we have introduced a procedure that uses a *home button* (Jensen & Munro, 1979). It has been effectively adapted to several different RT paradigms. The procedure divides the subject's response into two separately measurable acts, RT and *movement time* (*MT*). The simplest example of this procedure can be described for the CRT paradigm, as shown in Figure 4. At the beginning of a cycle, instead of the subject's having the index fingers of each hand readied for responding to the R_1 or R_2 buttons, the index finger of the preferred hand depresses H. Immediately on the appearance of the reaction stimulus (RS = S_1 or S_2), the subject removes the index finger from H and presses R_1 or R_2. The RT, also called *decision time* (*DT*) in this procedure, is the interval between the onset of the RS and the release of H. The interval between releasing H and

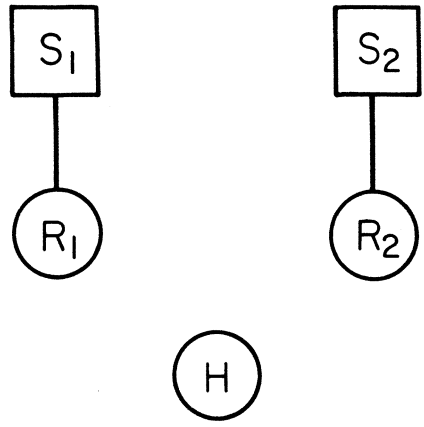

FIGURE 4. Choice reaction paradigm, using a home (H) button. The cycle begins by the subject depressing H with the index finger. As soon as S_1 or S_2 appears, the subject presses the appropriate response button (R_1 or R_2) as quickly as possible.

pressing R_1 or R_2 is the MT. The spatial distance between H and R is usually not more than a few inches. The same procedure can, of course, be used for simple RT and for CRT involving many more than just two S–R alternatives. (In subject-paced trials, each cycle is initiated by the subject's pressing down H.) The main advantage of the H procedure is that it permits RT to be measured by exactly the same form of response (i.e., simply raising the index finger of the preferred hand) regardless of the number of S–R alternatives. It is a remarkable fact that RT rather than MT increases as a function of the number of S–R alternatives, at least within a range of one to eight alternatives. The response alternative buttons should be so arranged that they are located at equal distances from H.

An apparatus, shown in Figure 5, for measuring both SRT and varying degrees of multiple-choice RT (all with maximal S–R compatibility), based on the H procedure, has been used extensively in our Berkeley laboratory (Jensen & Munro, 1979). The subject's console of the apparatus for measuring the subject's RT and MT consists of a panel, 13 × 17 in., painted flat black, and tilted at a 30° angle. At the lower center of the panel is a red pushbutton, $\frac{1}{2}$ in. in diameter, H. Arranged in a semi-circle above the H are eight red pushbuttons, all equidistant (6 in.) from H. One half an inch above each button (except H) is a $\frac{1}{2}$ in. faceted green light. Different flat black panels can be fastened over the whole array, to expose arrays having any number of light–button combinations. (We usually use one, two, four, and eight alternatives, which correspond to zero, one, two, and three bits of information, when information is measured as $\log_2$ of the number of S–R alternatives.)

The subject is instructed to place the index finger (of the preferred

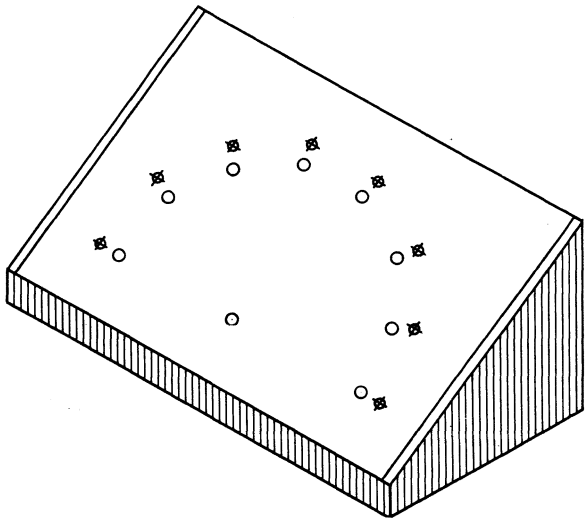

FIGURE 5. Subject's console of the RT–MT apparatus. Pushbuttons are indicated by circles, faceted green lights by crossed circles. The home button is in the lower center, 6 in. from each response button.

hand) on H. Then an auditory preparatory signal is sounded (a high-pitched tone of 1 s duration), followed, after a continuous random warning interval of from 1 to 4 s, by one of the green lights going "on," which the subject must turn off as quickly as possible by touching the sensitive microswitch button directly under it. The RT is the time the subject takes to remove his finger from the H after the green light goes on. The MT is the interval between removing the finger from H and touching the button that turns the green light off. The RT and MT on each trial are separately registered in milliseconds by two electronic timers.

Carroll (1980) has devised a useful method for representing, in a highly detailed fashion, both the task requirements and the hypothesized cognitive processes involved. Carroll calls this type of diagram the *dual time representation (DTR) of elementary cognitive tasks (ECTs)*. It is most useful for highlighting the precise (and often crucial) procedural differences between various chronometric paradigms as actually performed in the laboratory. It is illustrated in Figure 6 with respect to the RT–MT procedure just described. The DTR of another variant of the choice RT paradigm, used in a study by Keating and Bobbitt (1978), is shown in Figure 7 to illustrate how the DTR flow diagram depicts all the fine-grained procedure differences between tasks that can yield differences in results.

Carroll (1980) describes the DTR representation as follows:

> Objective (observable) stimulus and response events are shown along the central time axis that runs from upper left to lower right. The remaining space in the chart is available for other purposes. The upper triangle (above the diagonal axis) is used for representing presumed mental or "cognitive" processes, their duration and effects over time, and their interrelationships and interactions with stimulus and response events and with each other. The lower triangle can be used for such purposes as annotating stimulus variations, depicting repetitions of events (as by the "repeat signs" of musical notation), and showing measurement procedures (e.g., time measurements). The distances on a DTR chart are regarded only topologically, i.e., they show only temporal order relationships among events, but do not necessarily represent, to scale, the exact occurrence times or the durations of events.
>
> Various further conventions can be established in designing DTR charts. In representing objective events, those that are obligatory (i.e., that are always present and are characteristic of the task) are shown in solid-line boxes. Optional events are shown in broken-line boxes. Broken lines bordering the lower right of a box can be used to indicate that an event (e.g., the shining of a light) persists for an indefinitive period, or until some other event supersedes it . . . .
>
> "Cognitive" (nonobservable, but presumed) events may be shown in "cartouches" [boxes with rounded corners] placed in the upper triangle of the chart in such a way as to show assumed precursors and consequences of such events and their temporal relationships. . . . Lines, generally with direction of effect shown by arrows, show presumed causal connections and interactions of cognitive events with objective events and with each other. (pp. 13–14)

Double Stimulation. It is often of interest to measure the change in RT that occurs when the subject has to process two sources of information simultaneously rather than just one. Invariably, RTs to double stimulation are slower than RTs to a single stimulus. A simple example of the double-stimulation discrimination paradigm would be the simultaneous presentation of a tone (high or low) and a light (blue or yellow), with two response keys, and instructions to press (or release) the left key (with the left index finger) only when the high tone is sounded and to press (or release) the right key (with the right index finger) only when the blue light goes on. Such double tasks strain the subject's limited channel capacity and generally increase RT, as well as the error rate, considerably.

Double-stimulation tasks can also use successive stimuli, as in the study of processing-storage trade-off. For example, say the subject must respond "true" or "false" (by pressing keys labeled T and F) to simple addition problems (e.g., $3 + 4 = 7$) that are either correct or incorrect. Two such problems, labeled A and B, respectively, are presented one after the other in quick succession, immediately followed by the reaction stimulus (letter A or B) that post-cues the problem to which the subject must respond T or F. More complex variations of this paradigm have been

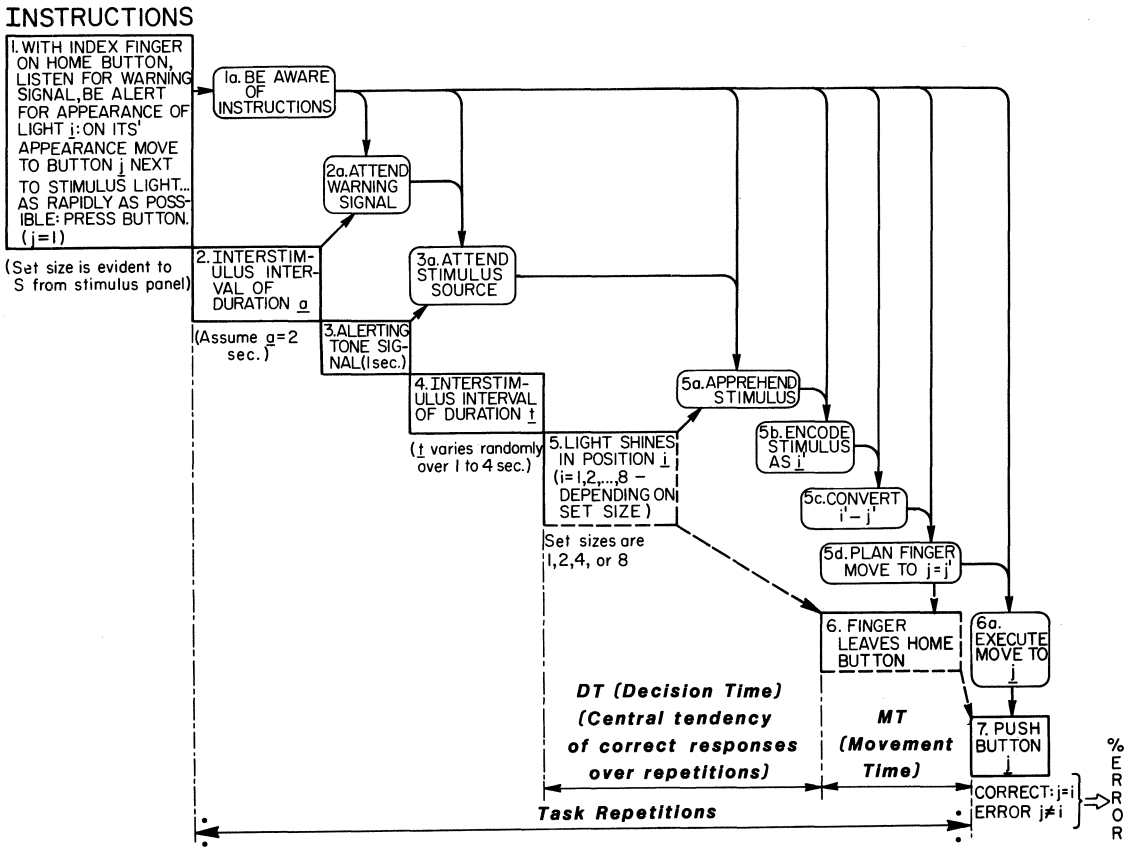

FIGURE 6. Carroll's dual time representation (DTR) of Jensen's RT–MT procedure. From *Individual Difference Relations in Psychometric and Experimental Cognitive Tasks* (p. 18) by J. G. Carroll, 1980, Chapel Hill, N.C.: L. L. Thurstone Psychometric Laboratory, University of North Carolina. © 1980 by J. G. Carroll. Adapted by permission.

