Chapter 11
The psychometrics of intelligence

A. R. Jensen

1. INTRODUCTION

1.1 The specificity doctrine

My research interest in human mental abilities grew out of the viewpoint espoused by Eysenck that I referred to in my introduction as the Galton–Spearman–Eysenck school. In several of his publications, Eysenck has contrasted this school of thought about the nature and measurement of mental ability with the other major approach stemming from the work of Alfred Binet (1857–1911).

At the behest of the Paris school system, Binet (in collaboration with Théophile Simon) invented what was probably the first practical test of intelligence. The Binet–Simon test was also the first mental test to be scaled in mental-age units. It soon proved highly useful for the diagnosis of mental retardation among children for whom schooling presented unusual difficulty. This was a major achievement and is unquestionably a landmark in the history of mental measurement. Binet was a distinguished experimental psychologist, Simon a physician, but neither one was a psychometrician in the technical sense of that term as we understand it today. In fact, the branch of psychology known as psychometrics had not even come into existence when Binet's test was published in 1905. Because the basic principles of psychometrics still awaited Spearman's formulation in terms of what is now known as classical test theory, Binet and his immediate followers could not possibly have conceived or described the "intelligence" measured by Binet's test except in terms of the test's superficial features, such as its specific item content (vocabulary, counting, form board, paper folding, esthetic judgment, arithmetic operations, matching forms, etc.), and the inferred mental faculties that the test items supposedly called upon (memory, discrimination, comparison, reasoning, judgment, etc.). As Binet's test was considered a measure of intelligence, and its overall scores (or the derived mental-age) accorded quite well with teachers' subjective estimates of their pupils' intelligence as judged from their classroom behavior and scholastic performance, psychologists naturally defined intelligence as the sum of all the kinds of knowledge and skills visibly recognized in

the items of the Binet test (or of other tests modeled after it). At a somewhat higher level of abstraction, intelligence was described in terms of all the various faculties that were surmised to be called for by the particular cognitive demands made by each of the different types of test items.

I have termed this prevailing conception of what is measured by intelligence tests the "specificity doctrine." From a strictly psychometric standpoint, it is easy to prove that the specificity doctrine is absolutely wrong, despite the fact that it is the prevailing notion among many clinical and applied psychologists—the very psychologists who use tests the most.

1.2 The fallacy of the specificity doctrine

Spearman was the first trenchant critic of the Binet approach and of the wrongly conceived specificity doctrine, arguing that its conception of intelligence and its method for measuring it were based on an arbitrarily chosen hotchpotch of items that include various kinds of knowledge and cognitive skills. There is nothing magical about the particular collection of the items that compose the Binet test. Any other equally diverse collection of items that tap many different bits of knowledge and skills would do just as well.

But why should this be so? Given the minimum requirement that the items range widely enough in their level of difficulty to produce reliable individual differences in the total score, why is the specific item content of the Binet test (or any other test) of so little importance? Spearman's complaint with Binet's signal contribution was not so much with the test itself as with the misconception about what it measures, a misconception, incidentally, that has endured among many psychologists up to the present day. Spearman's key discovery of what the Binet test (and all other tests of its type) essentially measures was completely unrealized by Binet. It is my impression that it is no better understood by many present-day psychologists, not to mention the popular media and the general public.

1.3 The signal and the noise

To make a long story short, Spearman was the first person to recognize that the total score on a mental ability test does not measure the ability to do this item or that, such as being able to define a particular word, know a particular fact, comprehend a given sentence, recall a string of digits, solve a particular puzzle, or copy a particular form. The observed individual differences in overall test scores that we represent as the *true-score variance* in some defined population (or a sample thereof) does not reflect individual differences in the specific abilities needed to perform correctly on this, that, or other items of a test. We can easily prove this by looking at the correlation between persons' scoring "right" or "wrong" on any given item and that item's correlation with any other

item in the test. In the best IQ tests, the average interitem correlation is typically between +.10 and +.15, and the correlation between any single item and the total score on a test composed of many diverse items is typically between +.20 and +.30. In other words, whatever it is that is measured by the test as a whole is measured only very slightly by each of the test items. The total variance of each item contains only a faint "signal," which reflects a source of variance (i.e., a factor) that is common to each and every one of the disparate items that make up the test. The vast bulk of the variance on any given item, however, is just "noise." What we can see with the naked eye by examining various test items is only this noise, that is, the specific bits of knowledge or skill called upon by a particular item. For any given item, the signal to noise ratio is typically about 1 to 99. But by aggregating the scores on each of a large number of diverse items to obtain a total score, the signal to noise ratio for the total score is typically greater than 9 to 1. The reliability of a test, that is, the percentage of its true-score variance, is about 90% of the total raw-score variance for most professionally developed standardized mental tests. The remaining 10% of the variance, or the noise, is the sum of the item variances. By examining each of the items to determine what the test measures, all that one really sees is the noise (or error component) of the total test scores. In other words, what we measure as individual differences, or variance, in total scores mostly consists not of *item variances*, but of twice the sum of all the *covariances* among items. In the total variance of test scores, these two sources of variance are generally in the ratio of 1 to 9. Unlike test items, covariances are not "things," or bits of knowledge, or specific skills. One can see the item content (the noise) by inspecting the items. But one cannot see the item covariances (the signal), which can only be determined by analytical calculations.

