4 PSYCHOMETRIC *G* AND ACHIEVEMENT

Arthur R. Jensen

Introduction

Education's traditional value of enhancing the quality of life for the individual need not be eclipsed by the growing recognition of its importance to the national welfare. A well-educated population is now deemed crucial in this technological era. The cultivation of excellence in the kinds of achievement that depend on an educated work force is an undisputed goal in all industrial societies. Regardless of differences in coutries' political and economic systems, we see implicit agreement with Adam Smith's dictum that the wealth of nations depends on the abilities of their people. Virtually every head of state appoints a minister of education. The government of Venezuela, in addition, even appointed a "Minister for the Development of Intelligence." Obviously, the modern world perceives the supply of educated intelligence as vitally related to the general welfare.

Our own nation's anxiety about the general level of attainment in our schools was voiced officially in *A Nation at Risk: The Imperative for Educational Reform*, a report of the National Commission on Excellence in Education (1983). The commission noted studies from the past two decades that indicate a general decline in the amount of scholastic learning, with achievement levels below those of many other industrialized countries. It also recognized the continuing inequality between majority and minority racial and cultural groups in the outcomes of schooling. Federal policies and programs to promote equality of educational opportunity, after some thirty years, have not yet massively impacted on the racial disparity in educational outcome or its correlated disparity in the job market.

Such formidable and complex problems obviously have too many layers and facets to be grasped from any single viewpoint. There are problems within problems, questions within questions, and each by itself is grist for study. The goal of any research addressed to these problems cannot be like that of the alchemist trying to discover the philosopher's stone, which would all at once answer our questions and remedy the problems highlighted in *A Nation at Risk*.

The extreme diversity and complexity of the problems dictate that the task for any one researcher must necessarily be quite limited. The only feasible tack for the individual researcher is to divide the problem—divide and divide, until some scientifically tractable part is in hand, even if only a small facet of the multifaceted problem. Unfortunately, any single investigator's limited part of the divided effort is dwarfed by the immediate larger problems, and politicians and the general public can get impatient with the plodding and piecemeal scientific approach. Researchers are easily accused of fiddling while Rome burns. But remedies for educational problems seldom arise from global or monolithic notions of their nature. Broad-brush prescriptions scarcely penetrate causal underpinnings.

A realistic goal, I would suggest, is not to create a Grand Solution, but rather to make many small and specific, yet socially consequential, improvements in the particular troubled aspects of schooling. Significant improvement in educational outcomes will most likely only result as the cumulation of many small positive effects of a great many causal factors. Some of these factors lend themselves to educational implementation. But there are also other influences outside the schools' domain that impact on educational outcomes (for example, the "social pathology" in the culture of poverty that blights many urban schools, as vividly depicted by Maeroff (1988). Positive change in this sphere will depend on social reforms and influences far beyond what the educational system can effect alone.

The Special Focus of This Chapter

This chapter, although inevitably related to a larger context, necessarily focuses on a relatively narrow aspect: the relevance of psychometric tests and cognitive psychology for children's scholastic achievement and adults' successful employment.

More exactly, my primary concern is not with the practice of psychometric testing or with questions of test validity, cultural bias, or the fairness of using tests for educational selection and hiring. These topics, although of great importance in their own right, are incidental to my present aim. Nor am I interested in making a case for the routine use of standardized tests in schools or anywhere else.

Psychometrics, however, is an essential tool for studying individual differences and the outcomes of instruction. Lately, psychometrics has been allied with theories and methods of experimental cognitive psychology in the study of information processing, with educational implications for benefiting students ranging widely in aptitude (Snow and Lohman, 1988). Research by personnel psychologists in the armed services indicates that psychometrics and information-processing theory can be brought to bear on selection and training. It has proved particularly important for enlistees who did very poorly in regular school but have benefited from appropriate training programs in the service, permitting them eventually to enter the skilled work force (Sticht et al., 1987).

The main question addressed here is not whether the use per se of psychometric tests is a source of problems, for instance, by limiting opportunity in education and employment. (Whatever the answer, it is not the issue here.) The main question addressed here is: Do mental tests, in fact, measure something that is intrinsic to the larger problems of education previously mentioned? If tests do measure some factors intrinsic to the problems, rather than factors that are merely symptomatic, we then should inquire how we might be able to get around these factors—or, if not get around them, at least take them into account as constructively as possible.

A further limitation: I shall focus here only on those factors that affect achievement in a normally calm, orderly atmosphere for learning and a desire on the part of both learners and teachers to cooperate. Talk of instructional techniques is pointless where discipline is grossly lacking and a defiant attitude toward school prevails. Possible applications of cognitive research are rendered impossible where educational aims are flagrantly obstructed, for example, by the growing social pathology that threatens many inner-city schools—behavior problems, drugs, teen pregnancy, parental indifference, truancy, school dropout, vandalism, gang intimidation, violence, and crime. Such misfortunes spell an altogether different order of school problems from those that stem directly from the inherent difficulty of the material to be learned or the considerable differences in aptitude reflected by psychometric tests. Such adverse conditions for scholastic performance would even block satisfactory achievement by students we would recognize as academically gifted under more favorable circumstances. What is too often lacking in school failures is not ability but the kinds of values and aspirations that inspire achievement.

Although we may be tempted to speculate about possible causal connections between these extrinsic behavioral problems and prior psychological factors, this will be eschewed in the interest of focusing more intensely on fewer and more closely interrelated issues. This decision should not be misconstrued as belittling the problem of school discipline. In certain schools discipline is undoubtedly the first order of business. No intrinsic educational improvement can possibly take place without it. The public was recently reminded of this ancient wisdom through the wide media coverage of Joe Clark, the dynamic New Jersey principal who ruled his ghetto high school with a bullhorn and baseball bat. (See the cover story in *Time*, February 1, 1988, 52–58). Clark's "get-tough" methods, including removal of habitual troublemakers, are hotly debated by school authorities, but all seem to agree with Clark's insistence that discipline is the sine qua non for pupils' standing a chance to benefit from their time in the classroom.

While this chapter pertains mainly to individual differences, it would be awkward to avoid any mention of racial group differences. The generally lower scholastic achievement of certain minorities, particularly blacks and Hispanics, is itself a leading topic in public discussions.

Although some of the psychological factors in average group differences could probably be discussed entirely at the level of individual differences, it is important to understand the connections between the phenomena associated with individual differences and those associated with group differences. The achievement gap between racial groups, even assuming the mean difference is analytically indistinguishable from individual differences of similar magnitude, itself generates another whole class of distinctive phenomena—educational, social, and political. These cannot be properly understood without inquiring whether group differences in school learning reflect differences in the same psychological processes that characterize individual differences of similar magnitude or whether they reflect differences in social, cultural, or linguistic factors that are superimposed on the individual differences within a particular group.

Hence, group differences, as well as individual differences, must be studied at the same basic level of analysis that experimental cognitive psychologists are now studying information processes related to scholastic skills. Investigation of group differences should not be constrained by thinking exclusively in terms of broad sociological factors that are largely beyond the school's control, instead of looking at the cognitive processes directly involved in reading, writing, and arithmetic.

Cognitive psychologists make the reasonable assumption that *all* people possess the same basic information processes, the very processes involved in scholastic performance. But people also differ from one another in the speed, capacity, or efficiency of these processes. Moreover, within any one person the various processing components differ in efficiency, and there is now evidence to suggest that differences in the efficiency of certain components are related to chronological age. Much the same thing can be said of conative or motivational factors, as well as of cognitive processes.

What researchers discover about the relation of elementary cognitive processes to manifest achievement by looking only at individual differences, however, may or may not be the same for group differences. It is not known, in fact, whether racial or other group differences in the speed, efficiency, or capacity of such basic processes even exist or are related to the observed group differences in scholastic achievement and psychometric *g*. But these are empirically researchable questions.

The Organization of This Chapter: An Outline

1. Since the primary concern of the National Commission on Testing and Public Policy is the impact of psychometric tests on opportunity allocation in education and employment, the first question necessarily must be: What do these tests actually measure? Why is whatever they measure educationally, socially, or economically important? What is their impact on minorities?

In virtually all the tests of concern, the chief "active ingredient" (*latent variable* in psychometric terminology) is something called the *g* factor. It inescapably holds center stage in this inquiry. Any discussion of human mental abilities that fails to recognize the central role of *g* would be like *Hamlet* without the Prince of Denmark.

2. Next we ask: What do we know about *g*? Is it measurable? What can we say about the characteristics of its distribution in the population? Then there is the question of the external validity of *g*: Is it related to real achievement? If so, how much? Can the answer to this question go beyond just a coefficient of correlation? This raises the question, Can we quantify achievement? What can we say about the characteristics of its distribution, and what bearing might this have on the validity of *g* for predicting achievement?

3. With the stage now set, we can ask: Are there cracks in the psychometric edifice through which lower scoring individuals or groups might to some extent escape the adverse impact of *g*-loaded tests—that is, without doing away with the use of tests altogether? (Realistically, this isn't going to happen.) Are there valid *psychometric* arguments by which we might to some extent be able to get around the observed correlation between *g* and achievement? What about the advantages (and disadvantages) of using within-group percentiles to reduce adverse impact where tests are used for selection? Can test coaching be effective?

4. Going beyond purely psychometric considerations, should educators try to raise *g* itself through psychological means? Why has *g* been so resistant to such efforts? We occasionally see a dramatic change in a child's IQ without apparent cause. Then, there is the puzzling finding that scores on *g*-loaded tests have shown a gradual rise in the entire population of the industrialized world over the past forty years, without known cause. Yet the possibility of intentionally raising the level of *g* in targeted groups, rather than merely training up performance on particular tests, remains in doubt.

5. If differences in *g* are not presently amenable to intentional change, then are there ways—in school and in occupations—by which we can appreciably reduce the apparent importance of *g* for success? Can other abilities or traits be substituted for *g*? I will look at the well-known ideas on this matter, the Level I—Level II notion, as well as aptitude × training interactions (known as ATI) in general, also: mastery learning, programmed instruction, training thinking skills, and directed learning based on hierarchical task analysis. How much does achievement depend on motivation? What in fact is motivation, and can it be intentionally enhanced?

6. Finally, I will try to interrelate psychometric *g*, learning, and achievement in a general model of information processing. Information-processing concepts afford a closer, more analytic view of the action level at which learning and achievement take place. The information-processing model, unlike factor-analytic models of the structure of abilities, posits a number of specific mechanisms that have different causal relationships to various kinds of learning. The component processes are independently measurable, at least in principle. Hence, besides its explanatory value (which is important from a scientific standpoint) an information-processing model also interfaces closely with instructional methods. Empirical research on information processing is already demonstrating the relevance of such constructs as *working memory* and *controlled and automatic processing* to instruction in reading and math. On the other hand, it is generally conceded that global measures of *g* (IQ tests, for example) have meager implications for improving instruction.

What Psychometric Tests Measure

Item Performance, Abilities, Factors, and g

A discussion of what our psychometric tests actually measure requires first that we be clear about the definitions of a few technical terms.

Definitions, of course, are not arguably either true or false in any real sense; they are conventions mutually agreed upon for the sake of precise communication, and they can be judged only on that basis.

On that basis, therefore, I must begin by dismissing the term *intelligence*. It will not be used in any of the ensuing discussion, except with quotation marks, and occasionally to warn readers against confusing it with other terms I hope to use with more precision than is possible for intelligence.

As elaborated elsewhere (Jensen, 1987a), I have been forced to the opinion that intelligence is not a scientifically useful term; it has no generally agreed upon meaning, and psychologists seem hopelessly unable to achieve a consensus on what this term should mean (see, for example, Sternberg and Detterman, 1986). Moreover, the word intelligence is fraught with many prejudices and emotional connotations that render it a stumbling block to serious discussion. It should be relegated to popular parlance and literary usage. The problem is only worsened, in my opinion, by talking about *multiple intelligences*. I am not urging that we should try to agree on a proper definition of intelligence, or that it needs to be redefined. That would completely miss the point, for it so happens that the term is simply unnecessary for our purpose, and no other term needs to be substituted for it. I believe that psychology, regarded as a natural science rather than a literary art, not only can get along without "intelligence" in its technical vocabulary, but is much better rid of it.

The few objectively definable key terms that are essential for understanding the subsequent discussion are *item performance*, *ability*, *cognitive factor*, *g*, and *process*. (I urge the reader not to skip over these definitions because some may differ from the meanings these terms have acquired in other contexts.)

Item Performance. Empirical psychometrics must have its basis in objectively observable behavior. But as yet there is no precise standard term for the observable units of behavior upon which the theory and measurement of mental ability must ultimately rest. It must be clearly seen that the superstructure of abstractions, interferences, and theoretical constructs in psychometrics is firmly grounded in objective reality. So to serve this essential purpose, I shall adopt the term *item performance* (IP).

By IP I mean an observable unit of behavior or some objective record of it. An IP is a single, narrowly circumscribed, overt act that occurs at one point in time. It could be a person's response to a single item on a test of any kind; it could be spoken, written, or registered manually, for example, by pressing a button on a reaction-time apparatus. But an IP is not necessarily a part of a test; it is any observable act. An IP is objective only in the sense that there is a high degree of agreement among observers that the IP (or a recorded trace of it) in fact occurred. A person's single act of lifting (or failing to lift) a two-hundred-pound barbell at least two inches above the floor for at least five seconds is one instance of an IP. But for the present exposition, the best example of an IP is a person's response to a single item on an objective test on one occasion.

Now, there are different domains into which IPs may be classified. We are here especially interested in IPs that can be classed in the *ability domain*. An IP is not itself an ability as I shall henceforth use this term. An IP qualifies for classification into the ability domain, however, only if it meets the following two conditions:

1. The IP must be an intentional or voluntary act. This excludes unconditioned and conditioned reflexes, tics, involuntary movements, autonomic emotional reactions, and the like.
2. An IP can be viewed as either a discrete or a continuous variable, or as a simple or complex performance. But to be classed in the ability domain, the IP (or a record of it) must be objectively classifiable or quantifiable in terms of a standard, for example: running the 100-yard dash in *x* seconds; recalling in correct order seven out of seven presented random digits; pressing a button in *x* milliseconds at the sound of a tone; correctly dividing two-thirds by one-third; or stating whether the words *generous* and *parsimonious* are synonyms or antonyms. Hence, IPs in the ability domain must in principle be objectively classifiable or measurable. Objective classification or quantification of the IP only means a certain specified high degree of agreement among observers of the IP or their readings of appropriate measuring instruments.

To summarize: An IP in the ability domain is a particular observable act which, in principle, can be referred to some objective standard. It is important to emphasize that an IP is not an abstraction. It is not a hypothetical construct or a latent variable underlying the observed act. The IP is the act itself, or some objective record of it. The number of different possible IPs is theoretically unlimited.

Ability. An ability is a psychometric abstraction. It is defined in terms of a number of IPs and the relation between them. IPs can represent an ability if they meet both of two criteria: (1) temporal stability (to some specified degree) and (2) generality (to some specified degree). These criteria can be established only on the basis of a number of IPs. To meet the stability criterion, a particular IP must be repeatable over some specified interval of time, which may be anything from minutes to years. Some criterion of stability is required to rule out accidental, adventitious, or chance IPs. On a simple reaction-time (RT) test, for example, the performance on each trial constitutes an IP, and a person may have an RT of 150 milliseconds on a single trial, although the person's median RT based on 100 trials distributed over a twenty minute testing session might be 350 milliseconds. The single trial of 150 milliseconds would scarcely be representative of the person's simple RT. The deviation of the single IP from the average of the IPs over a number of trials constitutes an error in the measurement of the ability that is common to all of the IPs and is best represented by their central tendency (mean, median, or mode).

To meet the generality criterion there must be evidence of some specified degree of consistency or correlation between a number of IPs from a class of highly similar IPs. A person's recalling a set of *n* digits after one presentation is an IP, but the IP might be idiosyncractic; for example, the digit series might have contained the person's telephone number, thereby making it much easier to recall than some other series of digits of the same length. We cannot known how stable or representative the person's digit span is without observing the person recalling different sets of *n* digits. Our criterion for saying that the person has the ability to recall a set of *n* digits may be, say, perfect recall on four out of five trials. Any particular IP may deviate to some extent from the average of some number of IPs of the same type. Such deviations constitute error with respect to the measurement of the ability that is general to this class of IPs.

For conceptual clarity, at this point, the term ability should be reserved for measures obtained from a homogeneous set of IPs (that is, IPs drawn from a quite narrowly specified domain) that meets specified criteria of stability and generality. (Later on, we will see that abilities can be conceptualized in terms of hypothetical factors differing in breadth or generality.)