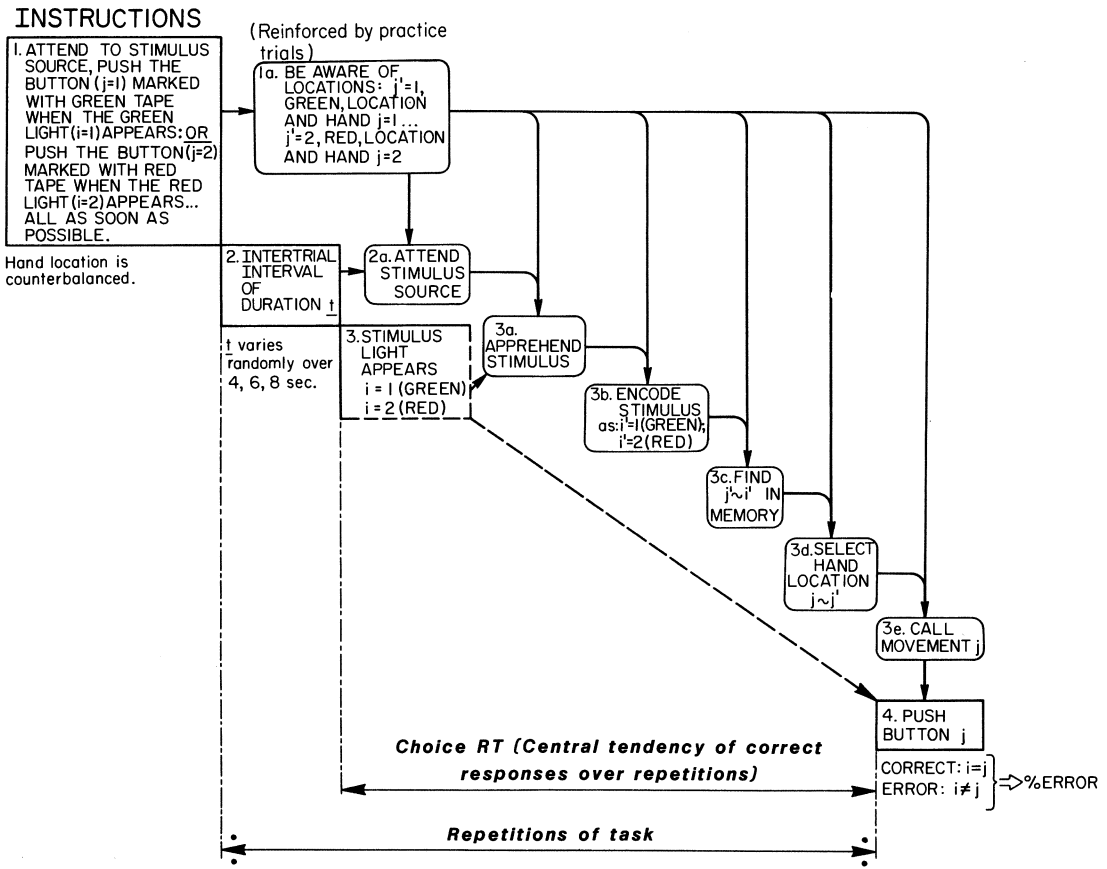

FIGURE 7. Dual time representation (DTR) of Keating's and Bobbitt's choice RT procedure. From *Individual Difference Relations in Psychometric and Experimental Cognitive Tasks* (p. 15) by J. G. Carroll, 1980, Chapel Hill, N.C.: L. L. Thurstone Psychometric Laboratory, University of North Carolina. © 1980 by J. G. Carroll. Adapted by permission.

used. (Double-stimulation paradigms, theoretical models, and sample experiments have been thoroughly reviewed by Kantowitz, 1974.) There is evidence that RTs measured by double-stimulation procedures are somewhat more highly correlated with IDs in general mental ability than are RTs to either stimulus presented singly, probably because of the greater information load and the strain on channel capacity occasioned by the double stimulation.

PROCEDURAL VARIABLES THAT AFFECT REACTION TIME

The results of RT studies are extraordinarily sensitive to a large number of factors in which procedures and subjects may differ. Investigators should be fully aware of all these factors in designing their experiments and in comparing the results of different studies. Variations in results of studies cannot be evaluated in the absence of explicit reports of the procedural and subject variables that are known to affect RT. A knowledge of these variables and of their interrelationships is also of importance theoretically, in that they afford an essential part of the network of empirical clues to the psychological and physiological nature of mental speed and its manifestations in performance on psychometric tests, as well as their practically significant correlates.

Preparatory Factors

The subject's state of expectancy and attention just before the RS is called *preparatory set*. It is largely a function of the experimenter's instructions to the subject (e.g., emphasizing speed *or* accuracy), and especially, of the PS and warning interval (*WI*). The RT will be shorter and less variable from trial to trial if a PS precedes the RS than if there were no PS. The duration and intensity of the PS should be sufficient so as to leave no doubt of its occurrence. We have found a constant PS duration of 1 s to be about optimal. The subjective intensity of the PS should not exceed that of the RS, especially if they are in the same sensory modality. The optimal WI is 1 to 2 s, but WIs in the range of 1 to 4 s are approximately equivalent in their effect on RT. Warning intervals shorter than 1 s or longer than 4 s result in a slower RT. A WI of constant length results in a gradual shortening of RT as the subject develops an expectancy of the precise occurrence of the RS. This expectancy effect can be overcome by using *random* WIs within the range 1 to 4 s. A PS of a different sensory modality than the RS (e.g., PS, auditory; RS, visual) is a decided advantage, especially when the subjects are young children or the mentally retarded. Distinct sensory modalities for the PS and the RS offer much less chance for

confusion and help to minimize the role of learning in the subject's RT performance.

Little attention is paid in the RT literature to the *intertrial interval (ITI)*, also called *afterperiod*, that is, the interval between the subject's last response to the RS and the reappearance of the PS. The ITI should always be distinctively longer than the WI. The risk of mental fatigue in complex RT tasks is reduced by the subject pacing the trials. Each cycle is initiated by the subject's pressing H, whereupon the preprogrammed cycle runs off automatically. There should be a constant interval between the subject's depressing H and the occurrence of the PS. (We have used a constant H–PS interval of 1 s with good results.)

Stimulus Factors

The RT varies according to the *sensory modality* of the RS because of differences in peripheral mechanisms. For example, visual lag is greater (by 30 to 40 ms) than auditory lag, probably because the former is initiated by a chemical process and the latter by a mechanical process. Also, central (foveal) vision results in a faster RT than peripheral vision. Tactile stimuli and a mild electric shock result in about the same RT as auditory stimuli.

Stimulus *intensity*, *area* (as of a light source), and *duration* are all positively related to a faster RT, but there is not a monotonic relationship at the extremes of these variables.

A greater *complexity* of the stimulus or a greater *number of alternatives* in the location or form of the RS or less *discriminability* of the alternate RS all result in a slower RT. Hick (1952) has noted that in CRT, the RT increases linearly as a function of the number of *bits* of information. A *bit* is defined in information theory as $\log_2 n$, where n is the number of choice alternatives. A *bit* can be thought of as the amount of information that will reduce uncertainty by one-half. Figure 8 shows this relationship, now known as *Hick's Law*.

Response Factors

The RT is facilitated by moderate increases in *muscle tension*, which, under normal conditions, is an index of cortical arousal. It has been found that the forearm muscles to the hand that executes the response become tense during the WI (see Woodworth & Schlosberg, 1954, pp. 30–32). The subject's *concentration* on the response to be made also speeds the RT.

When the subject's fingers are poised closely above the keys, ready to respond, *finger tremor* will affect RT in a variable fashion from trial to trial. Responses are synchronized with the tremor, so that RT is faster if the RS occurs when the tremor is in the downward phase of its movement. Control of the subject's *motivation* for fast response by means of a reward or

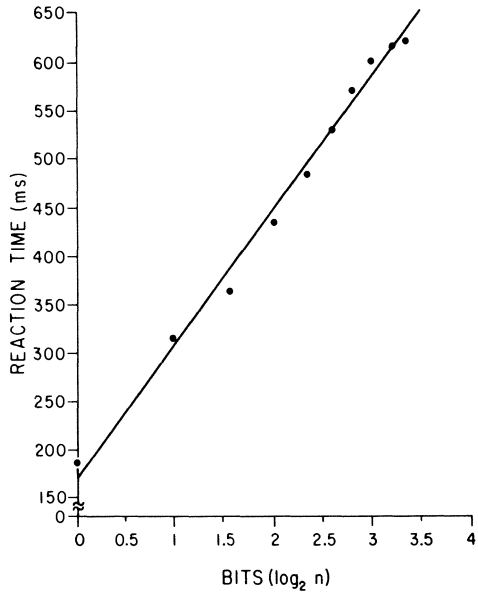

FIGURE 8. Mean choice RTs to stimulus arrays conveying various amounts of information scaled in bits; *n* is the number of choice alternatives. Data from Merkel (1885) as reported in *Experimental Psychology* (p. 33) by R. S. Woodworth and H. Scholsberg, 1954, New York: Holt, Rinehart and Winston. Copyright 1954 by Holt, Rinehart and Winston. Adapted by permission.

punishment or an immediate knowledge of the results will speed up RT beyond the subject's normal "best" effort without such incentives.

Practice Effects speed the RT (and lower the error rate in CRT) and are generally found over the course of many repeated trials, but they are often so small as to be practically negligible, especially for SRT. Practice effects become more prominent as the task requirements are made more complex. But practice effects differ greatly in magnitude and in number of trials to asymptotic performance, depending upon the task requirements and the characteristics, such as age and ability, of the subjects.

Extrinsic Organismic Factors

Anoxia, as at high altitudes, slows RT, and CRT is more sensitive than SRT. Stimulant drugs, such as caffeine, tobacco, and amphetamine, speed the RT. Depressant drugs generally slow the RT, but alcohol has a diphasic effect, at first speeding and later slowing the RT. In general, drugs that alter synaptic thresholds and hence synaptic connectivity also alter RTs in the predictable direction.

Intrinsic Organismic Factors

Age greatly affects the RT, which decreases in a negatively accelerated fashion from early childhood to maturity, plateaus between about 25 and 55 years of age, and gradually increases again in old age. The RT increases rapidly the few months before death, just as performance on IQ tests has also been found to deteriorate markedly in the few months before death.

Sex differences in RT generally are found to favor males slightly, but in studies in our laboratory, using H that permits separation of the subject's response into RT (or decision time) and movement time (MT), we have found a significant sex difference (males faster) only for MT. *Physical exercise* and general *fitness* speed the RT.

The RT varies throughout the day as a function of changes in body *temperature,* with higher temperatures making for faster reactions. The SRT varies about 9 or 10 ms per degree Fahrenheit change in body temperature in the normal range of diurnal variation. Woodworth and Schlosberg (1954) note that "the amount of [RT] change [with temperature] corresponds pretty well to what would be expected from the temperature coefficient of chemical processes, and suggests that the cerebral process in reaction depends closely upon chemical activity" (p. 38). It is theoretically noteworthy that CRT shows much greater shifts with change in temperature than does SRT.

Food intake also slows the RT, over and above the drop in temperature that follows eating. There is a post-lunch slowing of RT, which, of course, contributes to the variability among subjects who are tested at different hours throughout the day and may slightly attenuate the correlation between RT and other variables that are less sensitive to physiological state or are measured at some other time of the day.

Strangely, *body build* affects RT, with more slender persons having the faster RT. The index of body build found to show the highest correlation (about .3) with SRT is the ratio of (height)/(weight)$^{\frac{1}{3}}$.

In general, factors that slow RT also tend to increase its trial-to-trial variability.

CHRONOMETRIC APPARATUS AND TECHNIQUES

Apparatus

A great diversity of mechanical, electrical, and electronic equipment has been used to measure RT. The common aim is the precise and reliable measurement of very brief time intervals, and this has been largely achieved since the earliest studies of RT. What has changed in this field is not so

much the precision of the measuring instruments, but rather their dependability of operation and the silence, compactness, general efficiency, convenience, and ease of reading or recording measurements. The older chronoscopes, for example, required frequent adjustment and calibration, which is virtually obviated by modern electronic equipment using crystal timing devices that oscillate at constant frequencies ranging above 10^3 cycles per second (cps).

The RT should be measured in units of 10^{-3} s, or in milliseconds, with an error of less than .1%. This is routine with modern electronic timers.

Aside from the precision of the timers, there is virtually no standardization of the chronometric equipment used by experimenters. The reason for this lack of standardization of the stimulus and response modes is the great diversity of purposes served by chronometry in modern cognitive psychology. Every investigator adapts the experimental arrangements idiosyncratically to the requirements of a particular kind of study. The commercially produced RT apparatuses, which can be purchased from the suppliers of psychological laboratory equipment, usually have a sufficiently accurate timer, are relatively inexpensive, and, for any particular manufacturer, the S–R features are highly standardized. But these apparatuses are so simple and so inflexible as to be hardly adaptable to the great variety of experimental task arrangements required for mental chronometry experiments. The commercially available RT apparatuses are scarcely useful for anything but simple demonstrations and exercises in the undergraduate psychology laboratory. Hence, most professionals today use equipment that is custom-built to their specifications.