The item covariances originate from the aggregation of the small invisible signal reflected by each item. (The sum of the item variances, or noise, constitutes the test's error variance.) The test's true-score variance is solely attributable to the fact that the diverse items are all positively correlated with one another, however slightly, thereby producing a total of $n(n-1)$ covariance terms (where n is the number of items in the test). The sum of all the item covariances constitutes the test's true-score variance. But the important point I wish to make here is that the main source of the variance of interest when we use a test does not reside in the test items per se, but reflects whatever it is that causes items to be positively correlated with one another. The positive correlation between all cognitive test items is a given, an inexorable fact of nature. The all-positive interitem correlation matrix is not an artifact of test construction or item selection, as some test critics mistakenly believe. In fact, it is empirically impossible to devise a cognitive test that has nonzero variance and in which there are statistically significant negative interitem correlations. An imaginary test for which the average interitem correlation is either negative or zero is the psychometric equivalent of a perpetual motion machine.

1.4 Distilling psychometric g

The variance attributable to the sum of the factors derived from a factor analysis constitutes the *common factor variance*, of which the largest component is usually the *g* factor. (Other common factors typically found in mental ability tests are verbal, spatial, numerical, and memory.) The *g* factor, emerging as it does from correlations among elements rather than from the addition of elements, is not a compound or conglomerate, but rather is more aptly likened to a distillate.

Although the *g* factor is typically the largest component of the common factor variance, it is the most "invisible." It is the only "factor of the mind" that cannot possibly be described in terms of any particular kind of knowledge or skill, or any other characteristics of psychometric tests. The fact that psychometric *g* is highly heritable and has many physical and brain correlates means that it is not a property of the tests per se. Rather, *g* is a property of the brain that is reflected in observed individual differences in the many types of behavior commonly referred to as "cognitive ability" or "intelligence." Research on the explanation of *g*, therefore, must necessarily extend beyond psychology and psychometrics. It is essentially a problem for brain neurophysiology.

1.5 The Galton–Spearman–Eysenck school

In order to focus theory and research in this domain, Eysenck, probably more than anyone else, has emphasized the importance of D. O. Hebb's classic distinction between three different meanings of "intelligence," which are labeled Intelligence A, Intelligence B, and Intelligence C. Failure to observe these crucial distinctions only obscures theoretical discussion and creates spurious arguments. Intelligence A refers to the biological (i.e. genetic and neurophysiological) basis of observed individual differences in Intelligence B and Intelligence C. Intelligence B is any form of gross observable behavior that involves cognitive abilities as they are manifested in "real life" circumstances—in learning, problem solving, memory, general knowledge, verbal facility, quantitative reasoning, levels of mastery of vocational skills, educational and occupational level, income, social adeptness, intellectual interests and achievements, and so on. Intelligence C is the cognitive ability measured by psychometric tests, such as IQ.

These three conceptions of intelligence are of course not independent, and all are worthy of study in their own right. But Intelligences A and C are far more amenable to exact scientific study than Intelligence B. The class of variables that fall under Intelligence B is so unbounded, and the causes of individual variation at this broad-scope level of observation involve such a multiplicity of social, cultural, experiential, motivational, and specific contextual factors as to frustrate attempts to arrive at a scientifically satisfying

theoretical formulation. About all we can do with Intelligence B is to determine the degree of correlation some of its quantifiable aspects have with both Intelligences A and C, and possibly, by using the techniques of path analysis, we can test competing hypotheses concerning the causal relationships between Intelligences B and C.

Intelligence C, such as IQ, is itself problematic in that it reflects to some extent the particular item composition of the test, more or less, depending on the number and diversity of the items, which is technically referred to as the adequacy of psychometric sampling. Differences in psychometric sampling account for why various IQ tests are not perfectly correlated, although the correlations even among standard IQ tests that differ markedly in item composition is quite high, averaging about +.80 (which goes up to about +.90 after correction for attenuation [unreliability]). Because all cognitive tests, without exception, are positively correlated with one another and therefore, when factor analyzed together, yield a general factor, *g*, it makes sense to use *g* as the criterion of the adequacy of any given test as a measure of general mental ability. Thus a measure of Intelligence C, to the extent that it is correlated with *g*, is the best psychometric marker available for discovering which biological variables are involved in Intelligence A. In fact, we now know empirically that the larger the test's *g* loading (Intelligence C), the larger its correlation with biological variables (Intelligence A). In other words, the process of extracting the *g* factor from a large and diverse battery of tests screens out, so to speak, much of the variance in conventional intelligence test scores that is unrelated to Intelligence A. The *g* factor is probably closer to the biological substrate, Intelligence A, than any other measure we can derive from conventional psychometric tests. A figure often used by Eysenck represents the variance as the area of a circle and shows the causal influences (arrows) of different kinds of variables on Intelligences A, C, and B, which in Figure 11.1 are labeled as biological, psychometric, and social intelligence, respectively.

The Galton–Spearman–Eysenck school focuses primarily on Intelligences A and C. Its research program in this field is aimed first at discovering the relationship of psychometric *g* to biological variables, and second at ultimately explaining the causal nature of the relationship in terms of neurophysiological mechanisms. A number of possible approaches are available: reaction time (RT) and inspection time (IT) measures of the speed of information processing in elementary cognitive tasks, and physiological measures such as the average evoked potential (AEP), neural conduction velocity, the brain's glucose metabolic rate (GMR) during problem solving, and brain size estimated from head measurements or measured directly by magnetic resonance imaging (MRI). Each of these is worth pursuing, because the different kinds of data they provide, when viewed in conjunction, may serve as triangulation points for hypothesizing the nature of the information processes and their biological mechanisms that are reflected by the *g* factor derived from