The stability of a set of IPs is typically quantified by the test-retest reliability coefficient of the set. Generality can be quantified by the average intercorrelation between the IPs in the set or by an index of item homogeneity (which is monotonically related to the average item intercorrelation), such as proposed by Loevinger (1947). Highly homogeneous sets of IPs can be called *testlets*. A testlet, thus, is a measure of an ability having a specified stability (for example, test-retest reliability for a specified test-retest interval) and a specified generality (for example, an index of item homogeneity).

Some psychologists try to make a distinction between *ability* and *skill*, but this distinction is usually connected with notions of "innate" and "acquired" or, more generally, of latent and observed variables. These concepts have a necessary place at another level of analysis, but they should not get mixed up with the basic definitions of the observable IPs and the abilities (or factors) derived from their intercorrelations. Skill may be viewed as a subordinate category of ability in certain contexts, but at this point there is no need to make a distinction between the meanings of skill and ability.

The number of different abilities is theoretically unlimited. Questions pertaining to the origin or history of an ability, whether it is innate or acquired or some interaction of these influences, or whether it is an aptitude or an achievement—these are all separate issues and are wholly irrelevant to the present definition of ability.

Cognitive. In referring to the ability domain, we need some explicit criteria for deciding whether a given ability should be considered a physical ability or a cognitive ability. The question in some cases is not as simple as one might imagine, and there is probably no set of criteria that could unequivocally classify every conceivable ability as either physical or cognitive. Operationally, we can classify abilities in terms of correlations, including only those abilities in a given class that show positive correlations with one another larger than some explicit chosen value. This approach would generally distinguish between most physical and cognitive abilities, but there would surely be a good many ambiguous abilities. The seeming problem is probably best resolved by theorizing that every ability has both physical and cognitive components in varying degrees. The adjective *cognitive*, as in *cognitive ability*, would then mean simply that independent measures of individual differences in simple sensory and motor functions per se do not account for the major part of the population variance in the particular ability.

In psychometric discussions, ability means cognitive ability, unless qualified otherwise. Cognitive and mental ordinarily have the same meaning and are often used interchangeably. But cognitive is the more precise term, having to do specifically with knowledge or the process of knowing, which includes attention, perception, encoding and transforming information, learning, remembering, thinking, and all the other aspects of information processing.

Factor. The context in which *factor* occurs is the clue to whether one of its ordinary dictionary meanings applies, or its strictly statistical meaning (that is, as one of the conditions or classifications of variables in an experimental design or analysis of variance), or its specialized meaning in a variety of closely related multivariate mathematical techniques known as principal components analysis and factor analysis. The term factor is used in this chapter almost exclusively in the last sense.

The explication of factor analysis is beyond the scope of this chapter. Briefly, factors are hypothetical variables underlying observed or measured variables. They are thus latent variables, but no causal relationship between the latent and observed variables is necessarily implied. The relationship between the two is best thought of as a strictly mathematical one, in the nature of a mathematical transformation. The methods of factor analysis are the means for achieving such a transformation. The total variance of an observed variable, for example, can be analytically represented as a linear (additive) composite of a number of independent variances, each one attributable to a different hypothetical factor. Hence, it may be possible to represent most of the variance in a number of observed variables in terms of a considerably smaller number of underlying hypothetical variables, or factors; thus, we can speak in terms of n variables and p factors, when $n > p$. The system of mathematical manipulations involved in factor analysis permits one to determine the coefficients of correlation between (1) each of the n measured variables and (2) each of the p hypothetical variables, or factors "underlying" the n observed variables; such correlation coefficients are termed *factor loadings.*

The essential input for a factor analysis is the total matrix of correlation coefficients among all of the measured variables to be analyzed. The end product is a *factor matrix*, which shows the factor loadings (that is, correlations) of each of the n measured variables on each of the p factors. The part of the total variance of the n variables that is accounted for by the p factors is termed the *common factor variance*, which (since $p < n$) is necessarily less than the total variance of the n variables. The proportion of the total variance of any given one of the n variables that consists only of common factor variance is termed that variable's *communality* (symbolized h^2). (This symbol should never be confused with the identical but conceptually unrelated symbol for heritability used in quantitative genetics.)

Unlike observed variables, factors cannot be directly measured in individuals, but they can be estimated. Such an estimate of an individual's relative standing on a given factor is termed a *factor score.* By estimate is implied some degree of error, that is, a deviance of the estimated value from a hypothetical (hence not directly measurable) true value. The reason that individual factor scores are only estimates is essentially that the p factors account for less than the total variance of the n variables, and so any given variable's total variance is not fully accounted for by the p factors. But an individual's factor score is necessarily derived from a weighted average of the individual's standardized scores on each of the *n* variables. (The weights are related to the given factor's loadings on each of the *n* variables.) Hence, there is inevitably less than a perfect correlation between the estimated and the true (but unknown) factor scores. The size of the correlation between estimated and true factor scores is directly related to the average communality of the *n* variables and to the ratio *n*/*p*. When these values fall in the range typically found in factor-analytic studies, the minimum average correlation between estimated and true factor scores will range between about .60 and .90 (Gorsuch, 1983, 258–260). Factor scores are sometimes useful for certain research purposes but are hardly feasible (and are rarely encountered) in the practical uses of mental tests.

Distinction Between Factors and Item Performance. An *item* in a psychometric test is something that calls for a fairly specific response that meets a certain objective standard which some test takers can and some cannot attain. (An item which does not discriminate, that is, one which every test taker passes or every test taker fails, is obviously useless for the measurement of individual differences.) Hence a test item measures an IP, or item performance, as previously defined.

Although each single item on a psychometric test measures an IP, the raw score (the total number of items passed) on a psychometric test composed of a large number of items is at least one step removed from the IPs. The reason for this is not anything that is a logical, mathematical, statistical, psychometric, or methodological necessity per se. It is due purely to a fundamental *empirical* fact, namely, that *all cognitive performances are imperfectly but positively correlated with one another* in the general population. (This empirical generalization may not hold true in a group selected in such a way as to be highly restricted in cognitive abilities; the generalization fails as a direct function of the degree of restriction. Residents of an institution for the mentally retarded, for example, or the membership of the National Academy of Sciences are extremely restricted groups with respect to cognitive ability.)

Thus different test items that measure cognitive IPs show varying degrees of positive correlation with one another. It comes as a surprise to many that the average item intercorrelations in the best tests are very low—generally in the range of +.10 to +.20. Such low correlations indicate that any pair of items picked at random from the same test have very little in common (that is, only about 1 to 4 percent of the variance in one item can be predicted from the other). But the important empirical fact is that the items do measure something in common, however little that something is for any given item.

Hence scores on psychometric tests are measures of abilities of varying degrees of generality, depending on the homogeneity of the items. The higher the inter-item correlations (that is, the higher the homogeneity), the narrower, or less general, is the ability measured by the test score. The score on a psychometric test measures *common factors* plus some unwanted, uninformative "noise" called *measurement error*. This distinction between IPs and factors (or abilities) is an important one, because a good deal of argument and confusion about tests stems from the criticism of tests aimed at the level of single items. It is a common misconception that the IPs per se constitute the ability measured by the test score. In fact, most of the variance (that is, an index of individual differences) on any single item consists of the unwanted noise or error component, which contributes very little to the total variance of any test with high internal consistency reliability.

Psychometricians try to minimize the error component in the test scores. The higher the *internal consistency reliability* of test scores for tests of a given length, the smaller is their *error* (that is, variance contributed by item specificity), the less the test scores reflect any particular IPs, and the more they reflect common factors which define abilities. The total variance in test scores can be divided into two parts: (1) common factor variance among items (or twice the sum of all the item covariances) and (2) variance specific to each item (or the sum of the item variances). For most standard cognitive ability tests, the item-specific variance typically constitutes only about 10 percent of the total variance of the test scores. Increasing the number of test items (drawn from the same item population) reduces the proportion of the total variance in test scores that is attributable to error defined as *item-specific variance* and increases the proportion of what is termed the *true-score variance*. One definition of a test's reliability is the proportion of the total variance in obtained scores that consists of true-score variance. (Reliability and true-score variance are also defined in terms of the stability of obtained test scores between different points in time and are indexed by the test-retest correlation.)

But the main point that must be emphasized here is that the total variance in a distribution of test scores mainly reflects covariance among items rather than item performance per se. Hence, individual differences measured by test scores do not mainly reflect performance on this or that particular item on the test, but rather the total covariance among the items, and this cannot be described strictly in terms of any specific performance. Individual differences are measured not in terms of observed behavior (that is, IPs) but in terms of common factors, which are an abstraction, at least a step removed from the level of any specific performance. But the fact that individual differences in test scores reflect differences in an abstraction (namely, common factors) in no way diminishes their importance or objective reality. No less an abstraction is the force of gravitation, expressed by the *g* in Galileo's formula $S = \frac{1}{2} gt^2$.

Because common factors have different degrees of generality, they can be conceptualized hierarchically. There is a common factor in just two correlated items. Such a common factor would have relatively little generality and would thus be at the bottom of the hierarchy of generality. The common factor in a very large number of items selected at random from a vast pool of diverse items would have relatively greater generality than the common factor in just a few items selected at random from the same pool.

Here is how we would form a hierarchy of increasing generality. We begin with a large pool of diverse items. These items constitute the base level of the hierarchy. We administer all of the items to a large random sample of the population and obtain the matrix of correlations between every item and every other item. By inspecting the matrix of item intercorrelations we make up groups of items by selecting into any one group only those items that are correlated with one another more than, say, +.20. (An item that correlates less than +.20 with any other items is assigned to the group of items with which it has the highest average correlation.) Each of these groups of items, then, constitutes a testlet. Because the items within any one testlet are more highly intercorrelated than are the items across different testlets, each of these testlets is said to be comparatively more homogeneous than would be any set of less highly intercorrelated items. But whatever is measured by the total score on one of the testlets is something more general than what is measured by any one of the items in it.

Then we can go on and do the same thing with the testlets that we did with items—obtain the matrix of correlations between every testlet and every other testlet, and make up a number of groups of testlets (each such group termed a *test*), such that the testlet intercorrelations within each group of testlets are larger than the correlations across groups of testlets. Each group of such comparatively highly intercorrelated testlets constitutes a test. The test score in this case is simply the sum of the scores on the various testlets it comprises. The true-score variance of the scores on the test consists only of the covariances between all of its component testlets. Such tests are the third level of our hierarchy, and whatever the true-score on one of them measures, it is something more general than what is measured by any one of the various testlets that make it up.

We can go on in the same way to obtain a fourth level of the hierarchy, which we could label *super-tests*. Since there is necessarily a smaller number of groupings at each higher level of the hierarchy, we are finally left with a single group at the apex of the hierarchy; its true-score variance comprises only the covariances among the group of tests immediately below it in the hierarchy. It measures something even more general than tests at any lower level in the hierarchy. Whatever this something is, it seems quite removed from the various items of observable performance that formed the basis (the first level) of the hierarchy.

The pyramid-like hierarchical structure described in the preceding paragraphs serves as a rather easy-to-grasp nontechnical explication of *hierarchical factor analysis*, which is now almost universally considered the most appropriate type of factor analysis for research in the cognitive abilities domain. A technical discussion of it would take us too far off our main course, so readers must be referred elsewhere for more detailed information (Gorsuch, 1983; Gustafsson, 1988; Jensen, 1987b; Schmid and Leiman, 1957; Wherry, 1959). In brief, hierarchical factor analysis is a method for discovering the various latent (or underlying) factors (arranged hierarchically according to their level of generality) that account for the empirically obtained correlations among the multifarious cognitive abilities measured by testlets. The levels of the hierarchy, from the top down, are conventionally labeled as shown in table 4–1. All factors at higher levels than the first-order factors (also termed *primary* or *group* factors) are termed *higher-order* factors; their number depends on both the number and diversity of the variables (testlets or tests) that enter into the factor analysis. In the factor analysis of large batteries of diverse cognitive tests there are virtually never more than two levels of higher-order factors—usually consisting of two to four second-order factors and finally the single general factor, which is conventionally symbolized by an italicized lowercase *g*. In smaller batteries, such as the twelve subtests of the Wechsler Intelligence Scales, *g* comes out as a second-order factor. A hierarchical factor analysis of the Wechsler battery, for example, yields a *g* factor and the three first-order factors labeled verbal, spatial, and

Table 4–1. Levels of a Hierarchical Factor Structure

Level	*Descriptive Label*
5	General Ability Factor (*g*)
4	Second (or Higher) Order Ability Factors
3	First Order (Group or Primary) Ability Factors
2	Homogeneous Tests (Narrow Abilities)
1	Items (Observed Item Performance)

memory (Jensen and Reynolds, 1982). Primary factors are named in terms of the type of ability represented by the items that compose the tests in which the factors have their largest loadings.

It is important to understand that a general factor is not mathematically inevitable in a hierarchical factor analysis; it cannot emerge from the analysis unless it is actually latent in the set of variables that are factor-analyzed. It is possible to find (or construct) sets of variables that yield no general factor whatsoever when subjected to a hierarchical factor analysis. Hence, *the general factor identified by a hierarchical analysis is not a methodological artifact or mathematical necessity, but an empirical outcome.*

A hierarchical factor analysis is said to be *orthogonalized* when all of the variance that is common between any of the factors at a lower level is removed to the next higher level; this applies at every level throughout the whole hierarchy, going from the primary factors up to *g*. The result is that all of the factors in the hierarchy (both within and between levels) are perfectly orthogonal (that is, uncorrelated with one another). Methods for orthogonalized hierarchical factor analysis have been explicated by Schmid and Leiman (1957) and Wherry (1959).

Hence, the total variance on any one of the tests entered into the factor analysis can be viewed as the sum of a number of linearly independent components of variance attributable to common factors (that is, *g*, possibly other higher-order factors, and primary factors) and uniqueness (specificity + error). A test's specificity is that part of its true-score variance (twice the sum of all its item covariances) that it does not have in common with any of the other tests with which it was factor-analyzed.

The General Factor

The most important factor in the cognitive domain is the general factor, or *g*. This is so for a number of reasons. Probably the most important is the fact that *g* is more highly correlated with various indexes of learning, performance, or achievement outside the set of psychometric tests, from which *g* is derived, than is the case for any other factor or combination of factors (independent of *g*) that can be derived from the factor analysis of the same set of tests. In brief, *g* is the chief active ingredient in the concurrent and predictive validity of most psychometric tests in most of the situations in which tests are used. Also, the *g* factor accounts not only for a larger proportion of the common factor variance of various collections of diverse tests than any other factor, but it often accounts for more of the common factor variance than all of the other factors combined. For example, in a study of eighteen separate factor analyses of test batteries comprising anywhere from six to thirteen tests (averaging 11.1 tests), the *g* factor accounted on average for 4.3 times as much variance in test scores as all of the other common factors combined (Jensen, 1987c).

But it should also be noted that there is a great deal of uniqueness (specificity + random error) in tests. In the study just mentioned, for example, tests' uniqueness accounts, on average, for nearly one-half of the total variance in test scores. A test's specificity is usually problematic and is often virtually impossible to characterize precisely in psychological terms. Moreover, assuming a particular test was factor-analyzed among a large and diverse battery of other tests, our knowledge of the particular test's specificity would probably have no value for most of the practical purposes for which tests are generally used. For most of the criteria ordinarily predicted by tests, a test's predictive validity would probably be reduced to nil if its general factor and major group factors were partialled out.

The existence of *g* should not lessen the importance of other substantial group factors (for example, verbal, spatial, memory) and special talents (for example, musical, artistic, mechanical, motoric). But neither does the exisence of these group factors diminish the predominance of *g*. It is a popular misconception that every person has such large peaks and valleys across the total spectrum of abilities that it is virtually impossible to speak realistically of different persons as being higher or lower in abilities in some average or general sense. But the very existence and size of the general factor absolutely contradicts this notion. It is a logical corollary of *g* that the average difference between various abilities within individuals is smaller, in general, than the average difference between individuals in their overall average level of ability.

Now we must consider the three most commonly expressed doubts about the *g* factor. They are hardly compelling.

Different Methods for Extracting *g*. In the modern psychometric literature *g* is represented by any one of three methodologically and conceptually rather different methods: (1) as the first principal component (unrotated) in a principal components analysis, (2) as the first principal factor (unrotated) in a common factor analysis (also called *principal factor* or *principal axes* analysis), and (3) as the highest-order factor in an orthogonalized hierarchical factor analysis. It has been found to be true empirically (although it is not necessary mathematically) that the *g* extracted by any one of these methods is very highly correlated (usually above .95) with the *g* extracted by either of the other methods in the same set of tests.