The great disadvantage of each experimenter using his or her own custom-built RT equipment is the lack of standardization from one laboratory to another. Consequently, attempts to replicate experiments in different laboratories, using highly comparable groups of subjects, often result in replication of the same *relationships* among RTs obtained under various experimental conditions, but not the same *absolute values* of RT. This always poses the question of whether the absolute differences in RT found in different studies result from the use of nonstandardized apparatus or from a difference between subject pools. This is an especially undesirable state of affairs when chronometric techniques are used to study individual or group differences, for one of the potential advantages of chronometry, as compared with traditional psychometry, is that measurements can be made on an absolute or ratio scale. Nonstandardized RT equipment lessens this advantage when findings from different laboratories are compared. The significance of this observation has been impressed upon the present writer as a result of his contacts with other investigators who have duplicated the particular RT–MT equipment (shown in Figure 5) that was originally

devised by the writer for his own chronometric studies of the relationships between SRT and CRT and the *g* factor of intelligence tests (Jensen, 1982a,b). At last count, this apparatus, or the subject's response console, has been nominally duplicated in nine different laboratories in the United States. Despite the attempt to make the consoles as much like the original as possible, from the dimensions and descriptions provided, they all differ slightly, not in the physical measurements between the various S and R and H components or in the precision of the reaction timers, but rather in such subtle (but crucial) features as the lag in the microswitch response buttons and the pressure required. When we have tested the same subjects on different instruments, we find significant variations in the average RT because of these seemingly slight differences in equipment. Although the same *relative* differences and correlations with other variables replicate dependably, comparisons of absolute RT values between experiments performed even with only subtly differing apparatus is scarcely justified. (In order to compare RTs between groups that have had to be obtained in different localities validly, we have lent and transported our original apparatus to laboratories as far separated as Canada, northern and southern California, and Arizona.) It is clear that if S–R consoles are to be properly standardized, every component must be standardized. Ideally, prototype apparatuses that are intended to replicate experimental results with exactly the same absolute values of RT (except for sampling error) should all be constructed by the same manufacturer, using identical components, materials, and so forth.

The most expensive parts of any RT apparatus are the timers, but they are now also highly perfected and standardized, even when obtained from different manufacturers. They are the least of the problem, which resides in the lack of standardization of the stimulus display and the response console.

We have found the modern microcomputers, such as the Apple II and Apple III, to be a boon to mental chronometry. These computers are equipped with highly precise timing mechanisms and display screens and are programmable, so that the entire sequence of stimulus and response events can be run automatically. The whole program for an experiment lasting an hour or so can be stored on a magnetic tape cassette, and with a suitable (commercially available) attachment, can be read into the computer in just a few minutes. Also, with available attachments, all the subject's RTs can be recorded on magnetic tape and/or printed out on paper tape for a later detailed analysis. The computer can also be programmed to calculate summary statistics (e.g., mean, median, standard deviation) or frequency distributions of the subject's RTs over trials for each experimental condition—all available within a few seconds of the end of the testing.

The only component of the commercial microcomputers that we have

found inadvisable for the laboratory measurement of RT is the computer's keyboard as the subject's response console. The computer's keyboard should be reserved only for programming the computer and for giving it "instructions" by the experimenter. In the first place, if it is used as a response console, parts of the keyboard have to be masked, exposing only those few keys germane to the requirements of the experiment, and these exposed keys usually have to be relabeled. Also, the small size and close spacing of the keys tends to inhibit fast response. But the main disadvantage is that the relatively delicate, expensive keyboard mechanisms of computers are not ideally suited to take the constant "beating" a response console is subjected to when many persons are run through hundreds of trials in RT experiments. Therefore, we have devised special response consoles that can be connected by a cable to the computer. In addition to the RT–MT console shown in Figure 5, we have a general-purpose response console that permits the measurement of RT and MT in all chrometric paradigms that call for any form of binary response (e.g., yes–no, true–false, same–different, odd–even, red–green, + −). It consists of a panel with a home button and two response buttons; each button is equidistant from the others, with the apex of the "triangle" toward the subject. The microswitch buttons, which make instant contact with a very light touch, are about the size of a half-dollar; their centers are about 6.4 cm apart. Appropriate magnetized labels that can be easily changed are placed just above each response button. Inside the console is a small sound generator that delivers a computer-programmed "beep" as the PS. The elementary cognitive task, including the RS, is presented visually on an alphanumeric display screen attached to the response console. (Videoscreens are ideal for stimulus display.) The computer itself, which controls the experiment, need not be in view of the subject. In any case, its operation is silent and thus unobtrusive. The subject's console is shown in Figure 9. As a general rule, the home button and response buttons should be fairly large, to minimize the purely motor-skill aspects of the task, and should make contact with very little pressure, to minimize the effect of differences in finger strength and fatigability. Work, in the physical sense of Force × Distance, should be reduced to the absolute minimum in the response requirements of a chronometric apparatus. This becomes especially important when young children or elderly persons are tested.

Also, it is essential that the subject's console of any RT–MT apparatus be designed so that the RT timer will not register the subject's response if the subject's finger releases H before the RS appears. In other words, it should be impossible to activate the electrical connection between H and the reaction timer until the instant the RS appears. This arrangement helps to

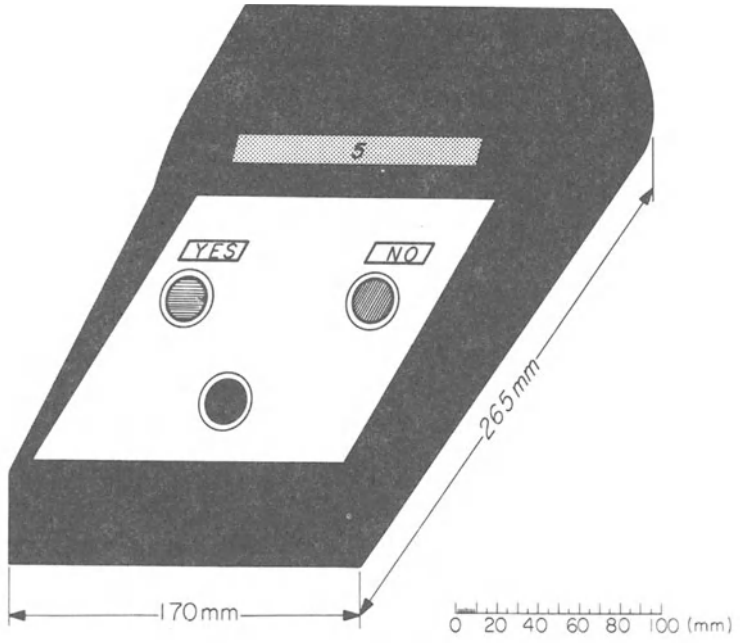

FIGURE 9. A binary response console, with an alphanumeric display unit. The lower button is the home, the upper two buttons are for binary choice responses (here labeled Yes and No) for indicating the presence (or absence) of the probe digit (shown in the display) in the "positive set," in S. Sternberg's short-term memory-scan paradigm.

prevent anticipatory flukes being included with authentic RT measurements.

Procedures

The most general procedural principle for RT studies has been succinctly stated by Nettelbeck (1980): "First, all subjects . . . must understand what is required, and no subject should be influenced or disadvantaged by factors in the experimental situation not accounted for—for example, insufficient practice, fatigue or undetected sensory or physical disabilities" (p. 384).

Subjects should be seated during testing to avoid fatigue. An adjustable chair or stool is advisable, especially for children of varying age, to ensure approximately the same physical relationship betwen the subject and the S–R console for all subjects.

In our experience, a session involving any one type of RT should not last longer than about one-half an hour when testing normal young adults. We usually make the testing sessions even shorter, often using the remainder of the hour the subject is in the laboratory to administer paper and pencil tests or individually administer such IQ tests as the Wechsler. The attentional requirements of RT tests make them more demanding and fatiguing than the usual psychometric tests, and unless the experimenter is explicitly studying persistence of attention and resistance to fatigue, the testing sessions should be kept short. Subjects can more easily take two rather different RT tasks, each lasting 15 min, than they can take either task alone if it lasts 30 min. Children and older adults and the mentally retarded must be given even greater consideration in this respect.

The RT varies throughout the day for a given person, and there are IDs in this variation. The best time of day for testing, therefore, is problematic when the main object of study is IDs in RT. Generally, the individual diurnal variations in RT simply constitute error variance in the measurement of IDs in RT. The only way it can be reduced is by testing the same subject on two or more days at different times each day and using the average RT over days (see the later section, *Reliability and Stability*). The least desirable time for measuring RT is any time within 1 hr or so after the subject has eaten lunch. Alcohol, drugs, medication, or illness of any kind may also affect the measurement of IDs in RT. It should be kept in mind that RT is considerably more sensitive to the subject's momentary physiological state than are psychometric tests.

Instructions to the subject are a crucially important part of the procedure for measuring RT. Variations in instructions can significantly affect the results, even when the testing procedure is the same in every other way. It is most important that the subject fully understand the task requirements and the features of his performance (e.g., speed and accuracy) that are being measured. The subject's ability to grasp and retain the instruction throughout the testing should not be a significant source of variance in the measurements. They are merely prerequisites for taking the test, and the experimenter must obtain evidence that all subjects are virtually equal in ability to comply with the task requirements, even if different subjects need different amounts of time for instruction and practice trials. Young children and retarded persons often need a demonstration of the required performance by the experimenter, so as to learn the procedure by imitation. Practice trials should be given until the subject performs confidently and consistently all the task variations that will be used in the experiment proper. For this purpose, we have made up brief practice sets that incorporate all the conditions of the experiment the subject will encounter. Subjects who cannot perform easily and consistently on the

practice set after a number of attempts (the number depending upon the time available and the supply of subjects) are dismissed as being unable to meet the minimal prerequisite skills to serve as subjects, whatever the reason. The task demands of most chronometric experiments are so simple as to be easily mastered almost immediately by normal young adults. For a new procedure, a pilot study with several subjects who are typical of those to be tested in the study proper should be performed to discover any problems that may arise in instructing subjects and to determine the effect of practice on the subjects' performance of the task. If there is a marked practice effect (i.e., improvement in performance) over the first n trials before an approximately asymptotic level of performance is attained, it is advisable to require n practice trials before beginning the experiment proper. (A typical learning curve can be plotted, with mean RT shown as a function of number of practice trials.) The reason for this requirement is that in chronometric studies we are usually more interested in the speed of reaction to various stimulus conditions than in the rate of learning the particular skills that are prerequisite for the subject's performance. Hence, a significant practice effect over trials usually indicates a source of variance that is extraneous to the experimenter's interest. The importance of measuring only RT performances that are close to asymptote, however, can be determined only by the particular purpose of the study.