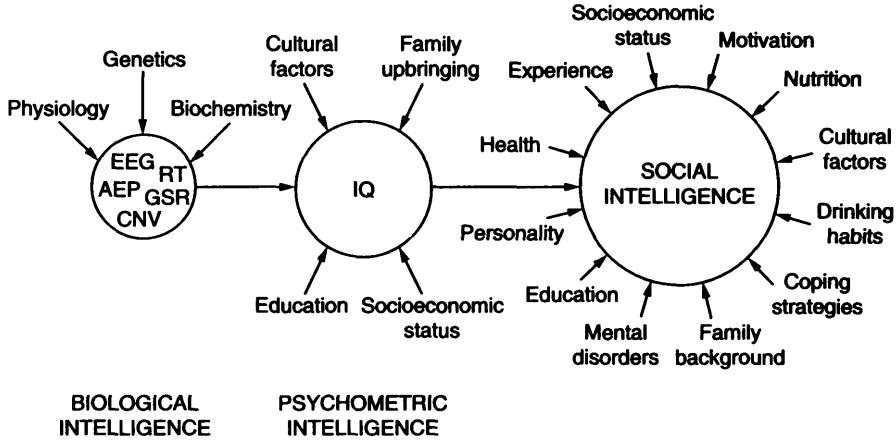

Figure 11.1. Eysenck's representation of three different conceptions of "intelligence." Biological intelligence is measured by the electroencephalogram (EEG), the average evoked potential (AEP), nerve conduction velocity (CNV), the galvanic skin response (GSR), and reaction time (RT).

conventional psychometric tests. For the most recent synthesis of Eysenck's whole philosophy about the nature and aims of research on intelligence, readers are referred to Eysenck (1994).

2. REACTION TIME AND PSYCHOMETRIC *g*

2.1 Introduction

Until 1964, I had never given any thought to reaction time (RT) in relation to intelligence. I had learned at some time in the past, probably as an undergraduate psychology student, that way back in "ancient history" Galton had hypothesized a correlation between RT and mental ability of the type previously described as Intelligence B (since there were no IQ tests at that time). I also knew that it was common knowledge among psychologists that Galton's idea on this point had long since been completely discredited empirically and that even to suggest that it might possibly still have some merit was a sign of ignorance or naïveté.

In 1964–1965 I spent the academic year of my first sabbatical leave from Berkeley in Eysenck's department at the Institute of Psychiatry, where seven years earlier I had spent two years on a postdoctoral fellowship from the National Institute of Mental Health. I recall that one day in 1964, during the mid-morning coffee break in the old Maudsley cafeteria, Eysenck joined several of us at our table and told us in considerable detail and with evident enthusiasm about an article he had recently read in a German psychology journal (Roth, 1964). It was a study of RT in which the subject, seated in front

of a display panel, was instructed to turn off a light as quickly as possible after its appearance on the panel by touching a button adjacent to the light. The light that went on (and that was to be turned off by the subject) was presented on the panel under each of several different conditions, referred to as set-size: the light appeared either alone (set-size 1), or appeared among sets of 2, 4, or 8 lights. In each set-size, any one of the alternative lights would go "on" at random. Each subject was given a large number of trials under each set-size and the subject's RT on each trial was measured in milliseconds (ms). (RT is the interval between the light's going "on" and the subject's touching the button that turns it "off.") When the RTs for each set-size were averaged over trials and over subjects, the phenomenon known as Hick's law was clearly evident. If n is set-size (i.e.,the number of alternative possibilities for which of the n lights would go "on"), Hick's law states that the increment in RT for each unit of increase in set-size is equal to the binary logarithm of the set-size, or $\Delta RT = k + \log_2 n$ (where k is the RT for $n = 1$). In information theory, the unit of information, termed a bit (for binary digit), is measured as the binary logarithm (log2) of the number of alternatives; a bit is the amount of information needed to reduce uncertainty by one-half. Thus, Hick's law expresses the linear relationship between amount of information and RT.

But what Eysenck found to be exciting about Roth's study was not that it confirmed Hick's law (about which there was little doubt, in any case), but that it showed rather marked individual differences in the slope of the Hick function (i.e., the linear regression of RT on bits) and, most important, these individual differences in slope had a significant negative correlation with IQ. Galton's original conjecture, via Hick's law, appears to have been substantiated. It made perfect sense theoretically that, if intelligence were conceived of as the speed or efficiency of information processing, the rate of increase in RT as a function of bits should be greater for persons of low IQ than for persons of higher IQ. It was also of interest to see that here was an exceedingly simple task that bore absolutely no resemblance to any IQ test and yet the RT parameters (median RT and the slope of RT/bits) derived from it were correlated with IQ.

Right then and there, for some reason that I cannot fathom, I felt that this was perhaps the single most interesting finding I had come across in my total acquaintance with psychology up to that time. But I did nothing about it. In retrospect, I think what I should have done was to drop everything else and immediately go to work on this Hick paradigm and its relation to IQ. But when I returned to Berkeley I was committed to completing other work in progress and I had research grants for experiments on the psychology of human learning. Then I was invited to spend a year at the Center for Advanced Study in the Behavioral Sciences, where laboratory work was not possible. And so it went, year after year. Although I went on thinking about the Roth experiment, of and on, year after year, I found no way to work it into my agenda.

Then, in 1967, Eysenck sent me a reprint (Eysenck, 1967) of what, for me, was one of his most important articles. In it, he put forth a theory of intelligence that was essentially biological and Galtonian, hypothesizing the speed of information processing as the basis of individual differences in performance on cognitive tasks. He also explained how Roth's 1964 study of the correlation between RT and IQ in the Hick paradigm jibed with Galton's ideas and brought research on individual variation in mental ability closer to its biological basis than was possible with conventional psychometric tests. My long-standing latent interest in this approach, then having been energized by Eysenck's stimulating article, I applied for a grant to construct the kind of RT apparatus needed to pursue this line of investigation.

2.2 The Hick paradigm

Figure 11.2 shows the subject's response console for the present model of my apparatus (in the first model, the lights and their closely adjacent pushbuttons were separate; in the present model, the pushbuttons themselves would light up, thereby maximizing the stimulus–response compatibility of the RT task).