The empirical reality of *positive manifold* (that is, the existence of nonzero positive correlations between all cognitive tests) is itself the only fundamental or necessary condition for inferring a general factor, and any correlation matrix displaying positive manifold will yield up a general factor by any method of factor analysis. The only exceptions are those methods, like Thurstone's (1947) multiple factor analysis, which necessarily and intentionally submerge the general factor by scattering all its variance among the (orthogonally rotated) primary factors. I have yet to find a bona fide empirical demonstration of correlations between cognitive ability tests which are negative or zero that are significantly replicable or cannot be explained by some combination of sampling error and restriction of range on *g* in the subject sample, Guilford's (1964) contention notwithstanding (Jensen, 1980, 224–226).

Although the various methods of factor extraction yield highly similar *g* factors, a hierarchical factor analysis is preferable for theoretical reasons which I have indicated elsewhere (Jensen, 1987b). Their explication is not essential for the present discussion.

Invariance of *g*. Although it is not a mathematical necessity, it is an empirical fact that the *g* factor is quite stable when extracted from different batteries of cognitive tests, provided the tests composing each battery are reasonably numerous and diverse in contents and task demands. In fact, the degree of invariance of *g* is a direct function of both the number and diversity of the tests. Also, a hierarchical *g* is generally somewhat more stable than either the first principal component or the first principal factor. I have found, for example, that estimated *g* factor scores derived from a factor analysis of just the six Verbal subtests of the Wechsler Adult Intelligence Scale (WAIS) are correlated .80 with the estimated *g* factor scores derived from a factor analysis of just the six nonverbal Performance tests. Yet there is no resemblance between the Verbal and Performance subtests in their information content or specific task demands.

A large-scale investigation of *g* invariance was conducted by Thorndike (1987). He began with sixty-five highly diverse tests used by the U.S. Air Force. From forty-eight of these tests, six non-overlapping batteries were formed, each composed of eight randomly selected tests. Into each of these six batteries was inserted, one at a time, each of the seventeen remaining "probe" tests. Hence, each of the six batteries was factor-analyzed seventeen times, each time containing a different one of the seventeen probe tests. The six *g* loadings obtained for each of the seventeen probe tests were then compared with one another. It was found that the six *g* loadings for any given test were highly similar, although the *g* loadings varied considerably from one test to another. The average correlation between *g* loadings across the six batteries was .85. If each battery had contained more tests from the same general test pool, it is a statistical certainty that the average cross-battery correlations between *g* loadings would be still higher. Thorndike's finding, which is consistent with similar studies, constitutes strong evidence that pretty much the same *g* emerges from most collections of diverse cognitive tests. This evidence also indicates that the invariance of *g* across test batteries does not depend on their having identical elements in common, in the sense of elements of test content. Even highly dissimilar tests (vocabulary and block designs, for example) can have comparably high loadings on one and the same *g* factor.

Just as we can think statistically in terms of the sampling error of a statistic, when we randomly select a limited group of subjects from a population, or of measurement error, when we obtain a limited number of measurements of a particular variable, so too we can think in terms of a *psychometric sampling error*. In making up any collection of cognitive tests, we do not have a perfectly representative sample of the entire population of cognitive tests or of all possible cognitive tests, and so any one limited sample of tests will not yield exactly the same *g* as another limited sample. The sample values of *g* are affected by subject sampling error, measurement error, and psychometric sampling error. But the fact that *g* is very substantially correlated across different test batteries means that the variable values of *g* can all be interpreted as estimates of some true (but unknown) *g*, in the same sense that, in classical test theory, an obtained score is viewed as an estimate of a true score.

Is *g* an Artifact? This question implies that *g* may have no significance or substantive meaning other than the mathematical technique used in deriving it. This is a false implication, for three main reasons.

First, a hierarchical general factor is not at all a mathematical necessity, and correlation matrices outside the cognitive realm can be found which yield no general factor. Therefore, the presence (or absence) of a hierarchical *g* is itself an empirical fact rather than a trivial tautology. It simply reflects the all-positive correlations among tests in the matrix, a condition which is not forced by any methodological machinations.

Second, as Lloyd Humphreys (1968) has argued, a highly replicable mathematical dimension that can be defined under specified conditions is real. It is real in the same sense that other scientific constructs (for example, gravitation, magnetic field, potential energy) are real and measurable, even though they are not directly observable or tangible entities.

Third, *g* is related to other variables and constructs that lie entirely outside the realm of psychometrics and factor analysis and have no connection whatsoever with these methodologies. For example, the degree to which various psychometric tests are *g*-loaded is highly related to their degree of correlation with variables such as the *heritability* of individual differences in the test scores, the *spouse correlations* and various *genetic kinship correlations* in the test scores, the *effects of inbreeding* (and its counterpart, *heterosis*) on test performance, *choice reaction time* to visual and auditory stimuli, *inspection time* (or the speed of visual or auditory discrimination), and certain features of the brain's *evoked electrical potentials*. (These studies have been reviewed in Jensen, 1987b.) No other factor that can be extracted from a collection of diverse cognitive tests shows as large or as many correlations with non-psychometric variables as does *g*. It is clear that *g* has as much claim to reality as theoretical constructs in other sciences. It is one of the major constructs in psychology, and one of the oldest and most well-established.

Spearman's Hypothesis

The *g* factor takes on further significance in the subsequent discussion of its central role in educational achievement through its connection with an hypothesis first suggested by Charles Spearman (1927, 379), the English psychologist who invented factor analysis and discovered *g*. Spearman noted that the average difference (in standardized score units) between representative samples of the black and white populations in the United States differ considerably from one test to another, and he commented that the size of these differences is directly related to the size of the *g* loadings of the tests on which the differences are found, regardless of the particular type or content of the *tests*.

I have formalized Spearman's notion, calling it *Spearman's hypothesis*, which states that the relative magnitudes of the standardized mean black/white differences on a wide variety of cognitive tests are related predominantly to the relative magnitudes of the tests' *g* loadings: the higher the test's *g* loading, the larger the mean black-white difference. This hypothesis, if true, would seem to have marked relevance for understanding the well-known black/white difference in scholastic performance. More generally, it would mean that understanding the nature of the statistical black/white difference in the cognitive domain depends fundamentally on understanding the nature of *g* itself.

A proper test of Spearman's hypothesis requires the following conditions:

1. The black and white samples must be fairly representative of their respective populations and should be sufficiently large so that there is small enough sampling error of the correlations among tests to yield stable factors; and the samples should not be selected on any variables, such as educational or occupational level, that would restrict the range-of-talent with respect to *g*.
2. The collection of psychometric tests should be fairly numerous to permit the extraction of a relatively reliable *g* factor.
3. The tests must be fairly diverse in content and task demands, both to ensure a stable *g* and to allow considerable reliable variation in the *g* loadings of the various tests.
4. The tests' reliabilities should be known so that the tests' *g* loadings (and also the standardized mean group differences) can be corrected for attenuation (that is, diminution because of measurement error).
5. The factor analysis must be carried out within either the white or the black sample (or both), but not in the combined samples, so that any differences between the samples cannot possibly enter into the factor analysis.
6. The similarity in the vector of *g* loadings extracted separately from the two groups must be sufficiently high to ensure that the same factor is represented in both groups, as indicated by a coefficient of congruence of .95 or above.

The statistical test of Spearman's hypothesis, then, is the rank-order correlation between the tests' *g* loadings (in either group) and the standardized mean differences between the groups on each of the tests (with loadings and differences corrected for attenuation).

I have investigated Spearman's hypothesis in eleven large data sets that meet these requirements, some more ideally than others (Jensen, 1985a, 1985b; Naglieri and Jensen, 1987). (They were the only published data sets [and hence are accessible to other investigators] that are appropriate for testing Spearman's hypothesis.) The hypothesis was borne out in every study. The larger the number of tests and the greater the dispersion of the tests' *g* loadings, the more strikingly the results accord with Spearman's hypothesis, namely, a large and significant positive correlation between (1) various tests' *g* loadings and (2) the sizes of the tests' standardized mean differences between the white and black samples.

My formalization (or reformulation) of Spearman's hypothesis, it is important to note, states that the variation in the mean black/white differences on various tests is associated predominantly (rather than exclusively) with the tests' *g* loadings. This weaker version of the hypothesis is dictated by the empirical finding that when we plot the linear regression of black/white differences on tests' *g* loadings, we find that certain tests consistently show moderate deviations from the regression line. Tests that have an appreciable loading on a *spatial* factor (block designs, object assembly, paper folding, comparison of rotated figures, and the like) consistently show a larger black/white difference than is predicted from the test's *g* loading. Tests with an appreciable loading on a *short-term memory factor* (digit span, verbal rote learning, digit symbol or coding) show a smaller black/white difference than is predicted by the test's *g* loadings. So far, these are the only two well-established psychometric factors that have been found to cause rather small but consistent perturbations in demonstrations of Spearman's hypothesis.

Some thirty scholars have published peer reviews of this work in *The Behavioral and Brain Sciences* (1985, vol. 8, 193–262; 1987, vol. 10, 507–537), but no one has refuted the empirical demonstration of my formalized statement of Spearman's hypothesis. Several critics, however, showed a misapprehension that the demonstration of the hypothesis was somehow a mathematical necessity or tautology rather than an empirical discovery, and that confirmation of the hypothesis was an inevitable result of the methodology for testing the hypothesis and thus purely an artifact. This is impossible, for two quite obvious reasons:

1. When Pearson correlation coefficients between tests are calculated, all information about the means and standard deviations of the tests (or their rank-order of magnitudes) is completely lost in the correlations. Consequently, nothing about the tests' means or their rank-order of magnitude can be inferred from the matrix of test intercorrelations. Ipso facto, nothing can be inferred about the rank-order of tests' means from the tests' loadings on *g* or any other factors extracted from the correlation matrix.
2. The test means of one or the other comparison group (either black or white) are experimentally independent of the data from the group that yielded the test intercorrelations and the *g* factor extracted from them.

These two self-evident statistical facts necessarily mean that the prescribed method for testing Spearman's hypothesis yields a result that cannot be an artifact or a tautology. If the hypothesis is indeed borne out, it must necessarily have the status of an empirical fact. (The only theoretically possible exception to this assertion would be in the unrealistic hypothetical case whereby the total variance of every test in the battery consisted exclusively of variance in *g* and variance due to random errors of measurement. In which case, any reliable group difference would necessarily be a difference in *g*, and variation in the group differences across the various tests would reflect nothing but variation in test reliability.)

Further analyses (Jensen, 1987c) of the data previously used to examine Spearman's hypothesis have revealed additional findings. Into each of eighteen independent correlation matrices, each comprising anywhere from six to thirteen tests (averaging 11.1 tests), with each matrix based exclusively on either a white or a black sample (but never a racially mixed sample), was inserted the point-biserial correlations of each of the tests in the particular matrix with the variable of race treated as a dichotomous variable (quantitized as black = 0, white = 1). Each matrix was factor-analyzed, with a minimum of three first-order factors extracted from each matrix. The average loading of the dichotomous race variable on the *g* factor was .55, whereas the average of the corresponding loadings on the three largest first-order factors (uncorrelated with *g*) was .24. In other words, the black/white variable generally had its major loading on the *g* factor. A spatial visualization factor is the only non-*g* factor that rather consistently rivals *g* in its loadings on the black/white variable (see also Naglieri and Jensen, 1987). Hence, the largest black/white mean difference is seen on those tests that are the most highly loaded on both *g* and a spatial factor. The smallest black/white mean differences occur on tests that are the least loaded on *g* and the most highly loaded on a short-term memory factor. Contrary to popular belief, the mean black/white difference on the verbal factor (independent of I) is nil.

Examination of 121 psychometric tests that were factor-analyzed in eleven studies also showed that the *g* loadings of various tests are distributed as a continuous variable extending over a wide range of values—from about .30 up to nearly .90. On the same set of tests, the black/white mean differences (expressed in standard deviation units) are also distributed as a continuous variable, ranging from close to zero up to about 1.3 standard deviations (SDs). From the linear regression of the mean black/white differences on tests' *g* loadings, the estimated mean difference on a hypothetically pure measure of *g* would be approximately 1.2 SDs.

Cognitive Processes

Cognitive *processes*, like factors, are hypothetical constructs. That is, they cannot be observed directly, but must be inferred from behavior. Conceptually, however, processes and factors are altogether different. A factor can arise only from variance and may be thought of as a dimension of individual differences. A process, on the other hand, is one of the operating mechanisms of the mind.

Presumably at some neurophysiological level, processes perform operations on mental representations or transformations of immediate stimulus inputs or the traces of encoded past inputs stored in either short-term or long-term memory. The identification and functional description of processes do not depend on an analysis of individual differences but can be sought through the experimental analysis of the performance of just one person. The usual methodology for this purpose is mental chronometry, or the study of the time course of information-processing in the nervous system (Posner, 1978). There are two main categories of hypothetical processes: elementary cognitive processes and metaprocesses.

Some of the elementary cognitive processes that have been identified are stimulus apprehension (simple awareness of some change in the stimuli impinging on the sensorium), encoding (selecting, recognizing, labeling, or categorizing a given sensory input), discrimination, decision or choice, and retrieval of information from short-term memory or from long-term memory. These elementary processes are inferred and measured chronometrically from a person's performance on a variety of very simple contrived situations known as elementary cognitive tasks (ECTs).

Metaprocesses, on the other hand, are the executive operations that deploy the appropriate elementary processes called for by a particular stimulus situation and govern the planning, sequencing, and execution of processes, as well as their automatization through rehearsal or repetition.

Processes are the operational basis, so to speak, of the cognitive abilities involved in any mental test item. Different items depend upon different processes, or overlapping combinations of processes, for their execution. A reductionist explanation of individual differences in a particular ability, or in the common factors (including *g*) derived from correlations among abilities, devolves upon an analysis of individual differences at the level of processes. It becomes a question of precisely which processes contribute to the variance in a particular ability or a given factor. Probably the prevailing theory of *g* is that it results from the existence of individual differences in some quite limited number of distinct elementary cognitive processes that enter into virtualy all cognitive abilities (see, for example, Detterman, 1987; Sternberg and Gardner, 1982). The measurement and analysis of processes and their relation to educational and occupational achievement are more fully explicated later in this chapter.

The External Validity of *g*

Measurement and Population Distribution of g

Discussion of the relationship of *g* to achievement involves questions about the measurement properties of both variables and the characteristics of the population distributions of these measurements.

An individual's standing on the *g* factor of a battery of tests can be measured by means of a factor score but this is seldom done. Conventional IQ tests, however, are always very highly *g*-loaded when factor-analyzed among a collection of diverse cognitive tests, and certain quite complex item-homogeneous tests, such as Raven's Progressive Matrices, are also very highly *g*-loaded. Scores on such tests may serve as a rough proxy for *g* factor scores, although there might be slight contamination by some group factors, most commonly either verbal or spatial

But raw scores (or any transformation or standardization of them) on multi-item tests are just like factor scores in one respect: there is simply no true or natural metric for test scores or factor scores as there may be for certain abilities that can be measured in physical units (for example, a sensory discrimination threshold and reaction time).

By a *true metric* I mean simply a scale with an absolute zero point, additivity of measurement units, ratio properties, and one that retains all these properties and has the same meaning across dissimilar phenomena measured on the same scale. This kind of measurement—only this kind of measurement—permits direct comparisons between dissimilar phenomena (for example, the moon has *x* times greater mass than a pint of water; or, the average distance between the planets Venus and Nepture is *x* times the diameter of a hydrogen atom). Psychometric test scores and factor scores unfortunately do not possess these ratio scale properties. Hence, we are more limited in the kinds of statements and inferences that can be made on the basis of test scores in any form. Many psychologists evince a surprising naiveté on this point. For example, to the best of my knowledge there are no existing tests or psychometric techniques that could render such statements as the following at all meaningful: "A person gains half of his or her adult level of mental ability by the age of five."; "The increase in intelligence (or *g*, or verbal, memory, spatial, and so on) between ages five and ten years is, on average, equal to *x* times the decrease in intelligence (or *g*, and so on) between ages seventy five and eighty."