Another important consideration is the relative emphases on speed and accuracy in the instructions. The speed–accuracy operating characteristic of a task depends on its complexity. The simpler the task, the less will be the effect on RT or on error rate of instructions that differentially emphasize speed and accuracy of responses. A speed–accuracy operating characteristic curve is shown in Figure 10. In this graph, the theoretical definition of RT is the minimal time required for correct response. It is seen that both RT and performance accuracy increase as accuracy is emphasized at the expense of speed. *Normal instructions* would be something like "We want to measure how fast you can respond without making errors." With these instructions, even highly practiced subjects will make 2% to 3% errors in fairly simple RT tasks, and the error rate will be considerably higher in complex tasks. Error rates are lowered if the subject immediately receives informative feedback as to whether each response was "correct" or "an error." It should be made clear to subjects that in addition to the measurement of RTs, the number of correct and error responses is recorded. Task difficulty and instructions should be adjusted in such a way as to maintain a low error rate and one that is fairly uniform across the various experimental conditions of the chronometric paradigm (e.g., the different numbers of light-button alternatives in the RT–MT paradigm). When error rates differ markedly across different experimental conditions, the interpretation of the cor-

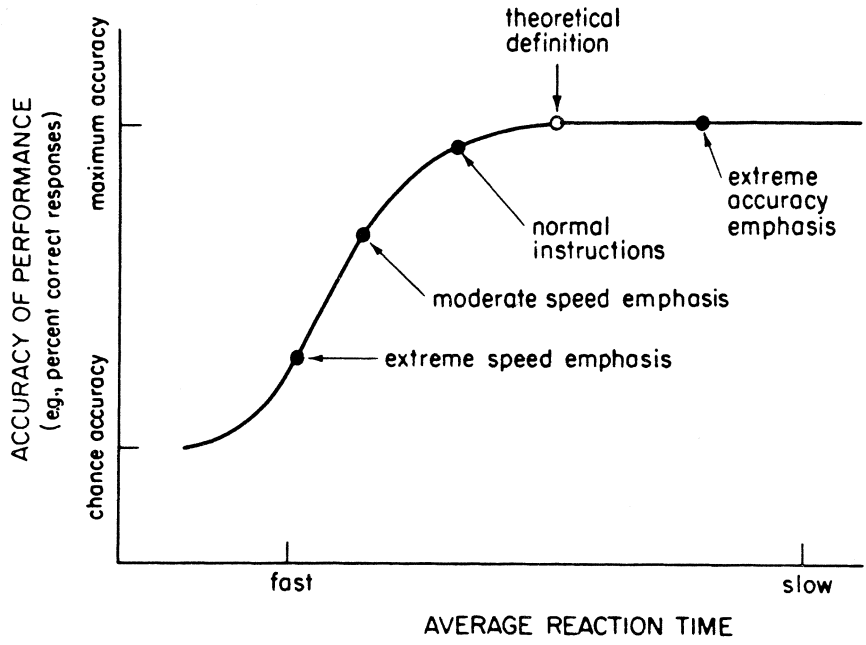

FIGURE 10. An idealized speed-accuracy operating characteristic. From "The Interpretation of Reaction Time in Information-Processing Research" by R. G. Pachella in *Human Information Processing: Tutorials in Performance and Cognition* (p. 59), B. H. Kantowitz, Ed., 1974, Hillsdale, N.J.: Lawrence Erlbaum Associates. Copyright 1974 by Lawrence Erlbaum Associates. Adapted by permission.

responding observed differences in RT becomes problematic. (See *Speed–Accuracy Trade-Off* in the following section.)

Finally, investigators should be aware that IDs and experimental effects can hardly be studied in-one-and-the-same procedure. Experimental psychologists are accustomed to thinking in terms of experimentally varying task conditions across subjects so as to randomize out certain unwanted sources of variance. This is rarely feasible in a single study of IDs and, moreover, it is usually undesirable. Beyond slight variations in instructions and preliminary practice to ensure that all subjects understand the task, the conditions must be *uniform* for all subjects. For example, the entire sequence of the particular S–R conditions over trials must be invariant for all subjects. Even if the sequence is *random,* it should be the *same* random order for everyone. Response repetition on successive trials (as contrasted with making different responses on successive trials) is known to affect RT, which is faster for a repeated response than for a varied response (Kirby,

1980). Hence, in measuring IDs in RT, it is essential that the conditions for sequential effects be the same for all subjects.

INDIVIDUAL DIFFERENCE VARIABLES DERIVED FROM CHRONOMETRIC PARADIGMS

Speed–Accuracy Trade-off

One of the prominent methodological problems in RT research concerns the relationship between speed and accuracy of response. (A thorough discussion of this problem is provided by Pachella, 1974.) In all but the simple RT paradigm, there is the possibility of errors, either in the failure to respond to the appropriate signal or in the selection of the wrong response in a choice situation. For a given task, the subject cannot maximize *speed* of response and *accuracy* of response simultaneously. Hence we speak of a *speed–accuracy trade-off*. The direction and degree of the speed–accuracy trade-off are influenced by the degree of complexity or difficulty of the task, the emphasis given to the importance of speed or accuracy in the experimenter's instructions to the subject, and individual differences among subjects. An objective index of the degree of speed–accuracy trade-off for a single subject is the point-biserial correlation between RT and response accuracy (scored 1 and 0 for correct and error responses, respectively) over trials. A *negative* correlation indicates a speed–accuracy trade-off. This index may be entered into a multiple correlation, along with other RT parameters, in studying the relationship among IDs in RT and psychometric test scores.

The speed–accuracy trade-off has always been of special concern to experimental psychologists who study RT because they are interested mainly in comparing average RTs obtained under different experimental conditions of task complexity, etc., which affect both speed and accuracy of response, and the relationship between speed and accuracy is almost always inverse when the same instructions for responding are used for all conditions. The problem lies in the interpretation of differences in RT among various experimental conditions when there are also differences in error rates. How much accuracy has been sacrificed for speed?

The speed–accuracy problem is generally less problematic to the differential psychologist than to the experimentalist. If IDs in speed and accuracy were *negatively* correlated, the differential psychologist would face the same trade-off problem as the experimental psychologist. But in fact, IDs in speed and accuracy are *positively* correlated. We have not found an exception to this generaliztion in our own work on IDs in RT or in any

studies reported in the literature. In other words, the speed–accuracy trade-off is only a *within*-subjects phenomenon, that is, speed and accuracy are *negatively* correlated *within* subjects between different task conditions. However, speed and accuracy are *positively* correlated *between* subjects within task conditions. These relationships may be easier to grasp in terms of Figure 11. On the *simple task*, persons A, B, and C are shown to have the same short RT and low error rate. On the *complex task*, the latent ability differences between persons A, B, and C are manifested as variation in their RTs and error rates. Their performances, as reflected jointly by RT and errors, will tend to fall somewhere on each of the arcs that describe the speed–accuracy trade-off; they are different for each person. If the same low error rate of the simple task is to be maintained for the complex task, the RT is greatly increased for all persons (vertical line, zero speed–accuracy trade-off). If the RT in the simple task is to be maintained in the complex task, the error rate is greatly increased for all persons (horizontal line, 100% speed–accuracy trade-off). So the arc for each person describes an *inverse* relationship (or *negative* correlation) between RT and error rate. But *between* persons, RT and error rate show a *direct* relationship (or *positive* correlation). The line marked *x* in Figure 11 indicates a fairly high

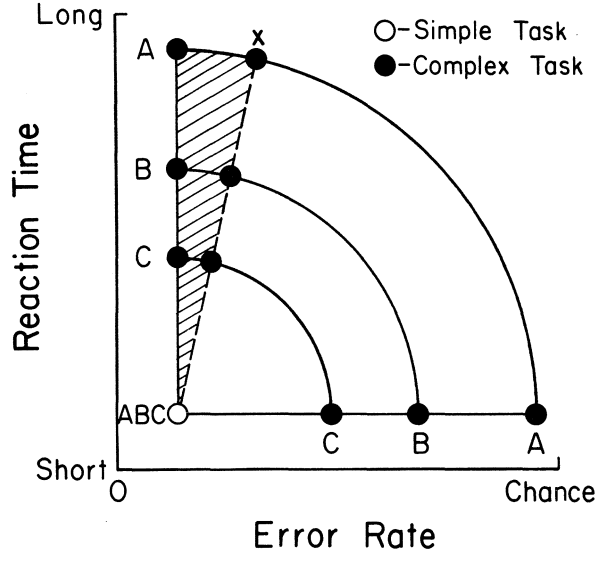

FIGURE 11. The relationship between RT and error rate for simple and complex tasks. The arcs describe the speed–accuracy trade-off for persons A, B, and C, who are shown here as performing equally well on the simple task. The shaded area represents the most desirable region of speed–accuracy trade-off for RT studies.

speed–accuracy trade-off for a typical RT study, if the error rate (on the abscissa) is assumed to range between *zero* and *chance*. Thus, the shaded area represents the most desirable region for performance when studying IDs in RT, in that it spreads out IDs in RT more than IDs in error rate.

As both RT and number (or percentage) of errors are ratio scales, Pearson's coefficient of variation ($V = \sigma/\mu$) (i.e., the ratio of standard deviation/mean) can be used to compare intersubject variability in RT and error rates. The more desirable condition for which procedures and instructions should aim is a larger V for RT than for errors. If the reverse is found, the investigator should question the procedures and instructions. A relatively high variability in errors often indicates that some subjects have not fully understood the task requirements or are too lacking in motivation or concentration to yield useful data. Subjects whose error rates are *outliers* by the some reasonable criterion (e.g., more than 3σ above the group mean) are probably better eliminated from subsequent data analyses.

It is especially important to take into account the speed–accuracy trade-off in studies in which subjects vary widely in age, because the speed–accuracy relationship interacts with age. Error rate decreases monotonically as a function of age, from early childhood to later maturity, whereas speed of response increases from childhood to early maturity and thereafter gradually decreases. Interestingly, in this respect, mentally retarded young adults resemble very old normal persons more than they resemble young children, that is, they have quite slow RTs, but relatively low error rates.

Several methods can be used to deal with errors in the treatment of RT data. Each method has advantages and disadvantages; none is ideal.

1. The central tendency (mean or median) of the subject's RT over trials can be based only on RTs for correct responses. The RTs for error responses are not used. This method is defensible only when error rates are very low (less than 4% or 5%) for every subject. With higher error rates, there is the risk that the subjects who have greatly sacrificed accuracy for speed are favorably overrated in terms of RT. A variation of this is to treat RTs for correct and error responses separately. If the correlation between RTs for correct and error responses is as high as the internal consistency reliability of either set, then there is no point in treating them separately.

2. The subject's RT is "adjusted" in terms of his or her error rate. This is accomplished by a regression equation in which RT is the dependent variable and error rate is the independent variable. The subject's "adjusted" RT score, then, is the difference between his or her *obtained* RT and *predicted* RT (using error rate as the predictor variable). Because the regression between RT and errors may be nonlinear, it is advisable to use a multiple regression equation, entering errors^1, errors^2, errors^3 (or higher

powers if necessary) as the predictor variables. The multiple prediction is justifiable only if the multiple correlation (R) between RT and the several predictor variables is significantly higher (after correction for bias or shrinkage) than the simple Pearson correlation (r) betwen RT and errors.

3. If the investigator is interested in the correlation between IDs in RT and some psychometric variable, error rate may be partialed out of the correlation. The error rate can act as a *suppressor variable* in such a correlation, that is, a variable, z, which, when partialed out of the correlation r_{xy}, results in a larger partial correlation, $r_{xy \cdot z}$.

Central Tendency of Response Time and Movement Time

Chronometric testing always involves repeated trials. In the study of IDs, we are interested in the central tendency of the subject's performance over n trials under a given set of task conditions. The number of trials (n) will depend upon the amount of testing time that is available and feasible in terms of the task demands and the degree of reliability deemed desirable for the purposes of the study.

Because there is an absolute lower limit to RT (and MT)—the so-called *physiological limit*—and RT theoretically has no upper limit, it is inevitable that the distribution of a subject's single RTs obtained in n trials will be *positively skewed*. In such a case, the *median*, rather than the *arithmetic mean*, is the preferred measure of central tendency, because the median is much less influenced by extreme values or outliers. The median has long been the usual measure of central tendency for the RT over trials of individual subjects. It should be remembered, however, that, unlike arithmetic means, medians are not additive, that is, the median value of the medians of each of two or more equal-sized groups is not equal to the median for the combined groups. For analyses in which this may be an important consideration, as in S. Sternberg's additive factor method, the arithmetic mean RT should be used instead of the median. Arithmetic mean is explicitly specified, because the *harmonic* mean (i.e., the reciprocal of the arithmetic mean of the n reciprocals of x) minimizes the effect of large outliers, and in a skewed distribution it has a value closer to the median than does the mean. But harmonic means are not additive. When additivity of RTs is an important consideration for subsequent analysis, only the arithmetic mean will do.

When the arithmetic mean is used, however, it is often advisable to apply certain uniform criteria for "cleaning up" each subject's RT data, to rid them of outliers—a practice known to statisticians as *Winsorizing* the distribution. It can greatly improve the reliability of the subject's mean RT over trials. (Winsorizing will have much less effect on the median.) Various methods can be used to Winsorize RT and MT data.

1. Eliminate all RTs of *less* than some specified value, as these are merely anticipatory flukes and not really measures of the subject's RT. True SRTs in alert young adults are rarely as short as 150 ms and certainly never shorter than 100 ms. One may safely use 100 ms as the cut-off for eliminating RTs at the lower end of the distribution. Winsorizing, or "trimming," the upper end is more problematic.

2. Eliminate all RTs (or MTs) of *greater* than some specified value. For normal subjects, we have used 999 ms as the cut-off in the one to eight light/button RT–MT paradigm; RTs or MTs that exceed 999 ms are not averaged, and the eliminated trial is repeated at the end of the scheduled trials to avoid a repetition effect.