The whole sequence of practice trials and stimulus presentations is run by computer, which also records response times (in ms) and errors on each trial. The one feature that differs importantly from the procedure used by Roth is the use of a "home" button, which makes it possible to separate RT from movement time (MT). RT is the time interval between the onset of the reaction stimulus (i.e., one of the buttons lighting up) and the subject's releasing the home button. MT is the interval between the subject's releasing the home button and pressing the lighted button to turn it off. Typically 15–30 trials are given for each set-size. From these data, five chronometric scores are derived for each subject: (1) median RT, (2) median MT, (3) the standard deviation of RT over trials (RTSD), (4) the standard deviation of MT over trials (MTSD), and (5) the slope of RT across set-sizes expressed as bits. Figure 11.3 shows the typical results for a group of subjects.

The characteristic features seen in every study (except those based on severely retarded subjects with IQs below 50) are (1) the significant linear slope of RT as a function of bits, (2) the near-zero slope of MT as a function of bits, and (3) that RT is much greater than MT even at 0 bits.

My program of research on the relationship between RT and psychometric *g* has extended over a period of more than two decades, and it would be impossible here even to summarize all of the findings within the allotted limits of this chapter. Most of these, however, are summarized in considerable detail elsewhere, along with references to most of the original studies (Jensen, 1982, 1987a, 1987b, 1993a). Here I will simply note some of the main findings, explain some of the main controversies that have since been more or less resolved, and note some of the open questions that await further research.

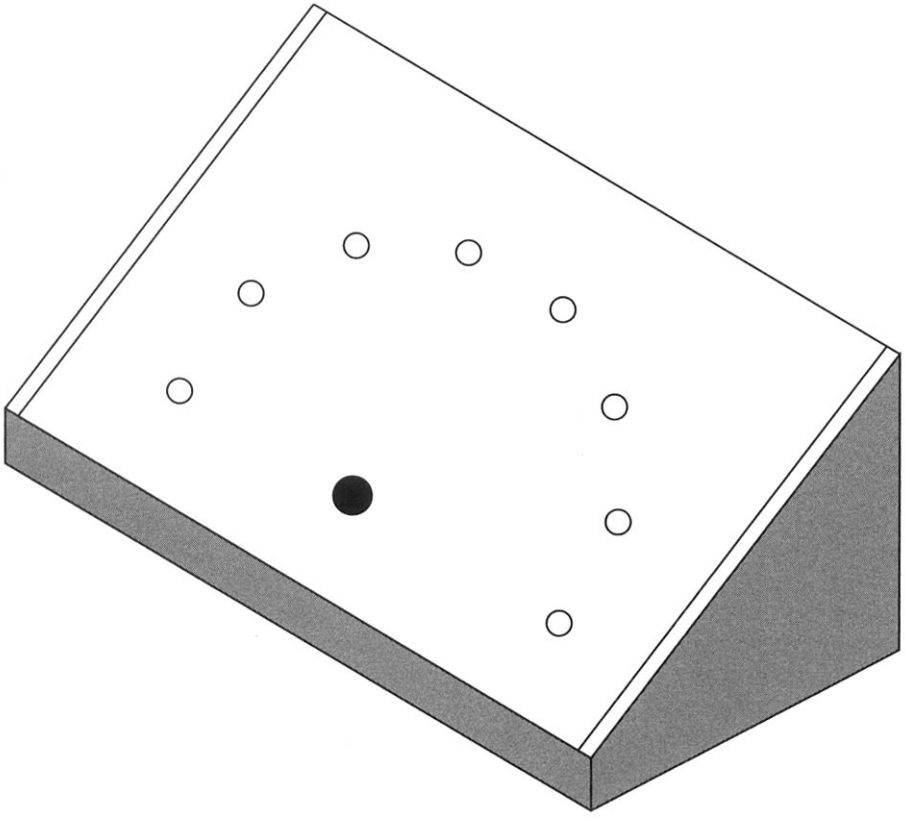

Figure 11.2. The subject's response console of Jensen's RT–MT apparatus. The panel is 13 in. × 17 in., painted flat black, and tilted at a 30 angle. At the lower center is the home button (black, 1 in. diameter), which the subject depresses with the index finger while waiting for the reaction stimulus. The semicircle of eight small circles represents translucent pushbuttons (green, 0.5 in. diameter, each at a distance of 6 in. from the home button); each button can be lit independently. Touching a lighted button turns off the light. Various plates can be placed over the console to cover some of the buttons, leaving either 1, 2, 4, or all 8 buttons exposed to view, making for four different ECTs, each with a different number of equally likely response alternatives. The binary logarithms of 1, 2, 4, and 8 exposed buttons are equivalent to 0, 1, 2, and 3 bits of information, respectively. A trial begins with the subject depressing the home button; 1 s later a preparatory stimulus ("beep") of 1 s duration occurs; then, after a 1–4 s random interval, one of the buttons lights up, whereupon the subject's index finger leaves the home button and touches the underlighted button. RT is the interval between a light-button going "on" and the subject's lifting the index finger from the home button; MT is the interval between releasing the home button and touching the underlighted button. In each trial only one of the buttons lights up, entirely at random from trial-to-trial.