The plain fact is that test scores (or anything derived from them, such as standardized scores, factor scores, normalized scores, mental age scores, and so on), when the total scores are not based on factor-homogeneous items to which responses are each measured on a single physical scale (for example, response times in milliseconds), can only represent at best an *ordinal* scale. Ordinality means that the only strictly interpretale information in the scores can be expressed only in terms of their rank-order. That is to say, any kind of numerical scores derived from tests do not have the properties of a true metric as in physical scales, and the most that any scores, even from the best-made tests, can actually permit us to do is merely to rank individuals on whatever amalgam of latent variables (that is, factors and uniqueness) are responsible for the total variance in the scores; neither does the test score variance itself have a true metric. So a test score of any kind should be thought of only as a *rank-order correlate* of some latent variable.

Psychometrics is by no means completely stymied by the limitations of ordinal measurement, however. Ordinal information still can be highly useful in scientific inference and for practical prediction. But scores that represent only an ordinal scale are meaningless outside the context of a clearly defined reference group. A particular test score is meaningful or useful only in terms of where it stands in the total distribution of scores in some reference population. Hence, standardized test scores are often referred to as *norm-referenced.* As explained earlier, although the scores in the reference population can have only ordinal (that is, rank-order) meaning, no matter how they have been converted or transformed, they are conventionally expressed most often as percentie ranks or as one or another form of (usually normalized) standardized scores (for example, z scores, T sores, IQ), which have certain known and convenient scale properties and distribution characteristics.

Distribution of Scores on *g*-Loaded Tests. In light of the foregoing discussion, it should not be surprising that nothing of fundamental empirical or theoretical importance is revealed by the frequency distribution per se of the total scores on any psychometric test composed of items, and this is true regardless of whether we are dealing with raw scores or standardized scores or any otherwise transformed scores. Therefore, it would be trivial and pointless to review the empirical test literature regarding the form of the distribution of test scores of any kind.

In a given population, the form of the distribution of raw scores (that is, number of items passed) is entirely a function of three interrelated item characteristics: (1) the average probability of getting the correct answer purely by chance, or guessing, (2) the average level of difficulty of the items (as indexed by the percentage of the population that fails them), and (3) the average correlation between items. Item difficulty is completely under the test constructor's control. Score increments due to chance guessing are a function of the number and quality of the alternatives in multiple-choice items and the nature of the instructions to subjects regarding the penalty for guessing at the answer instead of omitting response when uncertain (for example, total score based on number of right minus number of wrong answers). The item intercorrelations can be controlled to a considerable degree (but never completely) through item selection. Hence, in constructing a test it is possible, within broad limits, to produce almost any desired form of frequency distribution of the raw scores in a given population.

But if we have no basis for arguing that the obtained scores have true measurement properties, in addition to merely having a rank-order correlation with the latent trait that they measure—and this seems to be typically the case for psychometric test scores—the precise form of the obtained score distribution is essentially arbitrary. The very most that we can say in this case is that (within the limits of measurement error) our test scores have some monotonic relation to whatever the test really "measures." If only we could truly measure whatever latent variable accounts for the variation in the obtained scores on an absolute scale (that is, one having a true zero and additivity of scale intervals), the form of its population distribution could turn out to be quite different from that of the test scores we have actually obtained.

But certain forms of distribution are simply more useful than others, psychometrically and statistically, and it is this consideration that mainly determines the form of the distribution test constructors decide to adopt. The aims of maximizing the statistical discriminability of scores throughout a fairly wide range of talent and of obtaining a fair degree of internal consistency reliability (that is, inter-item correlation) are what largely dictate item selection. The test scores that result under these conditions of item selection typically (and necessarily) have a symmetrical and more or less "bell-shaped" frequency distribution. It is not truly the normal (or Gaussian) curve, although it usually resembles it. By juggling item characteristics, the test constructor can get a distribution that reasonably approximates the normal curve, or the scores can simply be transformed mathematically to approximate a normal distribution. (Such "normalized" scores are gotten by converting the raw scores to ranks, then converting these to percentile ranks, and then, by reference to a table of the areas under the normal curve, converting these to normal deviates, or normalized z scores.) The reason for thus normalizing a score distribution is not mainly theoretical, but statistical. The normal curve has certain uniform mathematical properties that make it extremely useful in statistical analysis and interpretation.

The argument is often made on theoretical grounds, however, that the main latent trait reflected by most complex cognitive tests, namely *g*, should be normally distributed in the general population. This argument, if accepted, justifies and indeed demands that IQs (or any other type of scores on any highly *g*-loaded tests) should be purposely scaled so that the form of their population distribution closely approximates the normal distribution. What can be said for this argument? It is rather disappointing. There are three main facets:

First, there is the argument by default: Unless there is some compelling reason to suppose that the form of the distribution of *g* is something other than normal, we might as well assume that it is normal—convenient statistically, but not very satisfying scientifically.

Second, there is the argument from the *Central-Limit Theorem* in mathematical statistics, which essentially states that the distribution of a composite variable representing the additive effects of a number of independent elements (components, causes, or influences) rapidly approaches the normal distribution as the number of elements increases. This should be the case for *g*, to the extent that we can argue on various theoretical and empirical grounds that individual differences in *g* are the result of a great many different additive effects—individual differences in the efficiency of a number of different cognitive processes, each of which is somewhat independently conditioned by polygenic inheritance interacting with a multitude of different environmental influences encountered throughout the course of development since the moment of conception. The population distribution of any variable with such multiple additive determinants, theoretically, should approximate the normal curve.

Third, there is the argument by analogy with human characteristics that actually can be measured on an absolute scale, such as height, brain weight, neural conduction velocity, sensory acuity, choice reaction time, and digit span memory (the number of digits that can be recalled entirely correctly after one presentation on 50 percent of the trials). We may reasonably presume that individual differences in each of these variables has multiple determinants, just as in the case of *g*. Indeed, we find that in very large samples of the general population the distribution of each of these variables (measured on an absolute scale) approximates the normal curve. Marked deviations from the normal curve usually occur in the regions beyond 2.5 or more standard deviations from the median of the distribution. These deviations can usually be explained in terms of certain rare extraneous genetic or environmental factors that completely override the multiple normal determinants of variation. This line of argument by analogy makes it quite plausible, but cannot prove, that *g* (or other complexly determined traits) is normally distributed. Also, the argument by analogy is weakened by the fact that not all complexly determined biological variables that can be measured on an absolute scale necessarily conform to the normal distribution. Age at death (beyond 5 years of age), for example, has a very negatively skewed distribution, because the mode is close to 75 years and the highest known limit of human longevity is about 113 years. (Below age 5, the age of death is distributed as a so-called *J* curve, with the mode immediately after birth.)

So probably the best answer we are able to give at present to questions concerning the distribution of *g* is that we do not really know the answer and cannot know it by any currently available means. But there is no good reason for not assuming that the distribution of *g* is approximately normal, at least within the middle range of about five standard deviations. Most psychometricians implicitly work on this assumption, and no argument has come forth that it adversely affects any practical uses of *g*-loaded tests. The question is mainly of scientific interest, and a satisfactory answer to it cannot come about through improved measurement techniques per se, but will arise only as part of a comprehensive theory of the nature of *g*. If we have some theoretical conception of what the form of the distribution should be in a population with certain specified characteristics, we can use random samples from such a population to validate the scale we have devised to measure *g*. The distribution of obtained measurements should conform to the characteristics of the distribution dictated by theoretical considerations.

The *g* factor as a theoretical construct will never be measured simply and directly as we can measure height with a ruler; but a rigorously testable and empirically substantiated theory of *g* would itself dictate the form of its population distribution, and our empirical measures of *g* can then be scaled so as to yield such a distribution. Although we have some notions of the kinds of experimental research and theory development that should advance us toward this scientific goal, a discussion of it here would be an excursive sidetrack. But one of the major features of this development is the use of mental chronometry in psychometric theory and research and the use of *real time* as the fundamental scale for mental measurement.

The Distribution of Achievement. There is no clear-cut or even real distinction between ability and achievement. Both are based on performance. Performances called *ability* and performances called *achievement* both yield a number of latent variables when factor-analyzed. Some of these latent variables, or factors, most notably *g*, are common to both abilities and achievements and usually constitute a large part of their variance.

A chief difference between the measurement of *g* and of achievement is that with tests of *g* our interest is mainly in the latent trait itself and not in the particular class of test items that reflect it and serve merely as vehicles for its measurement. In achievement testing, on the other hand, we are primarily interested in generalizing about the particular class of items in the achievement test. We want to know, for example, whether Johnny or Mary can add mixed fractions or do long division involving decimals. Probably because achievement has more obvious *face validity* (or *content validity*), it is not as problematic from a measurement standpoint as *g* or other ability factors. The achievement of figuring the unit price of grocery items is much more obviously consequential than solving Raven matrices, which are utterly trivial in themselves, although the *g* they reflect probably predicts more of the variance in individual differences in a great variety of achievements than any other known latent variable.

Test constructors generally aim for the same desirable psychometric properties in achievement tests that were earlier described for ability tests, with the same consequences in terms of scale properties. Again, the manipulation of item characteristics and standardization and transformation of scores produces the same kind of symmetric, bell-shaped distribution curve in the population for which the test is intended. Except for their narrowly specialized item content, standardized achievement tests are psychometrically indistinguishable from most standardized tests of intelligence or aptitude (tests in which *g* or other broad latent traits, rather than a particular class of information content or specific skills are uppermost). So the particular scale on which various aspects of scholastic achievement is measured by the standardized tests is no more than some transformation of what is basically just an ordinal scale, and the statistical features of the score distribution in the target population are an arbitrary artifact of the particular type of transformation.

There is reason to believe, however, that achievement should not have a normal distribution. In this respect, achievement might differ profoundly from *g* or from elementary cognitive processes. Certain aspects of achievement lend themselves to measurement on a cardinal scale, that is, they can be enumerated and the numbers represent a true scale, so the form of their frequency distribution in a given population is a meaningful phenomenon. Objectively countable achievements whose frequency distributions have been subjected to detailed statistical analysis are, for example, number of publications (within a given time interval) by individual research scientists and number of patents by inventors (Shockley, 1957). The distributions of these variables turn out to be very markedly skewed, with a long upper tail. In the total range, the mode is only about 20 percent of the distance above zero, and the median and mean are only slightly higher. The same thing has been found for number of compositions by composers listed in musical encyclopedias. Interestingly, earnings (but not inherited income) show the same kind of distribution in the population.

It is found that these skewed distributions when plotted on a logarithmic scale conform to the normal curve. In other words, the distribution of countable achievements (and of earnings) is long-normal. This result would be expected mathematically if achievement were a function of a number of more elemental variables, each of which is more or less normally distributed and all of which interact with each other in a multiplicative fashion. The product of two or more normally distributed variables necessarily has a skewed distribution, the skewness increasing with the number of variables.

One can only speculate about what these component variables may be (for example, *g*, energy level and effort, learned technical skills, amount of practice or experience, work habits, persistence, opportunity, and perhaps certain personality traits such as emotional stability, self-confidence, dependability, and the like). When each of these variables acts multiplicatively with all the others, the range of individual differences on each of the component variables need not be very large for their running product to have an extremely wide range of values. A person who is only one standard deviation above the average on each of the components would come out extremely far above the average on their product. From subjective impressions this appears to be the case for achievements. The variance of achievement seems to be much greater than the variance of the basic abilities or of any one of the component traits that seem to influence achievement. But there is presently no way to test this impression rigorously, since we are essentially comparing the coefficient of variation (CV = standard deviation/mean) of some measure of achievement with that of some other variable (for example, *g*) in a given population, and the coefficient of variation is meaningless unless the variables are measured on a ratio scale. How terribly handicapped psychometrics is by its limitation to ordinal scales! From the standpoint of rigorously testable theoretical development, the lack of ratio scales for the variables of greatest interest in the study of individual differences is probably the single greatest hindrance to scientific progress. One wonders how far the physical sciences could have developed without ratio scales.

But we can measure digit span on a ratio scale and thereby are permitted to illustrate some interesting points. The experimental and correlational analysis of such tractable and adequately measurable phenomena as digit span can possibly give us a better understanding of how individual differences in comparatively elemental abilities are related to achievements involving prolonged practice and application.

In the digit span test, a person sees or hears a series of digits presented at the rate of one digit per second and then immediately recites the digits in the order of presentation. The longest series that can be recalled on 50 percent of the trials is the measure of the person's digit span, and it is a ratio scale. We generally think of digit span as an ability. It is included in some standard IQ tests (Stanford-Binet and Wechsler) and it has a modest *g* loading (about .40). The range of individual differences in digit span is quite small. In young adults, the distribution of digit span closely approximates the normal curve, and most of the population falls within the narrow range of five to nine digits (hence the phrase well-known to psychologists: the "magic number 7 plus or minus 2"). Now we tend to think of achievements as involving learning and practice. So we can convert, so to speak, the relatively raw ability of digit span into a kind of ahievement by inducing people to engage in short periods of practice on digit series of increasing length, without any specific instruction, every day for several months. Virtually everyone shows a marked increase in digit span, with practice. Also, the variance of individual differences in digit span markedly increases during the course of practice. Some otherwise unexceptional individuals end up after a few months of practice with digit spans of even seventy or more digits! That is well beyond the span displayed by most professional stage performers of mnemonic feats (Ericsson, 1988). (Whether digit span becomes more or less *g*-loaded after extended practice has not been adequately studied.)

The great increase in the variance of digit span and the far-from-perfect correlation between digit span before and after practice suggests that besides the primary cognitive processes involved in digit span prior to practice, some secondary factor (or factors) comes into play during the course of practice. The great increase in variance is consistent with a multiplier effect of a secondary factor on the primary processes involved in the initial digit span. This secondary factor might be the acquisition of a strategy, such as "chunking" digits or forming meaningful associations to aid recall. Some persons adopt more effective strategies than others. The increase in the variance of individual differences with practice is commonly observed in many tasks that do not have a physiological limit or an intrinsic performance ceiling.

It is especially noteworthy that the great increase in digit span with prolonged practice does not show the slightest transfer to any other task, including memory span for symbols other than digits. The subject's letter span, for example, is not made the least bit longer even after months of practice and dramatic improvement on digit span. This indicates that the elementary cognitive processes involved in span memory have not been affected in the least by practice and that the development of certain metaprocesses or strategies specific to digits must account for the great increase in digit span. Also, the fact that digit span gradually decreases to its original status unless one keeps on practicing suggests that something more than simply knowing a particular strategy is involved (Gates and Taylor, 1925). Its efficiency seems to depend on the degree to which it has become automatized (a gradual process) by practice, and apparently the effects of automatization must be maintained by practice. These concepts are examined in more detail in the final part of this chapter.

Correlation Between g *and Achievement*

A significant correlation between two variables indicates that they have some factors in common. The fact that highly *g*-loaded tests show substantial correlations with many criteria of achievement means that these criteria must also be *g*-loaded to some degree. The most outstanding fact is the diversity of achievement criteria that are *g*-loaded. No other factors (independent of *g*) that can be measured with psychometric tests of any kind show as large correlations with as wide a variety of achievement criteria as does *g*. This is true as an empirical generalization regardless of differing theoretical conceptions of the nature of *g*.

The evidence on this point is now so vast that detailed documentation of its totality would be quite unfeasible as well as unnecessary. There are countless studies, and most of the main findings have been summarized quite comprehensively elsewhere (Jensen, 1980, ch. 8; Linn, 1982). Quotations from two of the leaders in psychometrics—Lee Cronbach and Robert Thorndike—provide the essential conclusions. Cronbach (with Snow, 1977) has stated that "general abilities are going to correlate with any broad index of later achievement" (500). "Measures of general mental ability or scholastic aptitude or academic achievement do predict learning of new material. The correlation is often in the range of 0.40 to 0.60, about equal to that found when grade averages are the criterion. This is true even in brief learning experiments, where the outcome measures very likely have low task-to-task reliability" (498). And Thorndike (1984) has concluded the following:

> Ability tests are of practical significance to us to the extent that they make it possible to predict the events of life outside of and beyond the testing situation. We may well ask, therefore, to what extent this prediction can be made from the common general factor, of which most or all cognitive tests partake, and to what extent it depends upon abilities specific to single tests or limited to groups of similar tests. I have carried out analyses of several extensive data sets in an attempt to answer this question. These analyses have led me to the conclusion that somewhere between 80% and 90% of all the variance in educational or occupational performance that can be predicted by an extended battery of ability tests is accounted for by the common *g* factor running through all the tests in the set. More limited group or specific factors appear to add not more than another 10% or 20% to the predictable variance when data are accumulated for various school subjects, various training programs, or various jobs. Thus, the notion of general intelligence, or general level of cognitive ability, is significant not only as a theoretical construct but also as the basis of most of the prediction of educational and occupational achievements that tests make possible. (2–3)

Consistent with Thorndike's conclusion is an exceptionally large set of validity data on the U.S. Employment Service's General Aptitude Test Battery (GATB), probably the most widely used test in personnel selection. The GATB consists of eleven quite diverse paper-and-pencil and performance tests. The composite score on the three tests with the largest *g* loadings (verbal, number, and spatial) provide the *G*-score of the GATB. My analysis (Jensen, 1984) of the GATB validity data shows that the average of 537 *G* score validity coefficients for 466 different occupations (with the performance criterion based on supervisor rating or on work samples) was +.27. When the criterion of job performance was predicted by the optimally weighted composite of the GATB tests (including *G*) that best predict for a given occupation, the average multiple correlation was +.36. This is a remarkably small increment (+.09) in average validity, considering that a multiple correlation mathematically must be non-zero positive and is somewhat spuriously inflated by sampling error, or so-called capitalization on chance. When the multiple correlation is statistically corrected for this bias (that is, the so-called correction for "shrinkage"), the *G* score, on average, predicts about 80 percent of the total criterion variance predicted by the optimally weighted combination of the various GATB tests, *including* the *G* composite. It seems a sound conclusion indeed that the predictive validity of tests or test batteries depends overwhelmingly on *g*. As yet no one has found (or even proposed) any other factor or combination of factors independent of *g* that could serve as a substitute for *g* in this respect.