3. Eliminate all RTs (or MTs) that exceed the subject's own *median* by some specified number of standard deviations, such as 3*SD*, with the *SD* based on the subject's own RTs over the *n* trials given to all subjects.

Intercept and Slope of Reaction Time

When the chronometric experiment consists of two or more S–R tasks of varying complexity, we usually want to characterize the subject's performance with respect to (1) an overall level or base level and (2) the amount of increase in RT as s function of task complexity. When there is an approximately linear relationship between RT and task conditions, the *intercept* and *slope* of the regression of RT on conditions efficiently describe the subject's performance. Figure 12 shows the mean RT and MT as a function of bits of information conveyed by the task conditions (one, two, four, or eight light-button alternatives, *n*) in the RT–MT paradigm (see Figure 5). The intercept and slope of the regression of RT on bits can be calculated for each subject. (Since we have never found a significant slope for MT, we now do not bother to compute its regression on bits, but obtain only the median MT for each subject.) Intercept and slope may also be calculated for the S. Sternberg memory-scan paradigm, in which RT is a linear function of the actual number of digits in the "positive set." (After being shown the "positive set," i.e., a series of from one to seven digits, the subject is shown a single "probe" digit and must respond *yes* or *no* according to whether or not it was a member of the "positive set." The RT is the interval between the probe and the subject's response.)

To determine how closely individuals conform to a linear relationship between RT and task conditions (e.g., bits in the RT–MT paradigm or set size in the S. Sternberg paradigm), one can compute the correlation (Pearson *r*) between RT and the task conditions. We have generally found the *r*s to be in the high .90s for the medians of individual subjects in the RT–MT paradigm, which clearly indicates that Hick's Law (i.e., the linear increase in RT as a function of bits) holds for individuals and is not merely

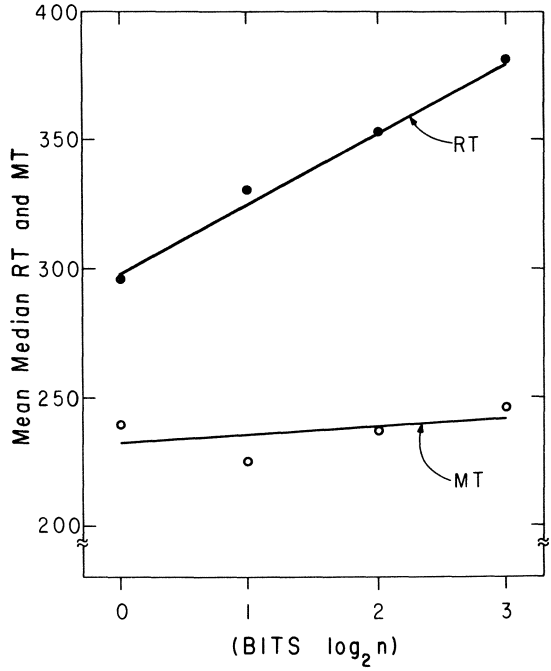

FIGURE 12. Mean median RT and MT on the RT–MT apparatus (see Figure 5) for 280 university students. Each subject's median RT (or MT) is based on 15 trials at each level of bits.

an artifact of averaging RTs over many subjects. Thus, an individual's RT in the RT–MT, or Hick, paradigm can be expressed in terms of the regression equation $RT = a + bH$, where a is the intercept, b is the slope, and H is the number of bits of information that must be processed for a correct response.

The a and b regression parameters call for distinct psychological interpretations. The *intercept* (a) is probably the most complexly determined feature of RT. It reflects not only the purely sensory and motor lags and peripheral nerve conduction, but also the apprehension and encoding of the stimulus and the preparation and initiation of the response, as well as all nonexperimental factors that may affect the subject's RT, such as the subject's general physiological state at the time. The *slope* (b), on the other hand, reflects such purely central processes as discrimination, comparison, choice, retrieval of information from short-term or long-term memory, and response selection. In terms of Hick's Law, the slope of RT on bits is the *speed* of information processing expressed as *milliseconds per bit*. The

reciprocal of the slope is the *rate* of information processing, which is conventionally multipled by 1,000 to express rate as *bits per second.*

The fact that MT is much shorter than RT, and that MT does not vary significantly or systematically with the amount of information to be processed, would seem to suggest that MT reflects only sheer speed of response after all the other functions involved in the intercept and slope have already occurred. Considerable doubt is cast on this simple interpretation of MT, however, by the fact that median MT, like the RT parameters, is correlated with IQ, which certainly involves central processesses. But RT and MT are not highly correlated with each other. *Within* subjects, the average correlation between RT and MT is *zero,* indicating that there is no trade-off between RT and MT (which would result in a negative correlation). *Between* subjects, we generally find a low correlation between RT and MT, mostly tin the range +.2 to +.4 for relatively homogeneous samples of young adults. It has also been noticed that RT (for 0 bit) is relatively greater than MT in groups with higher intelligence, as shown in Figure 13. The reason for this relationship between RT/MT and IQ remains speculative (Jensen, 1980a, p. 114; 1980b, pp. 286–289).

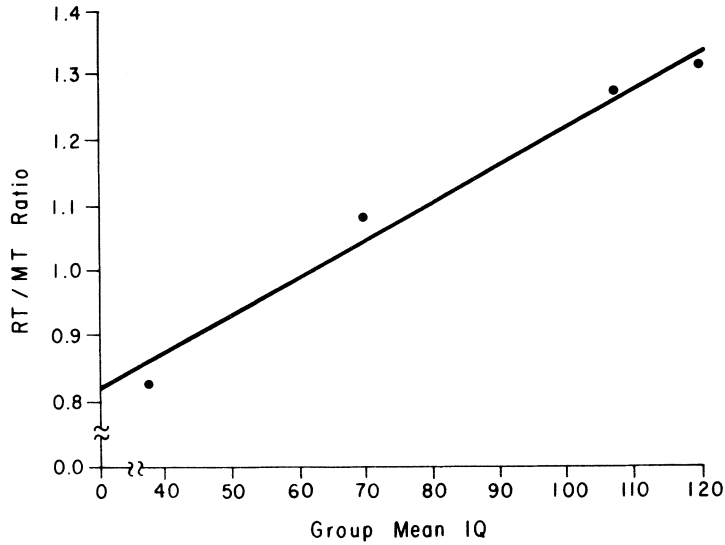

FIGURE 13. Ratio of mean of simple RT to mean MT as a function of the average intelligence levels of adult criterion groups: severely retarded (N = 60), borderline retarded (N = 46), vocational students (N = 200), and university students (N = 50).

Intraindividual Variability in Response Time and Movement Time

This has been a neglected variable in chronometric research, probably because it is less obviously a measure of "goodness" of performance than is speed of reaction *per se*; also, variability does not lend itself so neatly to such simple analytic techniques as Donders' subtraction method. Yet it now warrants our attention, mainly for three reasons: (1) There are reliable IDs in trial-to-trial intraindividual variability in RTs; (2) these IDs have been found to be at least as highly correlated with psychometric *g* as any other parameter of RT paradigms; and (3) intraindividual variability in RT seems to be a more fundamental phenomenon than RT itself, in the sense that it is theoretically easier to explain IDs in mean (or median) RT in terms of IDs in intertrial variability than the reverse. There is always a high *positive* correlation between IDs in the central tendency of RT and IDs in the intertrial variability of RT. If persons differ relatively little in the *shortest* RTs of which they are capable, but differ greatly and reliably in the *variability* of their RTs from trial to trial, they would also necessarily differ in the central tendency of their RTs and IDs in variability, and the central tendency of RTs would always be positively correlated. (Intraindividual variability is always more highly correlated with IDs in the *mean* RT than in the *median* RT over trials.) This is what we find. Hence, the causes of IDs in average RT may have to be sought in the causes of IDs in variability. It is also noteworthy that intraindividual variability in RT decreases markedly from childhood to maturity and increases again in old age.

Intraindividual variability in RT (or MT) is best measured as the standard deviation of the person's RTs over trials. It is symbolized σ_i (or when a distinction is required between RT and MT, $RT\sigma_i$ and $MT\sigma_i$). When RT is measured at a number of different levels of S–R complexity, and an overall measure of σ_i is obtained, it should be obtained *within* levels, so as not to mix up variability between mean RTs for different levels of task complexity with intertrial variability. The average of σ_i (symbolized $\bar{\sigma}_i$) over levels (or other conditions) should be obtained as follows: $\bar{\sigma}_i = \sqrt{\sum \sigma_i^2/n}$, where n is the number of conditions. (Note: Variances [σ^2] are additive, whereas standard deviations are not.)

The σ_i also increases as a function of task complexity, and for certain purposes it is useful to compute the intercept and slope of the regression of σ_i on the levels of complexity. In the Hick RT–MT paradigm, σ_i increases systematically as a function of bits of information in the stimulus array, as shown in Figure 14. Interestingly, σ_i increases in a perfectly linear fashion as a function of the actual *number* of light-button alternatives (i.e., the antilog$_2$ of bits).

The σ_i should be calculated *after* the RT data have been Winsorized by the methods previously described. This will appreciably improve the

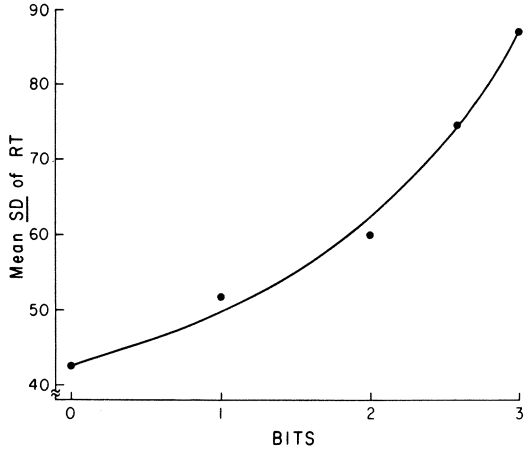

FIGURE 14. Mean intraindividual variabilty (measured by the σ_i of RTs in milliseconds on 30 trials) as a function of bits on the RT–MT apparatus, for 160 schoolchildren in grades four to six.

reliability of σ_i, which tends to have a lower reliability than the mean or the median RT.

Group differences in σ_i can be viewed more analytically by plotting RTs on each trial in their rank order of magnitude (from shortest to longest) for each subject averaging all RTs at each rank order over subjects. The RT data should first be Winsorized to minimize outlier flukes, such as by

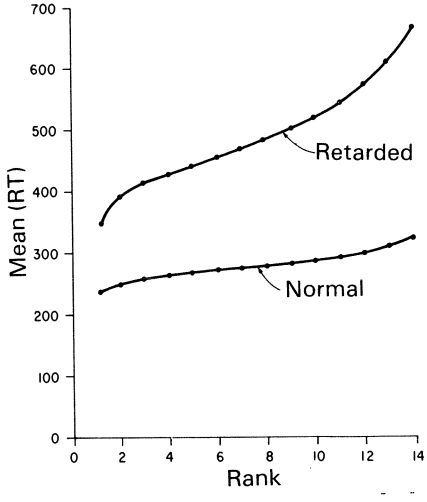

FIGURE 15. Mean simple RT (0 bit in the RT–MT paradigm) plotted after ranking each person's RTs on 15 trials from the shortest to the longest RT (omitting the 15th rank) for 46 mildly retarded and 50 normal young adults. (RT scaled in milliseconds.)

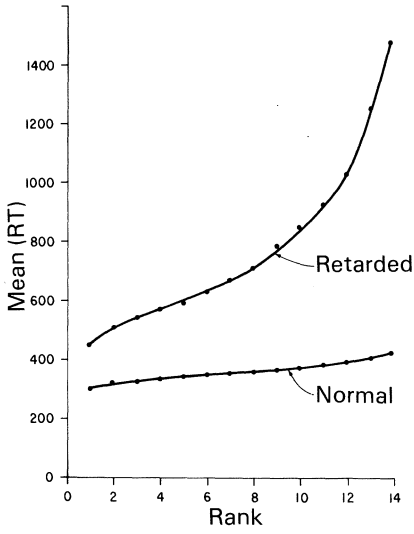

FIGURE 16. Mean choice RT (three bits in the RT–MT paradigm) plotted in the same fashion as in Figure 15.

omitting the one (or more) longest RT(s) for each subject. Figure 15 shows such a plot for normal and borderline retarded (IQs 60 to 80) young adults who were given 15 trials of SRT (one light button on the RT–MT apparatus). (The longest RT in 15 trials was eliminated for each subject.) This type of plot here reveals two theoretically important facts: (1) Retarded and normal persons differ (on average by about 100 ms) in SRT, even in their *shortest* RTs produced in 15 trials, and (2) the RTs are much more variable for retarded than for normal persons (as indicated by the marked divergence of the two curves). The group differences on SRT shown in Figure 15 are greatly exaggerated for CRT (with three bits of information), as shown in Figure 16.