2.3 Main findings

The basic generalization that follows from the results of this research is that Galton's hypothesis that speed of reaction and general intelligence are intimately related is amply substantiated. Innumerable studies, from my own

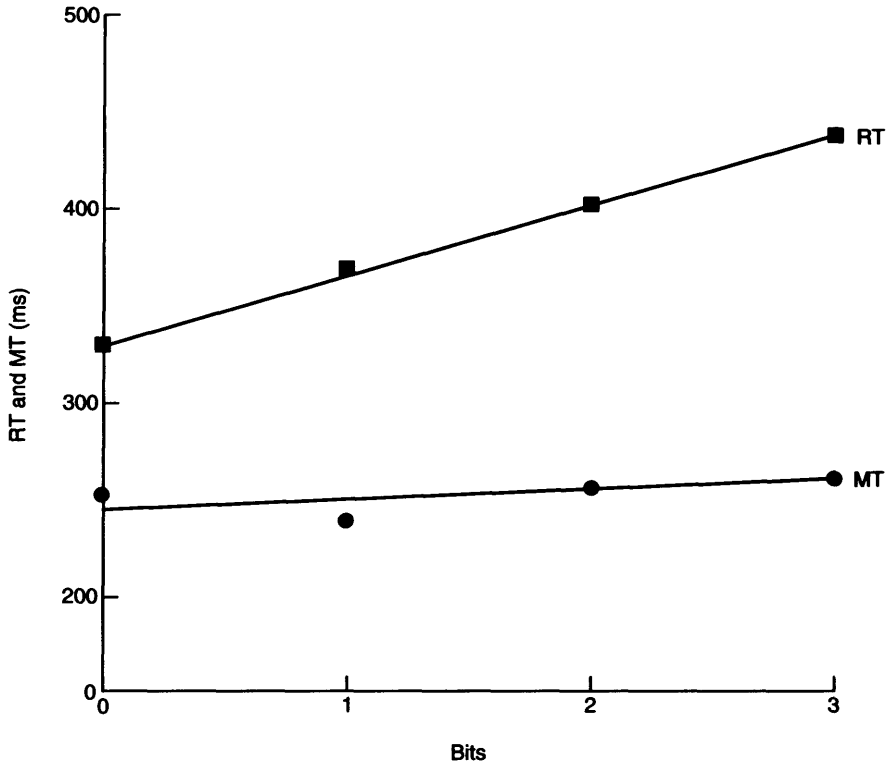

Figure 11.3. The median RT and MT obtained on the RT–MT apparatus (Figure 11.2) averaged over more than 1500 individuals. Note the significant positive slope of RT (RT = 336 + 34 BIT, r = .998) demonstrating Hick's law, which predicts a linear relationship of RT to the amount of information measured in BITs. In marked contrast is the nonsignificant slope of MT (MT = 245 + 4.3 BIT, r = .641) (data from Jensen, 1987, Tables 3 and 7).

laboratory and from many others around the world, have shown correlations, mostly in the range of .30 to .50, between response times and scores on untimed *g*-loaded tests, such as IQ. There is no longer the least doubt about the fact of correlation between RT and *g*. It shows up not only in the Hick paradigm (described above), but in every elementary cognitive task (ECT) that uses RT as the measure of proficiency. The correlations are largest for ECTs in which the RT is less than 1 s for adults or less than 2 s for elementary school children. Within this narrow range, the degree of correlation between RT and IQ is related to the complexity of the cognitive demands made by the ECT used to elicit RT. If the task is too complex and results in longer RTs, other factors besides speed of processing intervene to attenuate the correlation between RT and *g*.

Moreover, we have found that RT is more related to *g* than to any other factor independent of *g* that can be extracted from a battery of psychometric tests. RT correlates with various psychometric tests to the degree that the tests are *g* loaded.

RT in a variety of ECTs shows very large and highly significant differences between criterion groups that differ in mental ability not only as measured by psychometric tests, but as recognized by common sense, such as:

1. Between the institutionalized retarded or persons in sheltered workshops and the general run of "normal" persons.
2. Between precocious children who are succeeding in college at 12–14 years of age as compared with their age-mates who are in the usual school grades for their age.
3. Between students in a selective university and students of the same age in a much less selective vocational college.
4. Between young adults (ages 18–22) and much older adults (ages 65–80) of similar socioeconomic background and level of education (the older adults have slower RT, especially on the ECTs that are most highly correlated with *g*).

RT and MT do not measure the same thing. They are correlated only about .30, and MT is almost completely unrelated to experimentally manipulated variations in the complexity of any particular ECT, whereas the manipulation of complexity has marked effects on RT as well as its degree of correlation with *g*. Moreover, in a factor analysis of RT and MT measures from a large number of quite different ECTs, RT and MT very distinctly come out on different factors, and MT has no significant loadings on *g* or on any other factors with salient loadings on psychometric tests.

Intraindividual variability in RT, measured as the standard deviation of the individual's RTs over a specified number of trials (hence labeled RTSD), has lower reliability than the individual's mean or median RT, and yet RTSD is generally more highly correlated (negatively) with *g*. Higher IQ persons maintain more consistent RTs from trial to trial than persons of lower IQ. This shows up in the degree of skew in the distribution of RTs for a given individual. In a large number of trials, low IQ and high IQ persons differ relatively little in their fastest RTs, but they differ markedly in their slowest RTs—low IQ persons produce many more slow RTs, which makes their RT distribution more highly skewed. Median RT and RTSD are highly correlated, even when based on independent sets of RT data, but each measure is to some degree independently correlated with *g*, RTSD slightly more so than median RT. Eysenck, I believe, was the first to suggest that RTSD reflects neural noise in the brain, which impedes the transmission of information, which is also re-

flected in longer RT. However, the independent correlation of median RT and RTSD with IQ indicates that RT is not entirely derivative from RTSD, but that both aspects of RT—speed and consistency—are independently related to IQ.

2.4 Main controversies and possible solutions

In our studies we have almost always used nonspeeded or untimed tests (usually Raven's Matrices) to measure *g*, instructing subjects to attempt every item and to take as much time as they needed. Subjects are usually tested alone in a quiet room, so there is no chance of their being paced by observing other individuals who may complete the test sooner. This is important in order to rule out speediness of test taking as the common factor that might account for the correlation between RT and *g*. We have established beyond any possible doubt that the RT–*g* correlation is not a result of the speed factor often found in psychometric tests. The amount of time subjects take to complete a difficult cognitive test and their RTs on an ECT have a near-zero correlation. Moreover, RT is more highly correlated with scores on untimed tests than with scores on highly speeded tests. At one time, a number of psychologists thought we were merely rediscovering with our ECTs and measures of RT the clerical speed factor that had long since been identified by factor analyses of test batteries that included highly speeded tests. But the very tests that best measure the psychometric speed factor have the lowest correlations with RT and also with *g*. In brief, RT and RTSD certainly do not measure speediness of test taking.