But we also observe that the validity coefficients for highly *g*-loaded tests reported in the literature vary over an extremely wide range, from about −.20 to about +.80. There are ten principal causes of this variation which are important to recognize. (For more detail and references on each of these points, see Jensen 1980, ch. 8.)

Sampling Error of Correlation. A validity coefficient is simply a correlation coefficient between a predictor variable (for example, test scores) and a performance criterion (school grades, job supervisor rating), and, like any other statistic, it is subject to sampling error, which is inversely related to the sample size.

Reliability of the Test. Tests vary in reliability, although most published tests have reliability coefficients that are quite high (about .80 to .95), at least in the normative population. Hence, test reliability per se is not a major source of variation in validity coefficients. The square root of a test's reliability sets the upper limit of its possible true correlation with any other variable. (A detailed exposition of reliability can be found in Jensen, 1980, ch. 7.)

Reliability of the Criterion. Criterion reliability is much more problematic, and usually much lower, than test reliability. It depends largely on the type of criterion (see *Type of Criterion* below). Unlike test reliability, criterion reliability is a major source of variability in validity coefficients.

Range of Ability. Restriction of the range of ability lowers the estimated validity of a test. Some samples in which validity coefficients are determined are highly selected on the abilities measured by the test prior to the validity study, either by self-selection for opting to take the test or by imposed requirements that screen those who are finally tested and selected into the situation that yields data for estimation of the test's validity. Applicants to Ivy League colleges, for example, are not a random sample of high school graduates but are already preselected for relatively high academic aptitude, and those who are admitted are even more highly selected.

At every rung of the educational ladder, from elementary school to graduate or professional school, there is some degree of selection and consequent restriction of the range of individual differences in *g* among those who "survive." In my use of highly *g*-loaded tests in studies over the years with some three thousand students of University of California, Berkeley, for example, I find that compared with the distribution of *g* in the general population, the distribution of Berkeley undergraduates falls almost entirely within the upper quartile, and of graduate students, within the upper quintile. These highly select samples, therefore, have less than one-fourth of the variance in *g* found in the general population. This condition quite severely restricts the size of correlation that can be obtained between a *g*-loaded test (for example, the SAT) and some criterion measure of academic performance (grade point average, or GPA). It is a general observation that the more selective the institution, the lower the validity coefficients. The highest validity coefficients are seen in colleges with open admissions; the lowest I have seen reported are found in highly selective institutions such as Caltech and MIT. Some years ago I found that the Quantitative score on the Graduate Record Examination (GRE) had near-zero validity for predicting GPA among Berkeley graduate students in mathematics, but the severely restricted variance in test scores was not significantly larger than the test's error variance! Validity coefficients of the GRE for predicting grades in graduate school are typically close to +.30 (Kyllonen, 1986, 4). Tests that are psychometrically no better show much higher validities in less restricted populations. Kyllonen (1986) has reviewed the validities obtained on large samples in as many as fifty-seven different air force training courses over a period of twenty years, in which scores on heavily *g*-loaded military selection and classification tests were correlated with technical school GPA. The median validities ranged from +.42 to +.82, with a mean of +.61. Yet the air force has the most highly selected inductees of any branch of the armed services.

Selection for lower ability, of course, has the same effect on validity as selection for higher ability. For example, the validity of the Armed Forces Qualification Test (AFQT) for predicting navy enlistees' successful progression from apprenticeship training programs to specialized technical schools was about +.60 for enlistees in all navy classification categories based on the (AFQT), but it was only about +.30 for enlistees in category IV (the tenth to thirtieth percentile on the AFQT) (Cory, Neffson, and Rimland, 1980).

Heterogeneity of Criterion. It is often the case that the criterion measure is not really the same (although it may be nominally the same) for all members of the sample in which validity is determined. College GPA is a good example of a heterogeneous criterion. It is a composite of non-equivalent components for different students. Various courses and majors clearly differ in their intellectual demands, and students are highly varied in the curricula in which they obtain their grades. It has been found on several campuses of the University of California, for example, that the validity of the SAT is higher for grades *within* courses (that is, a homogeneous criterion, where everyone is graded on the same basis) than for overall GPA, a heterogeneous criterion (Goldman et al., 1974; Goldman and Slaughter, 1976). The heterogeneity of GPA, in combination with the somewhat restricted range of *g* in most colleges, imposes a ceiling on the validity coefficients of college selection tests, which are typically in the range of .40 to .50.

Initial Selection on Negatively Correlated Criteria. This exacts a heavy toll on validity. It occurs when persons are selected on two or more different criteria which, although they may be positively correlated in the total pool of applicants, are negatively correlated in the finally selected group. While the most prestigious colleges are in a position to select only those high school graduates who have both high GPAs and high SAT scores, many colleges cannot afford to be so choosy. They must select most of their students from among applicants who are high in either GPA or SAT scores while having relatively few applicants who are equally high in both criteria. This selection procedure necessarily "builds in" a negative correlation between the personal traits that make for high GPA despite mediocre ability (largely *g*) of the kind measured by the SAT, or that make for low GPA despite high ability. The effects of these combinations of traits carry over to students' performance in college and can markedly weaken the validity of SAT scores for predicting college GPA. There have been some extreme cases where graduate students have been selected in this way, with the result that the GRE showed zero or even negative correlations with performance in graduate studies.

Type of Criterion. Various criteria of performance differ not only in reliability but in the degree to which they depend on the cognitive factors measured by psychometric tests. A great deal of evidence supports the following generalizations:

1. In general, education in academic subjects and training in technical courses make greater *g* demands than most other situations in which *g*-loaded test validity is estimated. Therefore, it is not surprising that tests show their highest validities for the prediction of scholastic and training criteria. I have elsewhere (Jensen, 1989a) reviewed evidence that the general factor in learning tasks, where the criterion is the rate or amount of acquisition of new knowledge or cognitive skills, is the same general factor *g* found in psychometric tests. Hence, performance criteria that strongly reflect the acquisition of new knowledge and skills are among the most highly *g*-loaded and the most predictable by means of psychometric

tests. The highest predictive validities are found between *g* and scores on achievement tests. This is true even when the predictor test is highly *g*-loaded but does not contain any information content in common with the achievement test. Tests given prior to a specific course of training, before students have any knowledge whatsoever of the subject matter, can predict the final level of achievement in the course, with correlations as high as .70. The predictive validity is generally highest when all students have had the same instruction and the same amount of study time.

2. Validity is usually much higher when the achievement criterion is scores on an objective achievement test rather than teacher ratings or course grades. But even achievement tests differ in *g* loading, depending on whether they measure primarily the knowledge content of the course or measure primarily the use of this knowledge in making inferences, interpretations, or solving novel problems. Performance on the latter type of achievement tests is more predictable from *g*-loaded tests.

The higher predictive validity when the criterion is a paper-and-pencil achievement test is attributable in some part to what is termed *common method variance*. This means that the predictor and the criterion measures are based on highly similar procedures (for example, a multiple-choice format with separate machine-scored answer sheets) which themselves have little or no intrinsic relationship to the abilities or achievements being measured. So persons who are actually equal in achievement may differ in so-called test-wiseness, or familiarity with a particular type of test format. The importance of this source of variance in test scores is less, the more that test takers have been previously exposed to objective tests of various kinds throughout their schooling. For groups that have had such experience in taking tests, the gains resulting from special coaching and further practice in test-taking skills are seldom larger than the test's standard error of measurement (Cole, 1982; Jensen, 1980, 589–596).

Grades generally have lower validity than objective measures of achievement for four main reasons: (1) grades have lower reliability, (2) the grading scale is more coarsely graded than objective measurement scales, (3) grades usually reflect relative standing in a given class, and classes may differ considerably in average level of ability and achievement, and (4) the grades teachers give often reflect their feelings about pupils' personal traits (such as obedience, conscientiousness, effort, forthcomingness, neatness of written work, and the like), which have little or no correlation with either *g* or achievement. It has long been noted, for example, that girls get higher grades than boys in school—a difference not reflected in scores on objective tests of achievement.

Even when grades are averaged over a number of years, so that different teachers' idiosyncracies in grading are averaged out, the correlation between grades and *g* is far from perfect. A strong test of the overall relationship between *g* and grades was made by Gedye (1981), working with the longitudinal data of the Berkeley Growth Study. She extracted a general factor (and individual factor scores) from pupils' teacher-assigned grades in arithmetic, English, and social studies obtained in all grades one through ten. She also extracted a general factor (and factor scores) from the Stanford-Binet IQs obtained on the same pupils on six occasions between grades one and ten; so this is a rather ideal measure of *g*. The correlation between the general factor score for grades and the Stanford-Binet *g* is +.69. Corrected for attenuation (unreliability), the correlation is +.75. The fact that the corrected correlation is not higher indicates that school grades in academic subjects, although highly correlated with *g*, also reflect consistent sources of variance that are completely independent of *g*. The difficulty in studying this non-*g* variance in grades is that it seems to be attributable to a great many small (but relatively stable) sources (personality traits, idiosyncratic traits, study habits, interests, drive, and so on) rather than to just a few large and measurable traits. That is why attempts to improve prediction by including personality measurements along with cognitive tests have not shown much promise. There is no general factor (or even several broad group factors) in the noncognitive realm which, combined with *g*, would appreciably enhance predictive validity.

3. In personnel selection, *g*-loaded tests have much higher predictive validity when the criterion is a test of *job knowledge* than when the criterion is *supervisor ratings*. Probably little of this difference in validity coefficients is attributable to common method variance, that is, the fact that both the predictor and the criterion variables are measured by paper-and-pencil tests. Scores on a job-knowledge test obtained after employees have spent several months or more on the job reflect the amount of job-relevant information acquired through intentional and incidental learning while on the job. And the rate of acquisition of *declarative knowledge* is quite highly related to *g*. That is, persons with higher levels of *g* generally acquire, per unit of time, more information (especially of the kind that can be verbally articulated) from their experiences than do persons with lower levels of *g*.

Job knowledge, of course, is important to the extent that it is related to employees' actual proficiency on the job, and jobs differ considerably in the extent to which declarative knowledge plays a part. Specialized knowledge is the sine qua non of some jobs. In others, it is almost superfluous beyond some rather mediocre level. A specialist in some branch of experimental physics who is brought in as a consultant to advise a team of researchers in, say, the Lawrence Berkeley Laboratory, is sought expressly for her or his exceptional fund of specialized knowledge and problem-solving expertise, without which he or she would be of no value as a consultant. However, a gardener working on the Berkeley campus, with its rich variety of flora, might acquire an encyclopedic knowledge of botanical taxonomy and horticultural science, although such knowledge would not be essential for performing the gardening chores.

Not surprisingly, supervisor ratings are more highly correlated with job knowledge than with *g*. But supervisor ratings are a problematic criterion because they reflect all the factors that influence person perception, and the relevance of these factors varies greatly from one job to another. Factor analyses of a variety of supervisor ratings along with various cognitive tests, including job knowledge, show that ratings contain reliable (and, for some jobs, quite valid) components of variance that are entirely independent of the variance attributable to *g* and job knowledge (Campbell et al., 1973).

In general, however, job-knowledge tests are more highly correlated with actual proficiency on the job, as measured by objective work samples, than are supervisor ratings. There is no reason to beleive that whatever favorable personal qualities that might enhance effectiveness on the job and are reflected in supervisor ratings would be negatively correlated with either *g* or job knowledge. What evidence I have found, in fact, indicates a slight positive correlation between noncognitive personal qualities (as rated by supervisors) and job knowledge (Campbell et al., 1973).

Although *g* is most highly correlated with job knowledge, it is important to note that large-scale meta-analyses of both civilian and military personnel data show that *g* contributes to variance in actual job performance (assessed from work samples) independently of job knowledge. That is, even when workers are statistically equated on job knowledge, *g* is still significantly correlated with job performance (Hunter, 1986). As Hunter explains, "Ability [*g*] predicts performance above and beyond its prediction of job knowledge because it measures the ability to innovate and prioritize in dealing with situations that deviate from those encountered in prior training" (358).

Prediction between versus within Occupations. Here I will briefly summarize points I have fully documented elsewhere (Jensen, 1980, chap. 8). It has often been noted in reviews of test validities in personnel selection that validity coefficients for predicting job performance within specific occupations are rather disconcertingly low, for the most part in the range of .20 to .30. The reason usually given for this is restriction of range of ability within occupations. But this is only one factor in a quite complicated picture and probably a minor one at that. It so happens that predictive validities of *g*-loaded tests are actually somewhat higher in occupations that have a more restricted range of ability than in occupations with a very wide range of ability. (The reason for this seeming paradox is explained in *Job Complexity and* g *Validity* below.) When we analyze *g*-loaded test scores from persons in a very large number and extremely wide variety of occupations, we find that approximately one-half of the total variance in scores exists between the means of the various occupations and approximately one-half of the total variance exists within occupations (that is, individual differences among persons within any given occupation). From this empirical observation, it follows statistically that if we rank-order occupations so as to maximize the correlation between their ranks and their mean scores on *g*-loaded tests, the correlation between individuals' test scores and their occupational ranks would be the square root of one-half, or approximately +.70. In other words, *g* predicts occupational status with a validity coefficient of +.70. And this degree of correlation is just what is actually found in studies in which many different occupations have been ranked, not on any psychometric criteria, but in terms of their prestige (in the eyes of the subjects who do the ranking), their desirability, and people's subjective judgments of the amount of intelligence they think is required for successful performance in the occupation. (These three criteria, when based on the pooled ranks by a large number of persons, are amazingly consistent with one another and are highly stable throughout the industrialized world and from one decade to another.) It seems impossible to avoid the conclusion that what people ordinarily mean by "occupational status" is quite highly related to psychometric *g*.

On the other hand, the observed correlations between *g* and measures of proficiency within given occupations are usually very far below +.70, even though one-half the total *g* variance exists within occupations. This means that, in general, *g* is much less able to predict occupational *performance* than occupational *level*. The main reason (aside from the forms of attenuation of the validity coefficient previously mentioned) is that, once employees are up to the minimum level of qualification for performing in a given occupation, a host of other factors independent of *g* becomes at least as important as *g* for successful job performance (or the perception of effectiveness by supervisors and co-workers). Most nominal occupational categories accommodate a surprisingly wide range of *g* above the minimum level, or *threshold*, for a given occupation. This threshold level can be estimated from the mean *g*-loaded test scores of persons in a given occupation whose scores are at the first percentile of the distribution of scores in that occupation. The *g* "threshold levels" across a wide variety of occupations vary considerably more than do the mean levels of *g* across occupations, and the very top levels of *g* across occupations show surprisingly little variation—only about one-seventh as much (in IQ units) as we see at the threshold level. Some very high-*g* persons are found in some very low-*g* occupations, but no very low-*g* persons are found in high-*g* occupations. (The evidence is reviewed by Jensen, 1980, 343–45.) This widely recognized threshold property of *g*, with respect to both education and occupations, is probably responsible in large part for people's anxiety and antipathy concerning tests of mental ability.