The *relative variability* of RT is indicated by Pearson's coefficient of variability, $V = \sigma_i/\mu_i$. In this case, μ_i is the individual's mean RT over trials; V, like σ_i, is found to be positively correlated with mean (or median) RT and negatively correlated with intelligence level. Thus, slower (and less intelligent) subjects how greater intraindividual variability in RT in terms of both *absolute* variability and variability *relative* to their own average RT.

STATISTICAL TREATMENT OF CHRONOMETRIC DATA

Reliability and Stability

It is convenient in chronometric research to distinguish clearly between *reliability* and *stability* of the RT or MT measurements or the parameters

derived from them, such as intercept, slope, and intraindividual variability. *Reliability* refers to the "internal consistency" of the measurements across trials within a single test session. *Stability* refers to the consistency of measurements (derived from all trials) across test sessions. The stability coefficient will usually be more informative if the sessions are at least one day apart. The main reason the distinction between the coefficients of *reliability* and *stability* is important is that they differ greatly for RT (much less so for MT). The reliability of RT, when based on 15 trials or more, is usually very high—as high as the reliability of good psychometric tests, that is, above .90. The day-to-day stability of RT, however, is generally much lower than the reliability. Stability coefficients for RT mostly range from about .50 to .70, when test sessions are one or two days apart, because of the sensitivity of RT to slight changes in a subject's physiological state.

Correlations between RT (and its derivatives) and psychometric variables cannot be properly evaluated without knowing the reliability and stability of the measurements. Stability is probably the more important, because it is this nonrandom, physiological state source of variability that is most likely responsible for attenuating the correlation between RT and other variables. But the reliability coefficient is needed to evaluate the stability coefficient. Since the stability cannot be higher than the reliability, we want to be sure that a low stability coefficient is not the result of low reliability, since it is usually easier to improve the reliability (by increasing the number of trials in a session) than to improve the stability of the measurements (by increasing the number of test sessions).

For a chronometric technique that is to be used in a series of studies to measure IDs, it is advisable to determine the reliability and stability of all the derivative measurements in at least one sample that is typical of the study population.

Reliability of IDs in RT (or MT) is best measured by *coefficient alpha* (α) (Cronbach, 1951). It can be derived from a two-way ANOVA of the subjects × trials matrix. The three sources of variance are *between trials, between subjects* (*BS*), and *within subjects* (*WS*). Coefficient α, then, is derived from the mean squares (*ms*), thus:

$$\alpha = (BSms - WSms)/BSms$$

Coefficient α is the reliability of the *mean* of n trials.

Because the *median* is more popular than the mean in RT work, how can we determine the reliability of the median? There is no very satisfactory way. However, we can reason as follows. Coefficient α is the average of all possible split-half reliability coefficients (boosted by the Spearman-Brown formula). Therefore, we can determine the split-half reliability of the median by splitting the number of trials into two equal sets in various ways (e.g., odd–even trials, odd–even pairs of trials, odd–even triplets or purely

random sets), determining the median within each half, and correlating the two medians over subjects. We have done this analysis for 50 subjects given 30 trials on each of the four levels of bits (0, 1, 2, 3) on the RT–MT paradigm; we found that the S-B–boosted, split-half reliabilities for the median are as high as or higher than the same split-half determination for the mean. Therefore, it would seem reasonable to assume that coefficient α does not overestimate the reliability of the median and is probably the best estimate we can obtain, short of the wholly unfeasible prospect of calculating every possible split-half reliability coefficient. Coefficient α is undoubtedly more dependable than any *single* split-half determination.

Stability of the mean or median RT (over trials) is obtained from the Pearson correlation of these statistics between test sessions (boosted by the Spearman-Brown formula) to obtain the reliability of the composite score for two sessions. (The *composite* mean or median is the mean of the means or medians across sessions.) If there are more than two sessions, the reliability (coefficient α) can be computed from a two-way ANOVA, with the sources of variance being between sessions, between subjects, and within subjects (WS). The reliability of the composite of n sessions then is

$$\alpha = [BSms - WSms]/[BSms + (n - 1)WSms]$$

where n is the number of sessions.

All the essential reliability and stability coefficients information can be obtained from a three-way ANOVA of RT data obtained by administering the task (or a parallel form of it) for $t + 1$ *trials* on each of $d + 1$ *days* to $s + 1$ *subjects*. The full ANOVA design is shown in Table 1. The reliability coefficient α for the composite scores, derived from the mean squares (MS) in Table 1, is $\alpha = (MS_S - MS_{ST})/MS_S$. The *stability* coefficient for the composite scores is $\alpha = (MS_S - MS_{SD})/MS_S$.

TABLE 1. Analysis of Variance of RT Data for Calculating Coefficient Alpha for Reliability and Stability

Source	SS	*df*	MS
Between days (D)	SS_D	d	MS_D
Between trials (T)	SS_T	t	MS_T
D × T	SS_{DT}	dt	MS_{DT}
Between subjects (S)	SS_S	s	MS_S
S × D	SS_{SD}	sd	MS_{SD}
S × T	SS_{ST}	st	MS_{ST}
Within subjects (W)	SS_W	sdt	MS_W

The reliability of complexly determined parameters, such as the intercept and slope of the regression of RT on bits and intraindividual variability (σ_i), can only be estimated by the odd–even split-half method. These parameters are calculated separately on the odd and even trials for each subject. The Pearson r between the odd and even sets (boosted by the Spearman-Brown formula, i.e., boosted $r' = 2r/(1 + r)$) estimates the reliability of the particular parameter based on all the trials. Reliability and stability coefficients are generally much lower for these complex parameters than for the mean or the median RT.

We have discovered that when there is no significant practice effect, the RT over *trials* conforms perfectly to the basic assumption underlying the use of the Spearman-Brown formula, namely, that increasing the number of measurements by a factor of n boosts the reliability (r) such that the boosted reliability (r') is equal to $r' = nr/[1 + (n - 1)]$, *provided* that the measurements in the additional trials by which the total number of trials is increased are *equivalent* (but not necessarily identical) to the original set of measurements. Two crucial tests of the equivalence of RTs over all trials are tests of the *homogeneity* of all the *covariances* between trials and of all the *correlations* between trials. In other words, we test the null hypothesis (H_0) that all the covariances between trials are equal, and we test the same hypothesis for correlations. Statistical tests, based on chi square, for the homogeneity of covariances and correlations have been provided by Wilks (1946) and Lawley (1963), respectively. When these tests were applied to RT data from the RT–MT paradigm, they completely failed to reject the null hypothesis. (The obtained chi square was less than 1/70th as large as the chi square required to reject the null hypothesis at the .05 level of confidence.) It was concluded that neither the covariance matrix nor the correlation matrix was significantly heterogeneous, but rather appeared as if the RTs on each trial were a random sample from the total distribution of all RTs the given subject could produce during that particular testing session. There is naturally some limit to this generalization because testing cannot be prolonged to the point of fatigue without damaging the equivalence of trials. Such equivalence, or homogeneity, of RTs was not found when these statistical tests were applied to the matrix of covariances or of correlations of RT obtained in 10 sessions, each on *different* days, two days apart. The matrix of correlations between days closely resembles a *simplex*, i.e., a matrix in which the correlations systematically decrease as the number of days between test sessions increases. This simplex pattern of the correlation matrix indicates that for individuals there is some systematic change, or nonequivalence, of the RT across days, even though there is no overall significant or appreciable day-to-day variation in mean RT, mean σ_i, mean intercept, or mean slope for the group as a whole.

Factors that generally tend to decrease reliability and stability for any given number of trials are low-ability subjects (for whatever reason—age, IQ, etc.), greater task complexity, insufficient practice, variable experimental conditions across subjects, and non-Winsorized RT data.

Relationship of Chronometric Variables to Psychometric Variables

Much of our theoretical interst in chronometric variables stems from their relationship to psychometric variables, particularly general intelligence, or *g*. Verbal, numerical, spatial, and other group factors found in psychometric tests, as well as tests of scholastic achievement, are also of interest.

There are two main ways to demonstrate a relationship between a chronometric variable x and a psychometric variable y: (1) test the significance of the difference between the means on x of two or more criterion groups selected from discrete regions of the distribution of y (e.g., IQs 80–90, 100–110, 120–130) and (2) compute the correlation between x and y obtained from a sample with continuously distributed scores on y.

The first method is most economical in exploratory studies, when we are seeking those chronometric paradigms and variables that are most strongly related to psychometric variables. When significant differences are found on various chronometric variables between psychometrically distinct criterion groups, the magnitudes of the differences can be compared in terms of standard scores or mean sigma ($\bar{\sigma}$) units, where $\bar{\sigma}$ is the average within-group σ for all groups; that is, each of the group mean differences based on raw measurements is divided by $\bar{\sigma}$, so that all differences are expressed in terms of the same standard units. Given the standard deviation (σ) of raw measurements within each of n groups, the mean sigma is

$$\bar{\sigma} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + \ldots N_n\sigma_n^2}{N_1 + N_2 + \ldots N_n}}$$

where N is the number of subjects in a group. The mean difference between groups expressed in $\bar{\sigma}$ units may be corrected for attenuation (unreliability of the measurements) by dividing it by the square root of the reliability, that is, $\sqrt{r_{xx}}$.

The correlation coefficient is the most satisfactory method for expressing degree of relationship, but its interpretation and generalizability rest heavily upon a number of conditions.

1. The form of the distribution of the psychometric measurements (y) will determine the generalizability of r_{xy} to some population. If y is not

randomly sampled from a designated population, or if its frequency distribution departs significantly from the population distribution of which it is supposedly a sample, the correlation coefficient r_{xy} may be used to determine whether there is a significant relationship between x and y, but beyond the fact that r_{xy} is significantly greater than zero, its magnitude is meaningless with respect to any population; it is not generalizable. Such a correlation can be useful in exploratory research to discover those particular chronometric variables with the possibly closest relationship to the psychometric variable of interest. For such exploratory work, it is economical to test a small sample with a wide range on the psychometric variable and an approximately rectangular distribution, that is, the frequencies of each score are more or less evenly distributed over the entire range of scores. It would be as if we had a total of 60 subjects, with one subject at every IQ point over the range from, say, IQ 70 to IQ 130. The correlation of such IQ data with any other cognitive variable would, of course, be much higher than it would be in a large random sample of the population in which the distribution of IQs between 70 and 130 would closely approximate a Gaussian distribution.

2. Random or representative samples of particular natural populations (e.g., sixth-traders in a middle-class neighborhood, institutionalized retarded with IQs between 50 and 70, college students) can yield correlations that can be generalized to their respective populations, but these correlations *understimate* the true correlation in the *general* population. The reason, of course, is that almost any natural group from which we obtain our study sample has a more restricted variance than that of the general population. This is especially true for measures of intelligence, scholastic aptitude and attainments, and most other cognitive variables. If we know the standard deviation of the psychometric variable in a broader sample of the general population, such as the normative group for most standardized tests, we can use this *SD* along with the *SD* and the obtained correlation r_{xy} of our more restricted sample to obtain an estimate of what the correlation would be in the unrestricted population; the formula is given by McNemar (1949, p. 126):

$$R = \frac{r(\Sigma/\sigma)}{\sqrt{1 - r^2 + r^2(\Sigma/\sigma)^2}}$$

where R is the correlation in unrestricted sample, r is the correlation in restricted sample, Σ is the *SD* for unrestricted sample, and σ is the *SD* for restricted sample.

3. When estimates of the reliability and stability of the chronometric variables are available, the correlation can be corrected for attenuation to estimate the theoretical error-free correlation between x and y. The stability

coefficient will usually afford the more realistic correction. Simultaneous corrections for attenuation and restriction of variance are not advisable. Each correction in effect adds an increment to r_{xy}. If each of these added increments were completely independent, there would be no problem. But they are indeterminately nonindependent; the reliability coefficients used to correct for attenuation are themselves decreased by the restriction of variance. Hence, simultaneous correction for attenuation and restriction causes some indeterminate degree of overestimation of the true correlation in the unrestricted population.