How, then, can we explain why RT on ECTs that are so simple that all subjects can perform them with very few or no errors are substantially correlated with performance on untimed tests that involve complex reasoning and that measure individual differences in terms of the number of correct answers, which reflects the level of item complexity and difficulty at which the subject can successfully perform?

At least one part of the answer involves what cognitive theorists refer to as the capacity of working memory, that is, the amount of information that can be retained and manipulated in conscious awareness before any information is lost through interference from new input or decay of memory traces. The importance of processing speed in the operation of working memory stems directly from the capacity limitation and the rapid decay of information in short-term memory (STM). The limited capacity of working memory severely restricts the number of operations that can be performed at any one time on the information that enters the system from external stimuli or from retrieval of information stored in primary memory or in long-term memory (LTM). Quickness of mental operations is advantageous because more operations per unit of time can be executed without overloading the system. Also, because there is rapid decay of stimulus traces in the sensory buffers and of information

in STM, there is an advantage to speediness of any operations that must be executed on the information while it is still available. To compensate for limited capacity and rapid decay of incoming information, the individual resorts to rehearsal and storage of information into LTM, which has a relatively unlimited capacity. But the process of storing information in LTM itself uses up channel space, so there is a "trade-off" between the storage and the processing of incoming information. The more complex the information and the more operations that are required on it, the more time that is necessary, and consequently the greater is the advantage of speediness in all the elementary processes involved. Loss of information due to overload interference and decay of information that was inadequately encoded or rehearsed for storage and retrieval from LTM results in a failure to grasp all of the essential relationships among the elements of a complex problem needed for its solution. Speediness of information 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 capacity. The extreme simplicity of the ECTs, for which RT is therefore the only reliable source of individual differences, permits us to measure the speed of information processing when the capacity of working memory is not threatened by the complexity of the task. Increasing the complexity of the ECT, as in going from 1 bit to 3 bits in the Hick paradigm, increases the RT–*g* correlation. But when the task complexity is so great as to exceed the capacity of many subjects' working memory, then the number of erroneous responses, rather than RT, becomes the stronger correlate of *g*.

It is noteworthy that RT correlates only with *g* and not with any other psychometric factors independent of *g*, even when the reaction stimuli of the ECT are specifically designed to be either spatial, or verbal, or numerical and the psychometric tests with which the RTs are correlated are either spatial, verbal, or numerical. Statistically remove the general factor from the three types of psychometric tests and their correlation with the verbal, spatial, or numerical RTs is virtually zero.

Thus RT is a rather ideal tool for experimentally studying the task variables that cause a given task to be more or less *g* loaded. RT is highly sensitive to rather subtle experimental manipulations of the quantity of the information load of the task, which can be varied experimentally without in the least altering the type of information content or any of the stimulus or response aspects of the task. Increasing the task demands (resulting in RTs of not more than about 1 s) is found to increase the *g* loading of the RT parameters. In the Hick paradigm, for example, one-bit increments in the reaction stimulus cause, on average, only about 30 ms. increments in RT. The correlation of RT with IQ increases linearly from about −.20 for 0 bits of information to about −.30 for 3 bits. These are small but significant correlations, and their linear slope as a function of bits is also significant.

2.5 Open questions

The original experiment by Roth (1964) based on the Hick paradigm reported a significant correlation (−.39) between IQ and the slope of RT as a function of bits. This was a theoretically important finding, which meant that as the information processing demand of the task was increased, the advantage of higher IQ increased, as reflected by the higher IQ subjects' relatively faster RT. The majority of later studies, however, failed to replicate the relatively large correlation between IQ and RT slope originally reported by Roth, and this aspect of the RT–IQ relationship was more or less dismissed as if it had been totally discredited. True, the N-weighted mean correlation based on 32 independent studies was only −.165, though it is significant at $p < .001$. But the test–retest reliability of the slope measure (determined in six studies) is only .39. When corrected for attenuation, the RT slope correlates with IQ about −.26. For comparison, in the same 32 studies, the disattenuated mean correlation between IQ median RT is −.24 ($p < .001$), and the mean disattenuated correlation between IQ and RTSD (i.e., intraindividual variability of RT) is −.34 ($p < .001$). Thus when test–retest reliability is taken into account, the slope parameter shows, on average, about the same correlation with IQ as the median RT, and RTSD has the largest correlation with IQ.

Groups that differ in their average IQ show sizable mean differences in the Hick slope parameter, always in the theoretically predicted direction. For example, school students in regular classes and students in academically "gifted" classes differ .50 to .70 of a standard deviation (SD) in mean slope; university students and vocational college students differ about .50 SD; and black and white vocational students differ 0.34 SD (all of these differences significant beyond the .01 level).

Thus the Hick slope parameter accords with the theoretical implications first noted in Roth's study and elaborated upon in Eysenck's (1967) article. The measure of RT slope, like median RT, appears to be correlated only with the *g* component of a test's variance. When each of the 12 subtests of the Wechsler IQ battery was correlated with slope in the Hick paradigm, there was a rank-order correlation of −.83 ($p < .01$) between each of the subtests' *g* loadings and the degree to which each subtest is correlated with slope. That is, a test's correlation with slope proved a highly valid predictor of the test's *g* loading. If the three parameters of RT—median RT, slope, and RTSD—can be interpreted as measures of the speed and efficiency of information processing, then we may conclude that the speed and efficiency of information processing is at least a part of *g*. When RT parameters from a number of different ECTs are combined (either by multiple regression or by simple addition of their unit-weighted *z* scores), the RT–IQ correlations approach −.60.