Job Complexity and *g* Validity. If we could factor-analyze a great variety of occupations the way we can factor-analyze tests, we would find that occupations differ in their *g* loadings, which we could think of as the occupations' *g demands*. Hence the predictive validity of *g*-loaded tests for all types of job performance criteria improves as a function of the job's *g* demands. This is mainly characterized by the complexity of the job (Hunter, 1986, 344–45). Job complexity is related to the degree to which successful performance depends on both declarative and procedural knowledge, the making of fine discriminations, decisions, judgments, thinking, problem solving (especially the transfer of already acquired expertise to novel problems), and continual study and learning in order to keep up. Cosmology would seem to be more complex in this sense and to make greater *g* demands than cosmetology, for example.

Even at the very simplest levels of performance, such as the difference between simple reaction time (that is, RT for response to one signal) and choice RT (a response to only one of two possible signals), psychometric *g* is slightly but significantly more highly correlated (negatively in the case of RT) with the slightly more complex task. I have discovered this phenomenon within every segment of the total distribution of ability—the mentally retarded, average schoolchildren, super-gifted children attending university at age twelve, unskilled factory workers, navy recruits, vocational college students, university students, and members of Mensa (Jensen, 1982, 1987d).

For many jobs at the very lowest level of *g*-demands and the least personal responsibility in the whole occupational hierarchy, one finds test validities of zero, or even *negative* validity coefficients as large as −.20 or so. Such validity coefficients (usually based on supervisor ratings) do not reflect actual job proficiency per se so much as they reflect other behaviors that employers consider desirable, such as stable personality, dependability, interest, job satisfaction, low absenteeism, steadiness on the job, good attitude, and duration of employment. Employers have a usually well-founded reluctance to hire persons they regard as over-qualified for a particular job. Among employees in jobs with exceedingly small *g*-demands and little responsibility in terms of decisions or supervision, there is a slight negative correlation between *g* and other desirable personal traits. The employee turnover rate in low-*g* jobs, for example, increases as the employees' ability exceeds the minimum level of ability needed to do the job.

Length of Experience on the Job. The importance of the predictive validity of *g* for job performance would be fortunately lessened if it were found that the correlation between employees' *g*-loaded test scores and their quality of job performance steadily diminished with their length of experience on the job. This possibility—that differences in job performance between workers of high general ability and workers of low ability tend to fade the longer they remain on the job and gain more experience—is known as the *convergence hypothesis*. A recent large study that tested this hypothesis failed to detect significant convergence of high- and low-ability groups as a function of time spent on the job. Schmidt (1988) summarized the results as follows:

> We found that for all three measures of job performance—work sample tests, job knowledge tests, and supervisory ratings—the difference in performance between the top and bottom ability halves remained constant out to at least five years on the job. (Beyond that point the data were inadequate to draw conclusions.) Thus our findings disconfirmed convergence theory. It appears that initial ability differences produce job performance differences that are quite lasting. (286)

There is much research evidence to show that tasks with a degree of complexity that requires continual information-processing and on which performance cannot be completely routinized or automatized will continue to correlate with *g* indefinitely over time (Ackerman, 1987). It is questionable if even the most menial jobs can become so wholly routinized through practice that indexes of performance would cease to reflect individual differences in *g*.

Psychometric Aspects of Mitigating *g* Differences

The reality of *g*, the fact that it can be reliably measured and has useful predictive validity in education and employment, is overwhelmingly substantiated by psychometric research.

Furthermore, it is impossible to ignore the wide range of individual differences in *g* and the statistical differences between racial groups in the distribution of *g*. These are facts, regardless of controversy concerning the causes of individual and group differences. Causal questions need not be considered here. The immediate fact, on which there is a general consensus, is that *g* differences, regardless of their cause, have obviously important consequences for education, for employment, and for the quality of life.

Adverse Impact

Psychometric tests themselves are not responsible for creating the observed differences but are simply a more precise and standardized means of identifying and measuring behavioral differences that have been observed informally throughout human history. But our society's increasing use of tests, especially in educational and employment selection, has highlighted the phenomenon known as *adverse impact.*

Adverse impact refers to the disproportionate selection of applicants from groups that differ statistically in the characteristic measured by the selection test. In the simplest selection procedure, those individuals whose test scores fall above a given *cut-score* are the applicants who are finally chosen, regardless of their group membership.

When the means of two groups with overlapping score distributions differ, selection based on a common cut-shore necessarily results in proportionally fewer persons being chosen from the distribution with the lower mean. The higher the cut-score, the larger is the disparity between the proportions of the two groups that are "favored" by the selection procedure.

An index of adverse impact is the ratio of the proportions selected from the higher and lower groups. A ratio of 1:1 would indicate the complete absence of adverse impact. If the distribution of scores is approximately normal in each group, one can estimate with fair precision the index of adverse impact for any given cut-score from a knowledge of the group means and standard deviations (SDs). For example, if the means of two normal distributions differ by as much as one (within-group) SD and the cut-score is at the mean of the higher group, the index of adverse impact is 3:1. If the cut-score is moved up to one SD above the mean of the higher group, the index is 7:1.

In an applicant pool that includes representative samples of the black and white populations, the degree of adverse impact (for blacks) of any given cut-score depends on the factor composition of the test. In accord with Spearman's hypothesis, the larger the test's *g* loading, the greater will be the adverse impact. (Also, inclusion of a spatial-mechanical factor increases adverse impact for blacks, on average.) Attempts to reduce adverse impact, either by reducing a test's *g* loading, or by directly minimizing the group difference by means of item selection techniques expressly aimed at fulfilling this purpose, have been found so greatly to impair the test's validity within either group as to make such a test practically useless (Jensen, 1980, chaps. 11 and 14). Few psychometricians any longer consider this a promising solution.

Given the fact of adverse impact, the next question should be. Are there strictly psychometric and statistical aspects of the phenomenon that, if viewed properly, could permit us to lessen its severity?

Blind Alleys. Before considering the above question in a positive light, I should first explicitly dismiss those commonplace reactions to the problem that, in my judgement, the evidence indicates are conclusively unpromising.

1. The popular claim of *test bias* now carries no weight in the sense that if we could get rid of biased tests and substitute perfectly unbiased tests, the problem of adverse impact would be removed or diminished. It is simply paranoia to believe that psychologists, from the time of Binet to the present, have had a vested interest in producing or using biased tests. In recent years, psychometric researchers, test publishers, and the armed services have worked assiduously at devising methods for detecting and eliminating cultural biases from tests. The expert consensus is that these efforts have largely succeeded. (For an introduction to this now vast literature, see Arvey, 1979; Jensen, 1980; Reynolds and Brown, 1984; Wigdor and Garner, 1982.)

What little bias may exist in some few modern tests is generally so small and inconsistent in direction that its complete elimination would have a negligible effect on adverse impact. More often than not, in fact, the complete elimination of bias would have the effect of increasing the degree of adverse impact on blacks. It is a sound empirical generalization that most tests currently used in education and employment have useful validity in virtually all American-born and American-educated groups in our population. From two articles that contain excellent reviews of the evidence, here are the summarizing statements regarding the research on test bias by Robert Linn, a leading psychometrician, and Frank Schmidt, a leading personnel psychologist:

> Whether the criterion to be predicted is freshman GPA in college, first year grades in law school, outcomes of job training, or job performance measures,

> carefully chosen ability tests have *not* been found to under-predict the actual performance of minority group persons. Contrary to what is often presupposed, the bulk of the evidence shows either that there are essentially no differences in predictions based on majority or minority group data, or that the predictions based on majority group data give some advantage to minority group members. (Linn, 1982, 384)

> We now know not only that cognitive employment tests are equally valid and predictively fair for minorities, but also (1) that they are valid for virtually all jobs, and (2) that failure to use them in selection will typically result in substantial economic loss to individual organizations and the economy as a whole. (Schmidt, 1988, 281)

2. The idea of *banning* the use of tests in the private sector would be completely unrealistic without also contemplating drastic changes in the laws that regulate private enterprise. If employers find that the economic benefit of using tests in employee selection exceeds their cost, it is predictable that they will use tests. The government could exercise its power to ban the use of tests in its own agencies, but it is these very agencies, particularly the civil service and the armed forces, that best appreciate the economics of testing. Tests are used in personnel selection, training assignments, and promotions because the real economic advantages calculated under a wide range of various reasonable assumptions substantially exceed the costs of developing and using appropriate tests. For officials or taxpayers to surrender these advantages, alternatives to testing would be required for selection. So far, no alternative has been suggested that promises to be at least equal to tests in cost-effectiveness or as meritocratic for all classes of applicants.

3. *Depreciation of tests' validity coefficients* by squaring them (thereby making them seem much smaller) is a common but technically improper and misleading way of belittling the value of tests for selection. A validity coefficient, of course, is just a correlation coefficient, and the most predictable conditioned reflex among psychologists (and social scientists generally) is to square every correlation coefficient they see. We all have learned (quite correctly) that the squared correlation coefficient (termed the *coefficient of determination*) indicates the proportion of *variance* in variable x that can be "accounted for" (or "explained" or "predicted" or "attributed to") by its linear regression on variable y. But this is a quite misleading and scarcely useful interpretation of a validity coefficient.

The proper interpretation of a test's validity, as originally shown in a now classic paper by Brogden (1946), is that the validity coefficient itself is the average proportional gain in the criterion performance that results from the use of the test for selection. For any given selection procedure, and assuming the nature of the criterion is fixed, the selection test's validity coefficient itself is a direct measure of the average improvement in criterion performance (that is, quality of work, worker productivity, and so on) on a ratio scale of 0 to 1, where 0 represents the average performance if the same number of persons had been selected at random from the same pool of applicants, and 1 represents the average performance if the same number of persons had been selected from the same pool of applicants by means of a hypothetically perfect predictor, that is, a test with a validity coefficient = 1. These relationships can be most easily understood in terms of Brogden's formula:

$$r_{xc} = (\mathbf{T} - \mathbf{R})/(\mathbf{P} - \mathbf{R}),$$

where r_{xc} is the tests' validity coefficient (the correlation r between the test scores x and the criterion measures c), **T** is the average performance of persons selected with the test, **R** is the average performance of persons selected at random, and **P** is the average performance of perfectly selected persons, as if $r_{xc} = 1$.

In light of this generally accepted meaning of the validity coefficient, the usefulness of even a quite low validity coefficient (.20 to .30) cannot be regarded as trivial in many situations where efficiency of training (or low failure rates), or competence, quality of work, and productivity are considered especially important, in terms of cost, or safety, or urgency of time, or competition in achieving a goal. However, a test's actual *utility* in any particular situation also depends on other factors besides its validity.

Utility versus Validity

Utility and validity are clearly related, but they are importantly different concepts. In some cases, an argument can be made against the use of a selection test on the grounds of its utility, even when there can be no argument about its validity, which may be commendable. But validity is only one of several elements that determine utility, which is a more complex concept than validity. Although validity is essential for utility, from a practical standpoint, utility is the more crucial.

Utility (a term borrowed from the concept of "marginal utility" in economics) is a function of four independent elements: (1) *validity*, (2) *base rate*, (3) *selection ratio*, and (4) the *cost-effectiveness of alternative methods*. Validity has already been defined above.

The *base rate* is some indicator of the quality of the total applicant pool with respect to the criterion, such as the proportion of all applicants who are capable of satisfactory performance on the criterion, or more precisely, the (hypothetical) mean and SD of all applicants on the criterion. The base rate is determined by whatever conditions "pre-select" those who enter the applicant pool—how potential applicants were informed, requirements of age, education, or experience, and the many personal factors that influence self-selection.

The *selection ratio* is the proportion of the available applicants that are selected. It is determined by the test's cut-score, which is governed by supply and demand or by some required absolute standard of performance.

The higher the base rate and the higher the selection ratio, the lower is the test's utility for any given level of validity. If all applicants were selected, obviously the test's utility would be zero, regardless of its validity. The more severe the selection (that is, the lower the selection ratio), the greater is the utility. For example, given a validity of .50 and a selection ratio of .05, the selectees, on average, can be expected to perform on the criterion 1.04 SD above the mean of applicants selected at random.

How much practical difference 1.04 SD makes with respect to some absolute standard of performance will depend largely on the base rate. If it is quite high, then despite a range of individual differences, nearly all applicants would perform quite satisfactorily, and the practical advantage of using tests for selection might be only a trivial improvement over random selection.

This is where consideration of the time and money cost of testing must be weighed against the performance gain, or benefit. Also, less costly selection procedures with comparable validity may be readily available. Among category IV navy personnel (tenth to thirtieth percentile on the AFQT), for example, it has been found that a few easily obtained items of biographical information (for example, high school graduate versus dropout) actually have higher validity than psychometric tests for predicting job performance (Cory et al., 1980).

Test utility falls drastically when the size of the self-selected applicant pool is very limited, as is the case for many technical jobs. Cronbach and Gleser (1965) claim that if a test is worth using at all, at least twice as many applicants should be tested as will be selected. With a test validity of .50 (which is relatively high) and a selection ratio of .50 (the maximum recommended by Cronbach and Gleser), the selectees would be expected to have an average level of performance on the criterion about 0.4 SD above a randomly selected group. This is not negligible. But many colleges and many employers cannot afford to be as selective as Cronbach and Gleser recommend. Yet, as the selection ratio increases above .50, the utility of testing plunges markedly.

A recent example of the application of utility concepts is an elaborately researched argument for dropping the use of the SAT in the college admissions process (Crouse and Trusheim, 1988). One of the few technically cogent critiques of the SAT, its argument is based entirely on the test's utility. The authors state,

> Unlike many critics, we do not question ETS's claim that the SAT measures important abilities that are related to educational and economic success. Rather, we argue that despite its ability to predict educational success, the SAT is unnecessary. This apparent paradox disappears when one recognizes that even when a test predicts college success fairly accurately, it may not *improve* prediction much when used to supplement information available from high schools about students' coursework and grades. Our argument develops the case against the SAT as a tool in college admissions, not against the test's validity in measuring individual differences important to educational success. (xii–xiii)

Furthermore, their analysis shows that the use of the SAT in college admissions increases adverse impact for blacks over and above what it would be with the use of high school grades alone. But no other broad generalizations on this topic would be feasible here, because the degree of adverse impact due to the SAT as compared to high school grades alone results from the complex interaction of many factors that vary widely across high schools and colleges. Detailed explications are provided by Crouse and Trusheim (1988, ch. 5) and by Gottfredson and Crouse (1986).

The kind of examination from a utility standpoint that Crouse and Trusheim applied to the SAT should be applied to current uses of other tests in other settings.

Nonlinearity of the Test/Criterion Relationship

Imagine this situation: As test scores increase by equal intervals, the criterion measure (grades, job performance, and so on) increases by ever *decreasing* increments; that is, the monotonically positive relation of the criterion to the test scores is a negatively accelerated curve. Hence, for test scores that range below a given cut-score, unit differences between scores would correspond to larger differences in criterion performance than would unit difference between scores that range above the cut-score.

The question then is: Does this type of nonlinear test/criterion function offer the possibility that an optimally placed cut-score would allow strictly *random selection* from the pool of all applicants whose test scores range above the cut-score without sacrificing an acceptable level of utility?

The stated premise makes the question merely rhetorical. The answer has to be *yes*, provided the cut-score is high enough that over the range of scores lying above it the corresponding increments of criterion gain are too small to be of practical consequence. From the standpoint of reducing adverse impact, however, this hypothetical possibility looks extremely unpromising when examined in the light of empirical realities.