4. The true degree of relationship between x and y will be underestimated by r_{xy} if the regressions are not linear. Scatter diagrams should be plotted and examined, and if there is any suspicion of nonlinearity, it should be confirmed by a suitable statistical test, such as a statistical comparision of the magnitudes of r_{xx}^2 and the squared correlation ratio, or eta^2, which is explicated in most statistics textbooks.

5. Outliers in the distributions of variable x or variable y will inflate the correlation. The distributions are best rid of outliers, or Winsorized, by some reasonable criterion before correlations are calculated. Another solution to the same problem, which has been suggested but which has little merit, is a *reciprocal transformation* of the x or y scores (or both). A reciprocal transformation of the scores (on both variables) will indeed minimize the effect of correlated outliers at the high end of the scale, but it also has disadvantages, and there is little else to recommend it. It should be noted that the correlation r_{xy} between variables x and y is not simply $-r_{xy}$, that is, the same correlation reversed in sign, by correlating $1/x$ and y, or x and $1/y$. The numerical value of r will differ, as well as its sign, and if x and y are linearly related, there will not be a linear relationship between one variable and the reciprocal of the other. For the same set of data, with linearity of the regressions of x and y and without any discontinuities or outliers on either variable, the correlation between x and y and the correlation between the reciprocals of x and y can be markedly different. Only the rank order correlation (Spearman's rho) remains invariant in magnitude under a reciprocal transformation (or any other monotonic transformation). But the rank order correlation also has the advantage of being little affected by discontinuities and outliers in the bivariate distribution and may be a useful safeguard when the Pearson r is a suspect for such reasons.

Multiple correlation, R, is called for when we want to determine the degree of linear relationship between an optimally weighted composite of chronometric variables (the independent variables) and a particular psychometric variable (the dependent variable). The independent variables need not be *experimentally* independent, that is, two or more of them may be derived from the same set of data, such as the intercept, slope, and σ_i of RT

and the mean and σ_i of MT in the Hick paradigm. The aim is simply to find the optimal set of predictors of the dependent variable. The stepwise order in which the variables come out in the multiple regression equation has virtually no theoretical significance and can be largely a matter of chance. The investigator may elect to force the order of the variables in the stepwise regression to determine if a particular variable adds a significant increment to R^2 over the the variance already accounted for by certain other variables.

If there are a number of psychometric variables, they may play the role of independent variables to predict a chronometric variable. A set of n psychometric variables often yields a greater R with a chronometric variable than is found for the converse relationship, probably because the typical psychometric variables involve more different cognitive processes than the typical chronometric variables. Complex variables are better predictors of simple variables than simple variables of complex variables. In all cases, however, the multiple R should always be corrected for bias (or "shrinkage") by the formula given in most statistics textbooks.

A *canonical correlation* expresses the degree of linear relationship between a number of independent variables and a number of dependent variables considered simultaneously. This is useful for testing a hypothesis concerning overall relationships between two sets of variables and for exploratory studies that seek those variables in each of the two domains that contribute most to the canonical correlation and, therefore, seem most promising for further experimental and correlation analysis. Unfortunately, there is no convenient correction for bias (or shrinkage) of a canonical correlation, and with a considerable number of variables and a relatively small sample of subjects, the canonical correlation will be spuriously inflated. (Dempster, 1966, has proposed a "jackknifing" method for the removal of bias from estimates of the canonical correlation.)

Age variance must be attended to in a chronometric study if the subject sample is at all heterogeneous in age. Both RT and MT are strongly affected by age in the range from early childhood to early maturity. For a sample from this age range having an age spread of more than about six months, it is advisable to control for age (in months) in all the subsequent statistical treatment of the chronometric data. The same consideration applies to age-heterogeneous samples over about the age of 30 years, beyond which age increasingly contributes to the variance in RT. With respect to the correlations between chronometric and psychometric variables, the partial correlation coefficient, with *age in months* partialed out, is called for. The regression of chronometric and psychometric variables on age is generally linear within relatively short age ranges. But if the subjects' ages range over more than about three years, one should test the correlation for nonlinearity. Usually, when there is nonlinearity, partialing out age, age^2, and age^3 will rid the correlation of all the unwanted age variance.

In a multiple correlation, R, one can enter age (or also age^2, age^3, etc.) ahead of any other variable in the stepwise regression, so that all the variance associated with age in the dependent variable is accounted for, permitting evaluation of the contributions of the remaining independent variables free of age effects.

Psychometric variables are often measured by such age-standardized tests as IQ tests, the IQs on which, at least in the standardization sample, are made to be uncorrelated with age. One now and then comes across the (mistaken) notion that if either variable x or y entering into the correlation r_{xy} is not correlated with age (a), it is unnecessary to partial out age, presumably (but mistakenly) because r_{xy} would remain unchanged by partialing out age when it has zero correlation with x (or y). Actually, in this situation, age acts as a *suppressor variable*, and partialing out age will increase the correlation, that is $r_{xy \cdot a} > r_{xy}$. If $r_{xa} = 0$ and $r_{ya} > 0$, then the partial correlation is

$$r_{xy \cdot a} = r_{xy} / \sqrt{1 - r_{ya}^2},$$

which is necessarily larger than r_{xy}.

One should not assume that scores on an age-standardized test are uncorrelated with age in any particular study sample. In sampling from regular classrooms, for example, one typically finds a low negative correlation between IQ and chronological age; that is, the younger children within any grade level tend to be brighter.

Factor Analysis of Chronometric and Psychometric Data

When a large number of variables is to be analyzed in terms of interrelationships, some type of factor analysis affords the most informative technique. The particular type of factor analysis to be used will depend, in part, upon the investigator's analytical purpose and theoretical stance. The writer has expressed his own views on these matters with reference to chronometric research in some detail elsewhere (Jensen, 1982b, pp. 263–268).

The factors that emerge from a collection of tests have greater generality than do the particular test scores, and factors are therefore of more general psychological interest. Factor analysis, in a sense, separates the psychologically more important sources of variance from the chaff of test specificity, which usually attenuates the correlations between psychometric test variables and the cognitive process variables reflected in chronometric paradigms.

From a theoretical standpoint, common factor analysis (or principal

factor or principal axes analysis) is preferable to principal components analysis, but the arguments on this issue are beyond the scope of this chapter. Factor analysis generally yields more clear-cut and more replicable results than principal components analysis, since each principal component contains some part of each test's *uniqueness* (i.e., that part of the test variance that is not shared by any other test in the battery), whereas factors reflect only common factor variance (i.e., only that variance that all tests or some subsets of tests share). Principal components, however, have the advantage that the *factor scores* (they should actually be called *component scores*) derived from them are completely determinate and exact, whereas factor scores derived from common factor analysis are mathematically indeterminate and are really *estimated factor scores,* which are imperfectly correlated with the indeterminable "true" or exact factor scores. The seriousness of this limitation of factor scores for most purposes, however, has often been exaggerated. If it is important that factor scores for different factors be perfectly uncorrelated, it is preferable that they be derived exactly from principal components and not estimated from factors. When principal components are *orthogonal* (i.e., uncorrelated), the factor scores derived from them will also be perfectly uncorrelated. The estimated factor scores derived from different perfectly orthogonal principal factors, however, may be (and usually are) correlated with one another. But in general, principal factors and principal components are, in fact, highly correlated, and rarely, if ever, would these two types of analysis result in substantially different conclusions.

As for types of factor rotation, this writer takes a definite position, which is based on the overwhelming evidence for a large general factor in the domain of cognitive abilities. R. Sternberg and Gardner (1982) have stated it well: "We interpret the preponderance of the evidence as overwhelmingly supporting the existence of some kind of general factor in human intelligence. Indeed, we are unable to find convincing evidence at all that mitigates against this view" (p. 231). The most obvious evidence for a general factor is the fact that all tests of cognitive ability, however diverse, show positive intercorrelations in any large, unrestricted samples of the general population—a fact of nature termed *positive manifold* by Thurstone (1947). This means that cognitive ability tests of all sorts have a common source of variance, which Spearman discovered in 1904 and labeled *g* for *general factor.* Therefore, any form of factor rotation that submerges the *g* factor, that is, distributes its variance among a number of rotated factors so as to obscure its identity completely, is simply an inappropriate factor model for research on mental abilities. This is precisely what is accomplished by what, at least until recent years, has been the most popular analytical method of orthogonal factor rotation, Kaiser's (1958) *varimax,* a

criterion for rotation intended to approximate Thurstone's concept of orthogonal *simple structure,* which of mathematical necessity absolutely precludes the emergence of a *g* factor, even from a correlation matrix that perfectly exemplifies positive manifold. Hence, orthogonal rotation of factors, by varimax or any other method, should be avoided in this field.

What is recommended? If the investigator is only interested in the general factor of a matrix, there are essentially two choices: (1) The first *unrotated* principal factor is a good representation of the general factor of the correlation matrix, particularly when the tests are diverse and no one type of test is overrepresented. (The first unrotated principal component [FPC] will scarcely differ from the first principal factor [FPF]. Congruence coefficients between the FPC and FPF and the correlation between FPC and FPF factor scores are generally above .95.) (2) Hierarchical factor analysis of the correlations among obliquely rotated primary, or first-order, factors will yield a single *g* factor for cognitive tests. This second-order *g* factor accounts for somewhat less of the total variance than is accounted for by the first principal factor, but it is usually very highly correlated with the FPF—an empirical generalization, not a mathematical necessity.

If the investigator is interested in other factors besides the *g* factor in his collection of variables, he should resort to a method of factor rotation that completely rids the remaining factors of any trace of *g* variance. The ideal method for achieving this is by means of a hierarchical factor analysis, using the Schmid–Leiman (Schmid & Leiman, 1957) orthogonalization transformation (cf. Wherry, 1959). This method, in effect, partials *g* (or any other higher order factor) out of the first-order factors, leaving them all perfectly orthogonal to one another and to *g* or all other higher order factors.

Factor analysis is used in chronometric research on IDs in three ways.

1. Factor analysis is used to identify the best *marker* or *reference* tests for different factors in a battery of psychometric tests properly selected to measure certain hypothesized cognitive ability factors of interest to the investigator. It is more economical to use the factor reference tests than to use the whole battery of tests that was required to identify the factors, for subsequent correlation with chronometric variables.

2. Factor analysis can be used to obtain factor scores to be correlated with chronometric variables. Factor scores are usually of more general interest, psychologically, than scores on any single test. The *specificity* of any given test may affect its correlation with a chronometric variable more than the factor the test supposedly measures and there would be no way of knowing this without the use of factor analysis. The use of properly derived *factor scores* obviates this problem.

3. Chronometric and psychometric variables can be factor analyzed

together to see which variables of both types have the largest loadings on the same factors. This is a reasonable procedure if there are a great many psychometric variables and just a few chronometric variables, because the factor structure will be predominantly determined by the psychometric variables. If about equal numbers of variables of both classes are factor analyzed together, however, there is the possibility that a clear-cut and interpretable factor structure will not be achieved, because of what is termed *method variance*, which is variance peculiar to each of the two classes of measurements. For this reason, most factor analysts recommend performing a factor analysis separately within each domain of variables—the psychometric and the chronometric. Factor scores obtained within each domain are then correlated *across* domains to reveal more clearly interpretable common sources of variance between the psychometric and the chronometric domains.

It is advisable in this type of study to minimize as much as possible the effect of a *speed factor* in the psychometric tests. Their correlation with chronometric variables should not be attributable merely to the speed of taking psychometric tests. Therefore, the psychometric tests used in chronometric studies should be administered with no time limit, or at least with a very liberal time limit. The tests should be viewed as *power tests*, and subjects should be urged to take all the time they need to attempt every item. There should be absolutely no sense of time pressure on the subjects. Research has already established, however, that the correlations between psychometric tests of *g* and chronometric variables are not attributable to a speed factor in the psychometric tests. Timed tests are no more highly correlated with chronometric variables than are untimed tests. The recommended precautions for unspeeded tests are still important, however, to rule out the overly simple interpretation of the observed relationship between psychometric and chronometric variables as being the result of a common test-taking speed factor.