2.6 The odd-man-out discrimination task

To increase the information processing demands of the RT task, using the same RT–MT apparatus shown in Figure 11.2, Frearson and Eysenck (1986) cleverly modified the procedure to create an odd-man-out discrimination task. All eight of the light-button alternatives are in view on the subject's response console. On each trial, three of the eight lights go "on" simultaneously; two of the lights are always closer together than the third, which is the odd-man-out. For example, if we imagine that the light-buttons are numbered from 1 to 8, the odd-man-out pattern would be such as 1, 2, *4*; or 2, 4, *7*, and so on. (With 8 light-buttons there are 44 possible odd-man-out patterns.) The subject is instructed to respond as quickly as possible to touch the odd-man-out button, which instantly turns off all of the lights. Again, RT is the interval between the three lights going on and the subject's releasing the home button. While the 3-bit condition of the Hick yielded RT–IQ correlations of about −.30 in samples of university students, the odd-man-out procedure resulted in RT–IQ correlations of about −.60, which approximates the average correlation between the individual subtests of the Wechsler scale and the Full Scale IQ, and is about the size of the correlation between the Wechsler IQ and the Raven IQ in our student population.

Using the odd-man-out procedure along with the Hick procedures (all using the same response console), we tested over 800 white and black school children in grades 4–6 to test a hypothesis of Spearman's using RT–MT measures rather than ordinary psychometric tests. Spearman (1927) had suggested that the size of the standardized mean black–white difference on various psychometric tests is directly related to the test's *g* loading. Spearman's conjecture had already been strongly borne out in 12 studies based on conventional tests (Jensen, 1985). The question was whether it would be borne out using RT in tasks that varied in information load (and hence in *g*) but contained no specific informational content. The tasks were so easy that the most difficult task in the battery (the odd-man-out task) could be performed with 100% accuracy by 4th to 6th graders with RTs of less than 1 s (the average being about 700 ms) The *g*-loadings of the various RT and MT measures derived from the Hick and odd-man-out paradigms were indeed correlated with the standardized mean black–white differences on these chronometric measures in accord with Spearman's hypothesis and the correlation was even higher than for conventional psychometric tests. The correlation between the *g* loadings of the chronometric measures and the mean black-white differences on those measures was .80, $p < .01$ (Jensen, 1993b).

*2.7 The non-*g *RT factor*

For a long time, researchers thought there was a virtually inexorable correlation ceiling, at about −.35, between RT and IQ. Research with the Hick

paradigm produced results that were generally consistent with this rather low correlational ceiling, even when the RT measures were corrected for attenuation or were made highly reliable by averaging RTs over a very large number of trials. (I found that increasing the number of RT trials by any given amount yields reliability coefficients that accord perfectly with the reliability predicted by the Spearman–Brown prophecy formula.) The ceiling, therefore, cannot be blamed on reliability. Increasing the information processing demands of the task, as in the odd-man-out procedure, surely breaks through the −.35 ceiling, but the odd-man RT has its own correlation ceiling, between −.5 and −.6. As already noted, if the processing demands result in RTs greater than about 1 s (in university students), the RT–IQ correlation markedly shrinks.

I hypothesized that the RT measured by any one procedure is much like a highly homogeneous psychometric test, that is, one in which all of the items are so equivalent as to be almost identical. We know that such an extremely homogeneous psychometric test, whatever its content and however reliable its scores, has a relatively low correlation with psychometric *g*. The most *g*-loaded tests are those that have quite heterogeneous items in terms of their types of cognitive demands. The Wechsler Full Scale IQ, for example, is much more *g* loaded than any one of its relatively homogeneous subtests. The items of the highly *g* loaded Raven Matrices superficially appear to be highly homogeneous, but actually the Raven items demand many different types of problem solving.

Therefore, according to my hypothesis, it should be possible to obtain much higher correlations if we combined the RTs from a number of quite different ECTs, each of which was simple enough to minimize erroneous responses and elicit RTs in the range below 1 s. As in ordinary psychometric tests, the specificity of each ECT should average out, allowing the emergence of the common factor in the RTs from each of the different ECTs, namely *g*. So we used a battery of different ECTs, each of which theoretically tapped a different information process (indicated here in parentheses): the Hick (stimulus apprehension and choice), the odd-man-out (discrimination), the Neisser paradigm (speed of visual scanning), the Saul Sternberg paradigm (speed of scanning STM), the Posner paradigm (speed of accessing information in LTM), the semantic verification test (speed of matching symbols with meanings), dual tasks (divided attention between two tasks, thereby straining working memory capacity), and inspection time (speed of making a simple visual discrimination). Indeed, the simple summation of the RTs and RTSDs obtained from all of these paradigms resulted in a correlation with psychometric *g* slightly greater than .60 and approaching .70 after corrections for attenuation and restriction of range in the college population. The correlations are scarcely larger between different standard psychometric tests, such as the Wechsler, the Raven, the Terman Concept Mastery Test, and the Multidimensional Aptitude Battery.