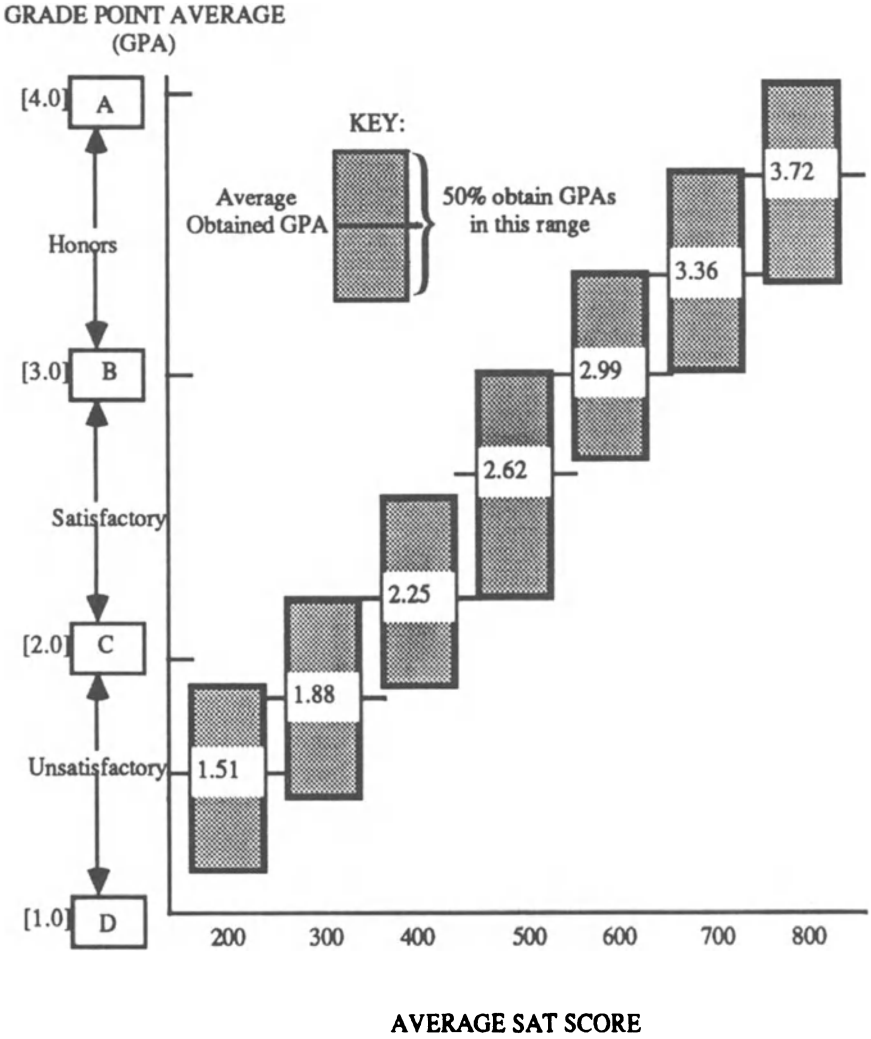
The problem has three main aspects: (1) the utility of *random selection* above a given cut-score as compared with *top-down selection* (selecting consecutively from the top score on down in the whole pool of tested applicants until the required number of selectees is obtained); (2) the placement of the cut-score so as to appreciably reduce adverse impact; and (3) the prevalence and consequences of a nonlinear relationship between test scores and criterion.

1. Provided there is a monotonic relationship between the predictor and criterion variables, then for selecting a given number of individuals from among a larger number of applicants, top-down selection has by far the greatest utility of any selection procedure that uses tests. It is always superior to random selection above a given cut-score unless the cut-score is so high that mostly the same individuals are selected by random selection above the cut-score as would be selected by the top-down procedure.

2. The problem with straight top-down selection, however, is that it has the greatest adverse impact of any selection procedure. The argument for random selection is simply the desire to reduce adverse impact. This could be accomplished to a socially significant degree only by setting a very low cut-score. With a very low cut-score, however, the utility of the test is drastically reduced. The typical result is that the random selection model does not select the ablest applicants from either population. Moreover, it has the added disadvantage (as compared with every other selection model) of maximizing the mean group difference among the selectees—an effect that only postpones adverse impact to subsequent decisions about retention and advancement. Also, the fact that there is usually a much higher proportion of majority than of minority persons in the applicant pool means that random selection above a very low cut-score will yield a relatively small number of minority selectees at the expense of getting a large number of substandard majority selectees.

3. Any form of monotonic relationship between predictor and criterion ensures some degree of validity. A linear relationship is not crucial, although it is convenient, because the validity coefficient reflects only the degree of prediction that is made possible by the linear regression of the criterion on the predictor. Nonlinear regression simply lowers the validity coefficient. But this in itself is a trivial problem. In the first place, for

every form of monotonic relationship (linear or nonlinear), top-down selection yields the same result. Moreover, linearity is by far the prevalent condition empirically, especially when criteria of a cognitive nature are regressed on highly *g*-loaded tests. The regression of college GPA on SAT scores is an example of almost perfect linearity, as shown in figure



Source: Manning and Jackson 1984.

Figure 4–1. Average College Grades for Students with Different SAT Scores

4–1. Such linearity comes about both "naturally" and by design. When tests are specially designed to have maximum discriminability in the range of ability most relevant to a particular criterion, the regression line is seldom significantly nonlinear. My previous search (Jensen, 1980, chap. 8) for a nonlinear relationship between scores on highly *g*-loaded tests and various achievement measures found no authentic evidence for nonlinearity. Extreme skewness of the distributions caused by "basement" and "ceiling" artifacts in the test scores or criterion measures usually account for the rare instances of significant deviation from linearity.

Even if nonlinearity exists in a particular case, it can be handled effectively in either of two ways: (1) if the only purpose is to improve prediction, then there is nothing sacrosanct about the particular scale of test scores or criterion measures. We are free to subject either one or both of them to whatever monotonic transformation comes closest to producing linear regression and maximizing the validity coefficient; and (2) if we suspect nonlinearity, we can simply enter each score (x) and one or more of its higher powers (x^2, x^3, and so on) into a multiple regression equation; the multiple correlation coefficient, then, is the test's validity coefficient, which reflects the predictive power made possible by the linear component along with the significant nonlinear components of the relationship between scores and criterion. Hence, there is no necessary loss in validity or utility as a result of a nonlinear relationship between criterion and predictor.

The question of whether individual differences in the criterion performance for all applicants, who are randomly selected from the range above a given cut-score on the selection test, are too trivial to matter can easily be answered by comparing the mean and SD of their criterion performance with that of exactly the same number of applicants selected from the same pool by the top-down method. In any particular situation, evaluation of the difference that is found, of course, would depend on the consideration of other factors as well.

An important consideration in this situation is the *standard error of estimate* (SEest), which is the overall average standard deviation of the dispersion of the criterion measure around the prediction line defined by its regression on the test score. (It should be noted that the standard error in estimating the criterion measurement for an *individual* is not the same as the SEest but may be considerably larger; the proper formulas for each type of error are given in many statistics textbooks and in Jensen, 1980, 379.)

Although the SEest has a perfect inverse relationship to the validity coefficient, it is often the case that the dispersion of the criterion measures does not show the property that statisticians refer to as *homoscedasticity* (that is, uniform dispersion of the criterion measure around the regression line throughout the full range of test scores; for example, going back to figure 4−1, the dispersion of GPA appears fairly homoscedastic). The dispersion of the criterion measure could be much greater in some part of the test score range than is indicated by the SEest, in which case the test would have little practical discriminability in that range.

So if, within the truncated range above some cut-score on the full range of scores, the average criterion difference between individuals in the bottom and in the top 10 percent of the truncated range is not statistically significant (say, at the 10 percent level of confidence), very little of the test's potential utility under top-down selection would bė lost by the random selection of applicants ranging above the cut-score. And if the cut-score were not too high, there would be an appreciable reduction of adverse impact.

Temporal Loss of Predictive Validity

In many situations, the predictive validity of test scores diminishes as the length of time between taking the test and measuring the criterion increases. This effect is clearly seen in the prediction of college GPA by the SAT or other high-*g* tests. Humphreys (1968), for example, obtained a composite score based on several tests given to students on admission to their freshman year at the University of Illinois, and he correlated this composite score with GPA obtained within each of the eight semesters between the freshman and senior years. In one analysis only the sixteen hundred students who had progressed through al eight semesters and graduated were used, so there would be no change in the range of talent over the eight semesters. For this group, the validity coefficients for predicting GPA showed a fairly gradual decline (from +.375 to +.173) between the first and the eighth semester. (Corrected for restriction of range compared to all entering freshmen, the validities over the eight semesters declined from +.47 to +.21.) An almost identical picture is seen when high school GPA was used as the predictor.

The largest part of the observed decline in predictive validity is most likely due to the increasing heterogeneity of the criterion (GPA) as students adjust their course loads, select their courses, shift their majors, and the like, to maintain satisfactory grades. (Not all of the decline in predictive validity is due to the nature of the GPA, however, because test-retest correlations also decline as a function of the interval between test and retest.) But the important point for our present concern is that a stringent selection policy (that is, a high cut-score on the predictor) eliminates some students who initially look unpromising but who would eventually succeed if given the chance. Hence, lenient selection and retention on probationary status for those who fail at an early stage would increase the number that finally succeed academically and are legitimately able to graduate. Humphreys (1968) states, "Senior performance is not predicted well enough from freshman information for one to be at all content with present college admission practices" (378). He recommends, "Perhaps admission tests should be validated primarily against staying in college versus dropping out" (378). Most selection tests are validated only against freshman GPA, but as Humphreys notes, "A good many students who are dropped at the end of first semester would do acceptable work later in college...one simply cannot predict well enough from freshman academic deficiency to senior performance" (378–379).

These conclusions, however, are liable to overly optimistic expectations of the probable consequences of very lenient admission standards. A later study by Humphreys (1973), based on the same data as the previous study (that is, Humphreys, 1968), somewhat dampens the rather common hope that the lowest level of ability (as indicated by the predictor variable) found among those who succeed on the criterion (when there has been no formal prior selection) should determine the level of the cut-score on the predictor variable for screening future applicants. In other words, if we find that there are some self-selected persons who, despite having quite unpromising scores on the predictor, have managed to succeed on the criterion, why should we not recruit and accept applicants at (and above) the same rather mediocre level of ability?

Humphreys (1973) investigated two very large groups of students who had been admitted by standard procedures to the University of Illinois. The "low" group (in terms of academic promise) consisted of students who where below the median (of all students) on both high school rank in class and a composite score on several college aptitude tests; the "high" group was above the median on both. (Humphreys notes that the "low" group is low only relative to university students but is above average in the general population.) What Humphreys found was that the standard entrance measures had considerably higher validity for predicting GPA within the high group than within the low group. But just the opposite was found for predicting dropouts among all students admitted; that is, the admission measures better predict dropout for low-promise than for high-promise students, although even in the high-promise group the predictors are related to dropping out.

Humphreys interprets these results in terms of *unmeasured variables* in the relatively few low-promise students who persisted successfully to graduation. Humphreys (1973) states,

> It is known that close to 100 per cent of students having the characteristics of high rank in high school class and high ACT composite scores attempt college. The rate of college going steadily decreases as the scores on both of these measures decrease. Thus the present low promise group has been selected by family and friends, or self-selected, to a greater extent than the high group on unmeasured variables. As the low group proceeds through college, attrition continues to take its toll and at a much higher rate than in the high group. Those who remain are again highly selected and only in part on the freshman entrance variables....[I]t is highly probable that the graduates who have the lowest scores on cognitive measures as freshmen have the highest levels on important noncognitive traits....It is also reasonable to believe that these noncognitive traits are important in later life as well as in college. (390–391)

The problem is that the favorable factors that permitted a small fraction of the low-promise group to persist to graduation are, in all probability, not *negatively* correlated with cognitive ability in the general population, so that when low-promise individuals in general are selected, as would be the case with open admissions or a very low cut-score on the selection test, the noncognitive traits that favor persistence in college are selected for no better than chance. Humphreys (1973) concludes,

> Unless open admissions were coordinated with a drastic change in curriculum content, standards of grading, and standards for degrees, the results would be catastrophic for students entering at levels of freshman predictors below present standards. The present low promise group was low only by relative standards. Their mean high school percentile rank in class was above the median; their ACT scores are undboutedly above the mean of all high school graduates. Yet less than 30 per cent survived to the second semester of the junior year overall and only about 15 percent were still in engineering. (391)

Within-Group Percentile Conversion

There is simply no selection procedure that is both strictly psychometric and "colorblind" that can reduce adverse impact to zero (that is, to the point of proportional parity for minority and majority selectees) without also reducing the utility of the selection procedure to zero. Therefore, if parity is required, strictly meritocratic selection must be forsaken, and individuals' group membership must be taken into account.

The one important question then is, Which procedure will incur the least damaging effect on utility? The unequivocal answer: Select from the top down within each group until the desired proportional parity is attained. This method is by far less damaging to utility than random selection above a low cut-score. For any desired minority/majority ratio of selectees, top-down selection within groups insures maximum utility. The typical percentage loss in utility (compared to top-down selection irrespective of group membership) is reportedly around 10 percent to 15 percent (Hunter, Schmidt, and Rauschenberger, 1984).

Some tests (for example, the GATB) now provide tables of *within-group percentile ranks* normed for various racial and ethnic groups. These permit easy conversion of raw scores to percentiles based on each applicant's group membership. Selection based on a common cut-score on the percentile scale then has virtually no adverse impact. The typically small loss in utility must be weighed against judgments about the long-term social benefits of expanding opportunity at the levels of education and jobs where minorities have been underrepresented. But such judgments, which are absolutely essential for determining policy, depend on considerations that lie outside psychometrics or statistics or science. They can only be discussed in terms of moral, social, and political philosophy. (I have expressed some of my opinions in this sphere in Jensen, 1975.)

Psychological Aspects of Modifying *g* Differences

This topic can be dealt with quite briefly. Not only is there a high degree of consensus among experts concerning the essential conclusions of the relevant research, but these conclusions encourage us to look in other directions than to changing *g* itself for dealing constructively with *g* differences. Two recent books (Detterman and Sternberg, 1982; Spitz, 1986) together afford a comprehensive review of the research and current theoretical viewpoints. In their final summary chapter of the volume edited by Detterman and Sternberg, Brown and Campione (1982) state the implicit gist of perhaps every chapter in the book:

> We now return to the contrast made in the title between training cognitive skills and raising intelligence. We would argue that although the participants may eventually be quite successful at training cognitive skills, their present papers are silent on the issue of intelligence and its modifiability. (226)

My reading of the literature on this subject has not found substantial evidence that the relative differences between individuals in *g* can be changed by any known psychological or educational techniques. The history of the most intensive attempts by many dedicated workers is well told by Spitz (1986), who arrives at essentially the same conclusion.

The problem is not that scores on specific tests cannot be raised significantly by some form of either direct or disguised "teaching to the test." There have been many such demonstrations. The problem is that the trained individuals show such a surprisingly narrow range of transfer of their training to other cognitive tasks. The fact that training-up performance on one particular type of highly *g*-loaded task may have no detectable effect on the subject's performance on a different type of highly *g*-loaded task indicates that *g* itself has been unaffected by the training. A striking example is digit span memory (which has a modest *g* loading) on which persons can improve with practice. With several months of daily practice, the average college student can increase her or his memory span from about seven or eight digits up to seventy or eighty! But when the student then is tested on memory span for random letters, it is found to be only about seven or eight letters, which is what it was before all the practice on digits span. Yet letter span also has some *g* loading. The training or practice on digit span not only had no effect on *g*, but it had no effect on memory span as a capacity independent of any particular content (Ericsson, 1988). This demonstration epitomizes the many attempts I have seen to "raise intelligence," assuming one accepts *g*, rather than performance on any particular test, as the sine qua non of intelligence (Jensen, 1989b).

Test coaching, which has become a big business with respect to the SAT, demonstrates much the same phenomenon. The average effect of one hundred hours of professional coaching on the SAT is a rise in the Verbal score of about 0.3 SD and in the Math score of about 0.5 SD. Although these coaching gains may improve some students' chances of admission to a particular college, the gain in actual academic achievement predicted by their trained-up test scores is practically negligible (Messick, 1982). In other words, the increment in test scores produced by coaching is apparently "hollow" with respect to the *g* factor that normally predicts academic achievement, and consequently the students with such inflated scores, on average, do not live up to them when it comes to actual performance in college.

Psychologists occasionally see a child who shows a quite large change in IQ over an interval of a few years. Does this mean that the level of a child's intelligence, or *g*, can be considerably altered? In any large cohort of children whose mental development is followed longitudinally, one finds reliable changes in their rank-order on indexes of *g*, and in some cases these changes, either up or down, are quite substantial. Moreover, they are often true differences in *g*, unlike the narrow contextual effects of special training or test coaching.

The causes of these seemingly spontaneous changes in rank, however, are rarely identified, so they can offer little support for claims that we should be able intentionally to bring about changes of a similar size through some kind of psychological manipulation. There is good evidence that a part of the observed irregularities—the spurts and plateaus—in mental growth rates is genetic, as indicated by the greater similarity between the mental growth curves of monozygotic than of dizygotic twins (Wilson, 1983). Similar effects are seen in the growth curves of other polygenic characteristics, such as height and weight. The effects of different genes that condition a given trait become manifested at different points in time during the course of the individual's development. The fact of spontaneous changes in *g*, therefore, is not a compelling basis for hopes of intentionally manipulating *g* by psychological means.