It is a general rule in factor analysis that all the variables entering into the analysis should be *experimentally independent*, which means variables based on measurements obtained from separate acts or observations, not from mathematical manipulations of other variables that are entered into the same analysis. If we give subjects two different tests, x and y, the scores are experimentally independent, but the "difference score," $x - y$, and the ratio score, x/y, are not independent. In the RT–MT paradigm, RT and MT are experimentally independent measurements, whereas the *intercept* and *slope* of the regression of TR on bits are not independent variables, because they are mathematically derived from the same set of measurements. The same thing is true for mean (or median) RT and σ_t of RT. The argument against including variables that are not experimentally in-

dependent into the same factor analysis is that the correlation between such variables may simply represent an artifact of their mathematical derivation, rather than a true psychological or causal relationship. (Of course, the correlation between *any* two variables may or may not represent a causal relationship.) In short, the interpretation of factors with significant loadings on two or more experimentally dependent measures is always suspect and problematic. Yet a factor analysis or components analysis that includes experimentally dependent measures may be performed, keeping this problem in mind, for the explicitly limited purpose of seeing which variables cluster together (i.e., load on uncorrelated factors), in order to select the one variable from each cluster that best represents the cluster, as indicated by the magnitude of its factor loading. If one needs to use such experimentally dependent measures as intercept and slope in a factor analysis and wishes to give an acceptably rigorous interpretation of the results, or base a theoretically important argument on them, then the RT data should be obtained in two separate test sessions, S_1 and S_2, so that the intercept and slope parameters can be obtained in experimentally independent sets of RT data. The factor analysis can be repeated, as well, the first analysis including the correlation between the S_1 intercept and the S_2 slope and the second analysis including the correlation between the S_2 intercept and the S_1 slope.

Such chronometric variables as RT, MT, intercept, slope, and σ_i are *positively* correlated among themselves, but are all *negatively* correlated with scores on various psychometric tests of ability, which are always *positively* correlated with one another. This condition can confuse anyone examining the results of a factor analysis that comprises both chronometric and psychometric variables, even though, of course, the mixture of positive and negative correlations could have no effect on the factor structure or the magnitudes of the factor loadings. It is advisable to avoid this unnecessary difficulty in "reading" the factor matrix by reflecting the signs of some variables in the original correlation matrix so that superior performance on any variable will always show a *positive* correlation with superior performance on any other variable, thereby allowing the appearance of positive manifold when it, in fact, exists.

Factor analysis has not yet been widely or rigorously used in chronometric studies of IDs, perhaps because other methods of data analysis are more economical and, with a limited number of variables, other methods are more defensible in the initial exploratory stages of this work. Practically all the chronometric studies that have used factor analysis or have produced data that would justify factor analysis (30 data sets in all) have been quite thoroughly reviewed, and in many cases factor analyzed by a uniform method, by Carroll (1980). Carroll (pp. 81–82) has noted the five

most common deficiencies of the factor analyses applied so far to the study of IDs in elementary cognitive tasks, which include chronometric variables:

1. There was little deliberate attempt to design sets of variables that would reasonably be expected to produce clear simple structures and/or test hypotheses about factors.
2. The variables included in the factor analysis exhibited too much overlap and experimental dependence on each other.
3. The analysis used only principal component techniques (analysis of total variance), whereas a principal factor procedure (analysis of only common factor variance) would have been preferable.
4. The data were either under- or over-factorized, in that there was slavish dependence on the Guttman-Kaiser rule that the number of factors analyzed be taken as equal to the number of eigenvalues in a principal component solution that are equal to or greater than unity.
5. The factors were rotated, if at all, only orthogonally, usually by Kaiser's (1958) varimax procedure, whereas the structure of the data may have suggested that the results could be clarified by the use of oblique rotations.

CHRONOMETRIC VARIABLE CORRELATED WITH PSYCHOMETRIC INTELLIGENCE

It is a seemingly remarkable and almost counterintuitive fact that chronometric variables derived from elementary cognitive tasks that include virtually no intellectual *content* that would be a source of IDs nevertheless show significant, even substantial, correlations with scores on complex psychometric tests of general intelligence and of scholastic achievement, the item contents of which comprise a great variety of acquired knowledge and skills (Carlson & Jensen, 1982; Jensen, 1979, 1980, 1981, 1982a,b; Jensen & Munro, 1979; Jensen, Schafer, & Crinella, 1981; Vernon, 1981, 1983; Vernon & Jensen, 1984). Therefore, psychometric tests of intelligence and achievement actually tap much more fundamental sources of IDs than the superficial aspects of the information content that can be gleaned from casual inspection of the test items. Thus, IDs in mental test performance must also reflect IDs in fundamental cognitive and even neural processes that lie below the level of information content and scholastic skills *per se*. Galton's original intuition would seem to be vindicated. But much research remains to be done. The prospect of measuring IDs in human intelligence in terms of IDs in such basic and content-irrelevant processes is still a major challenge for researchers in differential psychology and mental chronometry. Research aimed toward this goal is still exploratory. The

techniques are too undeveloped and too lacking in sufficiently substantiated theoretical underpinnings and construct validity for chronometric techniques to be recommended as replacements for standard psychometric tests of intelligence. Yet, judging from the buregoning research in mental chronometry in the study of IDs, the time does not seem far off—less than a decade, perhaps—when we will see the practical application of sophisticated chronometric techniques to individual assessment, at least as a valuable adjunct to the standard psychometric instruments used in clinical work, in the diagnosis and remediation of school-learning disabilities, and in educational and personnel selection.

REFERENCES

Baumeister, A. A., & Kellas G. Reaction time and mental retardation. In N. R. Ellis (Ed.), *International review of research in mental retardation*. Vol. 3. New York: Academic Press, 1968.

Boring, E. G. *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts, 1950.

Brebner, J. M. T. Reaction time in personality theory. In A. T. Welford (Ed.), *Reaction times*. New York: Academic Press, 1980.

Carlson, J. S., & Jensen, C. M. Reaction time, movement time, and intelligence: A replication and extension. *Intelligence*, 1982, *6*, 265–274.

Carpenter, P., & Just, M. Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, 1975, *82*, 45–73.

Carroll, J. G. *Individual difference relations in psychometric and experimental cognitive tasks*. Chapel Hill, N.C.: L. L. Thurstone Psychometric Laboratory, University of North Carolina, 1980.

Cattell, R. B. *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin, 1971.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297–334.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, *12*, 671–684.

Dempster, A. P. Estimation in multivariate analysis. *Proceedings of the Symposium on Multivariate Analysis*. New York: Academic Press, 1966.

Donders, F. C. Over de snelheid van psychische processen. *Onderzoekingen gedan in het Physiologisch Laboratorium der Utrechtsche Hoogschool*, 1868, 2nd series, *2*, 92–120. (On the speed of mental processes. Trans. by W. G. Koster, *Acta Psychologica*, 1969, *30*, 412–431.)

Ehri, L. C., & Wilce, L. S. Development of word identification speed in skilled and less skilled beginning readers. *Journal of Educational Psychology*, 1983, *75*, 3–18.

Eriksen, C. W., Pollack, M. D., & Montague, W. Implicit speed: Mechanism in perceptual encoding? *Journal of Experimental Psychology*, 1970, *84*, 502–507.

Eysenck, H. J. (Ed.) *A model for intelligence*. New York: Springer-Verlag, 1982.

Flavell, J. *The developmental psychology of Jean Piaget*. New York: Van Nostrand, 1963.

Galton, F. *Memories of my life*. London: Methuen, 1908.

Grice, G. R., Nullmeyer, R., & Spikes, V. A. Human reaction time: Toward a general theory. *Journal of Experimental Psychology: General*, 1982, *111*, 135–153.

Hendrickson, A. E. The biological basis of intelligence. Part I: Theory. In H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer-Verlag, 1982.

Hendrickson, D. E. The biological basis of intelligence. Part II: Measurement. In H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer-Verlag, 1982.

Hick, W. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 1952, *4*, 11–26.

Jensen, A. R. *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology*, 1979, *2*, 16–29.

Jensen, A. R. Chronometric analysis of mental ability. *Journal of Social and Biological Structures*, 1980, *3*, 103–122.

Jensen, A. R. Reaction time and intelligence. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning*. New York: Plenum Press, 1981.

Jensen, A. R. Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer-Verlag, 1982a.

Jensen, A. R. The chronometry of intelligence. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*. Vol. 1. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1982b.

Jensen, A. R., & Munro, E. Reaction time, movement time, and intelligence. *Intelligence*, 1979, *3*, 121–126.

Jensen, A. R., Schafer, E. W. P., & Crinella, F. M. Reaction time, evoked brain potentials, and psychometric *g* in the severely retarded. *Intelligence*, 1981, *5*, 179–197.

Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, *23*, 187–200.

Kantowitz, B. H. Double stimulation. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1974.

Keating, D. P., & Bobbitt, B. Individual and developmental differences in cognitive processing components of mental ability. *Child Development*, 1978, *49*, 155–169.

Keele, S. W. *Attention and human performance*. Pacific Palisades, Cal.: Goodyear, 1973.

Kirby, N. Sequential effects in choice reaction time. In A. T. Welford (Ed.), *Reaction times*. New York: Academic Press, 1980.

Lawley, D. N. On testing a set of correlation coefficients for equality. *Annals of Mathematical Statistics*, 1963, *34*, 149–151.

Libnet, B. Cortical activation in conscious and unconscious experience. *Perspectives in Biology and Medicine*, 1965, *9*, 77–86.

McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.

Merkel, J. Die zeitlichen Verhältnisse der Willensthatigkeit. *Philosophische Studien*, 1885, *2*, 73–127.

Nettelbeck, T. Factors affecting reaction time: Mental retardation, brain damage, and other psychopathologies. In A. T. Welford (Ed.), *Reaction times*. New York: Academic Press, 1980.

Pachella, R. G. The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1974.

Pöppel, E. Time perception. In R. Held, H. W. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of sensory physiology: Vol. VIII. Perception*. Berlin: Springer-Verlag, 1978.

Posner, M. I. *Chronometric explorations of mind*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978.

Posner, M. I., Lewis, J. L., & Conrad, C. Component processes in reading: A performance analysis. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Schmid, J., & Leiman, J. The development of hierarchical factor solutions. *Psychometrika*, 1957, *22*, 53–61.

Smith, E. E. Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin*, 1968, *69*, 77–110.

Spring, C. Perceptual speed in poor readers. *Journal of Educational Psychology*, 1971, *62*, 492–500.

Spring, C., & Capps, C. Encoding speed, rehearsal, and probed recall of dyslexic boys. *Journal of Educational Psychology*, 1974, *66*, 780–786.

Sternberg, R. J. *Intelligence, information processing, and anological reasoning: The componential analysis of human abilities*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Sternberg, R. J. (Ed.) *Advances in the psychology of human intelligence* (Vol. 1). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1982.

Sternberg, R. J., & Gardner, M. K. A componential interpretation of the general factor in human intelligence. In H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer-Verlag, 1982.

Sternberg, R. J., & Rifkin, B. The development of anological reasoning processes. *Journal of Experimental Child Psychology*, 1979, *27*, 195–232.

Sternberg, S. The discovery of processing stages: Extensions of Donders' method. (Attention and performance II.) *Acta Psychologica*, 1969, *30*, 276–315.

Thurstone, L. L. *Multiple-factor analysis: A development and expansion of the vectors of mind*. Chicago: University of Chicago Press, 1947.

Vernon, P. A. Reaction time and intelligence in the mentally retarded. *Intelligence*, 1981, *5*, 345–355.

Vernon, P. A. Speed of information processing and general intelligence. *Intelligence*, 1983, *7*, 53–70.

Vernon, P. A., & Jensen, A. R. Individual and group differences in intelligence and speed of information processing. *Journal of Educational Psychology*, 1984, *5*, 411–423.

Warren, H. C. (Ed.) *Dictionary of psychology*. Boston: Houghton Mifflin, 1934.

Welford, A. T. (Ed.) *Reaction times*. New York: Academic Press, 1980.

Wherry, R. J. Hierarchical factor solution without rotations. *Psychometrika*, 1959, *24*, 45–51.

Wilks, S. S. Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution. *Annals of Mathematical Statistics*, 1946, *17*, 275–281.

Woodworth, R. S. *Experimental psychology*. New York: Holt, 1938.

Woodworth, R. S., & Schlosberg, H. *Experimental psychology*. New York: Holt, Rinehart & Winston, 1954.