But a hierarchical factor analysis of the correlation matrix containing a number of conventional psychometric tests along with RT measures derived from various ECTs reveals why there is an inexorable ceiling to their correlation with psychometric *g*, regardless of how many different RT measures we may combine in a single score. The reason is that the variance of each RT measure based on a different paradigm does not consist only of *g* plus the specificity of each paradigm. There is another quite large factor besides *g* common to all of the various RT measures, which can be called a non-*g* RT factor (Jensen, 1994). (There is also variance that is specific to each RT paradigm.) In other words, RT tasks all measure *g* to some extent, but they also measure an RT factor that is unrelated to any factors measured by conventional psychometric tests. The total true-score variance of RT is divided between *g*, a non-*g* RT factor, and the specificity of the particular ECT. The *g* and non-*g* components of RT vary with the complexity of the ECT, the *g* component being larger in the more complex tasks. But the ubiquitous presence of the substantial and apparently noncognitive RT factor rather severely limits the practical usefulness of any ECT, or even the combination of several ECTs, as an alternative method for measuring the same *g* factor that we can measure quite accurately and efficiently with a standard psychometric test. The noncognitive RT factor, which seems to reflect individual differences in a purely perceptual-motor speed or coordination ability, may be of interest in its own right, and it is presently being researched by personnel psychologists in the Air Force for its possible predictive validity in the selection of recruits for pilot training.

2.8 Top-down, bottom-up and physiological explanations

Hardly a month goes by without some new ECT for measuring RT appearing in the psychological literature (particularly in the journals *Intelligence* and *Personality and Individual Differences*). In nearly every study there is found a significant correlation between RT and psychometric *g*, and each such study usually throws some light on the experimental variables that affect this correlation. What is still unclear is the precise basis of the correlation between RT and the *g* derived from nonspeeded psychometric tests—whether it is a matter of individual differences in RT being influenced by whatever higher-level cognitive processes are possessed by high-*g* persons (who are high-*g* for reasons having no causal relation to RT), or whether RT reflects differences in the speed and efficiency of the basic neural processes that cause differences in psychometric *g*. These two alternative possibilities are known as the "top-down" versus the "bottom-up" theories of the RT–IQ correlation. Different researchers prefer one or the other, but the issue has not been definitively decided empirically.

T. E. Reed and I had hoped that by finding a correlation between nerve conduction velocity (NCV) in the brain's visual tract, from the retina to the visual cortex, it would rule out the top-down theory, because the NCV is measured long before the neural impulse has reached the higher association centers that are necessarily involved in the kind of knowledge retrieval or problem solving typically demanded by untimed psychometric tests. There was indeed a significant correlation (−.27, corrected for range restriction in a college sample, −.38) between individual differences in NCV and IQ (Reed & Jensen, 1992). This finding clearly supports the "bottom-up" hypothesis, at least as regards NCV and IQ. But alas, it does not enlighten the issue regarding RT and IQ, because we found that the measure of NCV in the visual tract is not correlated with RT as measured by the Hick or the odd-man-out procedures (Reed & Jensen, 1993). This puzzle suggests new hypotheses, but there is as yet no compelling explanation. Speculation should be postponed, however, until a replication of these results insures their reliability.

3. CONCLUSIONS AND PERSPECTIVE

Clearly, much remains to be learned about the nature of *g* through further investigations into the causes of its relation to RT. Using a variety of experimentally manipulated ECTs to achieve the maximum possible correlations between RT and *g*, it is then possible to investigate the physiological correlates of these RTs with measures of NCV, AEP, and glucose metabolic rate in localized regions of the brain. There is good reason to believe that a program of research utilizing such techniques carried out by a number of independent laboratories dedicated to this common goal will, in the foreseeable future, realize what Spearman (1927) envisaged as the aim of research on human intelligence: "The final word on the physiological side of the problem [of *g*] ... must come from the most profound and detailed direct study of the human brain in its purely physical and chemical aspects."

I conclude by noting that it was entirely through Eysenck's influence that I began reading Galton and Spearman in the first place, and it is exceedingly improbable that my two decades of researching the connection between RT and *g* would ever have ensued had I not once heard Eysenck talk about Roth's experiment with the Hick paradigm and caught some of his enthusiasm for the subject.

REFERENCES

Eysenck, H. J. (1967). Intelligence assessment: A theoretical and experimental approach. *British Journal of Educational Psychology, 37*, 81–98.

Eysenck, H. J. (1994). A biological theory of intelligence. In D. K. Detterman (Ed.), *Current topics in human intelligence, Vol. 4, Theories of intelligence* (pp. 117–149). Norwood, NJ: Ablex.

Frearson, W. M., & Eysenck, H. J. (1986). Intelligence, reaction time (RT) and a new "odd-man-out" RT paradigm. *Personality and Individual Differences, 7,* 807–817.

Jensen, A. R. (1982). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93–132). Berlin: Springer.

Jensen, A. R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193–219.

Jensen, A. R. (1987a). Individual differences in the Hick paradigm. In P. A. Vernon (Ed.), *Speed of information processing and intelligence* (pp. 101–175). Norwood, NJ: Ablex.

Jensen, A. R. (1987b). The *g* beyond factor analysis. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 87–142). Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1993a). Spearman's *g*: Links between psychometrics and biology. *Annals of the New York Academy of Sciences, 702,* 103–129.

Jensen, A. R. (1993b). Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence, 17,* 47–77.

Jensen, A. R. (1994). Phlogiston, animal magnetism, and intelligence. In D. K. Detterman (Ed.), *Current topics in human intelligence, Vol. 4, Theories of intelligence* (pp. 257–284). Norwood, NJ: Ablex.

Reed, T. E., & Jensen, A. R. (1992). Conduction velocity in a brain nerve pathway of normal adults correlates with intelligence level. *Intelligence, 16,* 259–272.

Reed, T. E., & Jensen, A. R. (1993). Choice reaction time and visual pathway nerve conduction velocity both correlate with intelligence but appear not to correlate with each other: Implications for information processing. *Intelligence, 17,* 191–203.

Roth, E. (1964). Die Geschwindikeit der Verarbeitung von Information und ihr Zusammenhang mit Intelligenz. *Zeitschrift fur Experimentelle und Angewandte Psychologie, 11,* 616–622.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.