Another striking phenomenon that raises questions about the malleability of *g* is the apparent gradual secular rise in raw scores on IQ tests, as reported by Flynn (1984). American norms on the Stanford-Binet and Wechsler tests obtained at various times between 1932 and 1978, for example, show an average rate of increase in raw scores equivalent to three-tenths of an IQ point per year, for a total of about 13.8 IQ points. The same phenomenon has been found to varying degrees in Britain, the Netherlands, and Japan. The nature of this increase in test scores, aside from the fact that it seems to be authentic, is not at all understood, either by Flynn or by anyone else. How much of the change, if any, represents a true rise in *g* and how much is attributable to a general rise in the specific declarative and procedural knowledge content of the tests (as a possible result of the increasing universality and changing content of public education and the general access to radio and television) has not yet been sorted out. A major problem in research on this question is the lack of an absolute scale for the measurement of *g*.

It seems quite possible, however, that some part of the secular rise in test scores reflects a true change in *g*. Over the same period of time, in industrialized nations, there has also been a gradual increase in some other variables, such as physical stature, weight, and rate of physical maturation (for example, younger age at puberty and menarche), and this has been accompanied by an increase in birth weight and a decline in infant mortality. All these effects probably reflect improved nutrition and hygiene and the vast increase in inoculations to prevent the many nearly universal childhood diseases of earlier generations. (Each disease could take some slight toll on children's general physical and mental development.) Some such nearly universal causes are also suggested by the fact that the rise in test scores is about the same (in SD units) in all social classes and racial groups in the populations in which such a rise in test scores has been found. It is like a rising tide, which raises all ships equally and leaves their relative heights unchanged.

This constancy of the relative differences makes one question the inference favored by Flynn, that whatever unknown factors are the cause of the secular rise in test scores during the past fifty years or so are the same factors as those causing the social class and racial group differences (which have remained remarkably constant over the same period). There are reasons for suspecting that different causal factors are involved in the two phenomena, but because it would require too great a digression to explicate these here, the reader is referred to the exchanges among Flynn, Nichols, and Jensen in the volume edited by Modgil and Modgil (1987). But this whole matter of a secular rise in scores on all kinds of tests, especially those considered "culture reduced," is at present quite puzzling to everyone. It would be unfortunate to settle on any explanation prematurely; the needed further research on this puzzle might throw new light on the measurement and nature of *g*. Flynn (1987) has presented the most penetrating analysis of these theoretically troublesome aspects of *g*.

An explanation of the apparent failure to demonstrate an authentic change in *g*, by means of training or behavioral manipulation, would have to take into account the fact that, although the *g* factor is identified by means of psychometric tests, it is enmeshed with other variables completely outside the realm of psychometric tests. Hence, persons who are relatively high or how in *g* also differ in ways that could not be in the least inferred from examination of the highly *g*-loaded psychometric tests on which they differ. It is extremely improbable that the kinds of training typically seen in attempts to "raise IQ" would have any effect on many of these correlates of *g*. The fact that there are nonpsychometric, as well as psychometric correlates of *g*, could mean that both types of variables reflect some causal substrata that cannot be affected by direct behavioral manipulations, at least not by any kinds that have been tried so far. The "extra-psychometric" aspect of *g* is evident in findings such as the following:

1. Infants adopted shortly after birth have, as teenagers, IQs that are correlated with the IQs of their biological mothers, with whom they have not had any contact since shortly after they were born (Horn, Loehlin, and Willerman, 1979).
2. Cognitive tests are correlated with evoked electrical potentials of

the brain to the degree that the tests are *g*-loaded (Haier et al., 1983; Schafer, 1985).

3. Gifted children (IQ > 130) are faster than their average age-mates and their own siblings in reaction times to simple visual or auditory stimuli that have no intellectual content (Cohn, Carlson, and Jensen, 1985; Jensen, Cohn, and Cohn, 1989); they are more myopic than average children or their own lower-IQ siblings (Cohn, Cohn, and Jensen, 1988), and they have more allergies and are more often left-handed (Benbow, 1988).
4. Precise laboratory measurements of the duration of eye fixations on novel stimuli obtained on infants before they are six months old are correlated .5 to .6 with IQ obtained when the children are four to six years of age (Kolata, 1987).
5. Highly *g*-loaded tests are moderately correlated with "inspection time," or the time (independent of response time) required for making simple visual and auditory discriminations (Brebner and Nettelbeck, 1986).
6. Multivitamin and mineral supplements to the diets of some children increase their IQ (nonverbal IQ more than verbal IQ) by about five points within a few months—a finding consistent with other studies relating optimal neural functioning to adequate levels of thiamine, B vitamins, zinc, and iron (Benton and Roberts, 1988).

In view of such findings, it seems probable that the essential locus of control of individual differences in *g* will have to be sought primarily at a neurophysiological rather than a behavioral or psychological level. But that is largely unexplored territory and is most unlikely to be helpful at present. So, we are left with the question that educational psychologists necessarily must ask: Are there feasible means within our ken at present that might reduce the untoward effects of *g* differences in the overall economy of socially valued achievement and self-fulfillment?

Reducing the Effects of *g* in Education and Training

Before reviewing some of the approaches that my study of the research literature suggests are promising, I should indicate my guiding principles for realistic expectations for all of these approaches.

1. It now seems most unlikely that we will discover some new or previously hidden form of intelligence that will substitute for *g* and which we can tap into by some innovative instructional methods.

The "pop" psychology notions of developing children's right brains, left brains, creativity, hidden potential, or the like, are not backed by any evidence of promise for meeting the problems we are concerned with here. The hyper-development of a highly specialized ability or talent in the presence of very low *g* confers exceedingly little benefit, as can be seen most dramatically in the case of so-called idiot savants, who may have extraordinary powers of arithmetical calculation, artistic, or musical ability. Yet, they cannot earn a living by these talents and are never recognized as outstanding mathematicians, artists, or musicians. They have to be taken care of like any other mentally retarded persons.

Thus, the idea of cultivating other intelligences without considering the level of general ability is a blind alley. This opinion is not at all contradicted by the fact that persons we recognize as highly accomplished in any particular pursuit are never outstanding solely by virtue of their general ability, or *g*. Other exceptional personal assets are invariably a crucial feature in outstanding achievement, while some probabilistic minimum threshold level of *g* (depending on the type of achievement) is a necessary-but-not-sufficient condition. Consequently, the level of achievement of virtually all truly outstanding achievers is far more exceptional, in a statistical sense, than their level of *g*.

2. The only known dependable means for substantially reducing variance in overall achievement is by handicapping individuals at the upper end of the ability distribution by inadequate instruction, restricting the opportunity or time available for learning, hindering motivation, setting low standards, and the like. It is much easier to pull down the top of the distribution than to pull up the bottom. But restricting those with higher ability is so obviously unacceptable that it is mentioned here only to be cautioned against as an inadvertent possibility in our effort to make school more rewarding for the less able students.

Unfortunately, the very conditions described above as hindering the possible achievement of high-ability students exist in some schools, particularly schools in which the average achievement level is poor and teachers have become dulled to the special needs of the most able pupils. Then, for these able pupils, achievement may fall shamefully below their actual level of ability to achieve. Just how much wastage of potential achievement occurs for this reason is not known. But it is important that the underachievers be identified by suitable tests and other means for recognizing academic aptitude (or other talents), especially in disadvantaged groups in which academic aptitude is least apt to be recognized by parents, peers, and teachers. This is where appropriate tests and other methods for identifying talent can make a positive contribution. Baldwin

(1985, 1987) has suggested such methods and gives a promising description of their use and results. Much greater efforts along these lines should be encouraged. If the level of ability needed for superior achievement cannot be created by education, it is especially important that schools recognize those individuals who possess superior ability as early as possible and foster their potential.

Such measures, of course, would increase the range of individual differences in achievement. This is almost a basic "law" of individual differences in learning: By improving the conditions of learning, we cannot increase the mean level of performance without also increasing the variance. Experiments with computer-assisted instruction (CAI) for reading in the elementary grades gave striking evidence of this. While the achievement of all pupils receiving CAI showed an improvement over pupils receiving only conventional instruction, the CAI group also showed a much greater spread of individual differences (Atkinson, 1974). I have not seen a demonstration of a group's mean level of achievement being raised without its variance also being increased. (The exceptions involve learning tasks which have a fixed performance ceiling attainable by nearly everyone within the amount of practice time available to everyone.)

A closely related observation has been made by Bereiter (1987), one of the most experienced psychologists in the field of improving instructional methods:

> In my experience any instructional innovation that puts certain skills within the reach of previously failing children also makes it possible for the more successful children to acquire those skills at an earlier age. The resulting acceleration can easily increase the spread of differences. (334)

3. The aim of remedial efforts, therefore, should not be directed at trying to decrease individual differences but rather to increase absolute levels of achievement in essential knowledge and skills sufficiently to allow a larger percentage of the population to become self-sufficient and productive by ordinary societal standards. There will always be large individual differences in the kinds and amounts of knowledge people possess, but certain kinds of knowledge act as an "either/or" threshold for success in a particular society. Basic literacy and numeracy, for example, have become strong predictors of the following dichotomous classification: successful employment versus chronic unemployment. Both *declarative knowledge* (knowing *what* or knowing *that*) and *procedural knowledge* (knowing *how*) are products of learning, and in principle they can be taught. It is only a question of what, how, when, and how much of it to teach.

4. In large-scale evaluation studies (for example, Stebbins et al., 1977) of the variety of compensatory education programs developed in the 1960s and 1970s, the approaches that seemed to show the least promising outcomes emphasized cultural enrichment (that is, the provision of a variety of typically middle-class experiences to economically disadvantaged children), school services, general principles of developmental psychology, and multicultural education. The compensatory programs emphasizing these approaches had an overall slightly negative effect on scholastic achievement, as compared with matched control groups enrolled in schools without compensatory programs.

The one compensatory model that produced the most significant gains in achievement was *direct instruction*, that is, teaching directly the particular knowledge or skills in which the pupil is deficient. Direct instruction has not only proved sounder than other approaches empirically, but it makes more sense in light of what we know about the psychology of learning and cognition. The comparative results of compensatory and "follow-through" programs, however, are subject to controversy, as seen in the methodological critique by House et al. (1978) and the rebuttals to it in the same issue of the *Harvard Educational Review* (1978, no. 2).

Children (and adults) will learn what is directly taught, provided the teaching method elicits pupils' full attention and is not so sketchy or confusing that pupils have to discover for themselves what is being taught, and provided "information overload" is avoided by not trying to teach too much too fast for the pupils' rate of consolidation. Because of large individual differences in rate of consolidating new material, some children learn more than do others per unit of instruction or study time, given equal attention and motivation. For scholastic subjects, these individual differences mainly reflect two factors: *g* and the degree of mastery of prerequisite material. The second factor, which is subject to direct control in the instructional process, is the essential point of the *mastery learning* model of instruction (see below).

The principle of *direct instruction*, therefore, is uppermost in all of the approaches I will consider, both as a means to remedy existing deficiencies in essential knowledge and skills and to insure adequate mastery of pivotal prerequisites for the learning of new material.

Developmental Readiness

The whole concept of developmental or maturational readiness for particular forms of learning was almost totally eclipsed by the behavioristenvironmentalist influence that pervaded educational psychology in the 1950s. It is amazing how thoroughly the concept of readiness was ignored in the theories and practices of the compensatory education programs that arose in that period. Even the theories of Piaget, looming on the scene at that time, were interpreted in keeping with behaviorist theories. Some educationists expected to accelerate children's scholastic progress by attempting to teach them to perform Piaget's experimental tasks (used by Piaget only to investigate stages in cognitive development) a year or so before the age at which the children would normally be able to do them spontaneously without any instruction.

We now have solid evidence that there are large individual differences in children's readiness for school learning, from kindergarten through high school and beyond. At any given age, these differences are largely *g* differences, and they can make a big difference in the ease with which a child "catches on" to what is typically taught in the primary grades. Early on, these differences between children of the same age look very much like the typical behavioral differences one sees between younger and older children. This is why Binet proposed the idea of "mental age" as a metric for characterizing a child's mental status at a given chronological age. The average first-grader (age six years) has a mental age (MA) of six; just one SD below and above the average MA extends from MA five to MA seven, and two SDs extend from about MA four to MA eight, which embraces about the full range of developmental levels typically seen in the first grade. The range of individual differences in MA increasingly widens in each successive grade.

Psychologists Binet, Gesell, and Piaget have discovered or invented various developmental tasks which the average child finds easy to do at age seven or eight but which the average child of five or six finds either frustratingly difficult or altogether impossible. The very same thing is seen for children with MAs of five or six when compared with those with MAs of seven or eight, regardless of their chronological age. Now, it so happens that these developmental tasks have much in common with the typical learning tasks of the primary grades. Researchers at the Gesell Institute of Child Development at Yale University discovered that children's performance on these developmental tasks are good indicators of their readiness for learning what is typically taught in the primary grades, especially the basic elements of reading, writing, and arithmetic (Ilg and Ames, 1964).

The point is that, because of the wide range of individual differences in school readiness, some children are placed under much greater stress than others and are at risk of failure and lowered self-esteem right from the beginning of the primary grades. The slower child, under pressure to keep up with his or her peers and win the teacher's approval, begins to perceive the school as a punishing experience. School learning then is anything but pleasurable. The strongest reward or reinforcement for learning is the learner's own immediate perception of successful performance following effort to improve. We know from the classic studies by Pavlov and Thorndike that effort and performance not followed by reinforcement lead to extinction and inhibition of both the nonreinforced behavior and the effort that accompanied it. Attempted performance of the complex skills that evidence school learning can take this untoward course of extinction and inhibition when the material to be learned is beyond the child's developmental grasp. Also, when the child's efforts are frustrated by a lack of readiness for the assigned task, the well-known frustration-aggression hypothesis predicts that the child will "act out" with verbal or physical aggression, vandalism of school property, and the like—types of behavior most commonly seen in schools with a high percentage of pupils who are failing to learn.

Since practically nothing has been done with this hypothesis by way of its application to primary education, it would be unwise at this time to recommend more than an experimental approach to delaying instruction in certain skills to accord with a child's assessed level of readiness. Such a radical approach would probably have to be attempted under the auspices of a university laboratory school. The idea runs so counter to the popular push for earlier and earlier introduction of scholastic subjects that many parents predictably would object to having their children take part. The Scandinavian countries, however, have been doing this for over thirty years. Their schools do not even begin reading instruction before age seven. Yet the rates of illiteracy in these countries are among the lowest in the world. Delaying the age of reading instruction to age seven or eight has no adverse effect on the level of reading comprehension attained at later ages, as compared with beginning reading instruction at ages five or six.

The child's level of reading comprehension is highly related to the child's mental age. But MA acts as a readiness threshold for the initial acquisition of basic reading skills in much the same way that it acts as a readiness threshold for the child's ability to perform the developmental tasks of Gesell and Piaget. Many first-graders whose level of reading readiness at age six would cause them to struggle and fail in learning to read in first grade would learn more easily at ages seven or eight, and the obviously disadvantageous effects of frustration and failure in beginning reading would be prevented. The same applies to all other mentally demanding aspects of the primary curriculum.

There is probably no reason to worry about erring on the side of a little more delay than might actually be needed for some children, because the later consequences are nil, and there is still the advantage of minimizing risk of evoking negative attitudes toward school learning by an excess of early difficulties. In my travel in India, I was told by American missionaries, who formerly had been reading specialists in American schools and whose mission in India was to bring literacy to the children in totally illiterate villages, that most of the children introduced to reading for the first time between ages ten and twelve were reading, within a year or so, fully on a par with the average run of American children of the same age, although the latter had been exposed to reading since the first grade (age six). (The missionaries didn't try to teach children under age ten, because, as they put it, they wanted to maximize pupil output for the limited instructional input they could afford in any given village.) Of course, such anecdotal reports cannot qualify as bona fide research and must be weighed accordingly. But I would strongly urge obtaining some hard evidence on this matter. It could quell people's fear of the consequences of experimenting with gearing the curriculum and instruction to individual differences in pupils' readiness.

Aptitude x Treatment Interaction (ATI)

For educational psychologists, ATI has been the ardently sought Holy Grail. The gist of ATI is that the same type of instruction is not optimal for all levels of aptitude (or other learner characteristics) and that different instructional methods work best when they are appropriately matched to students' aptitudes. In terms of the regression of achievement on aptitude, the hoped-for effect of ATI is to reduce the slope of the regression line while not lowering the achievement of the high-aptitude students. Ideally, the optimum method of instruction for low-aptitude students would permit them to achieve at the same level as high-aptitude students, even when those with high aptitude get a different optimum form of instruction to maximize their achievement.

The big problem in ATI is discovering variations in instruction that will dependably produce substantial and desirable effects when the aptitude dimension is general ability, or *g*. This quest for effective ATIs has spurred a lot of research in the past twenty years. Its results, unfortunately, are confusing and practically impossible to summarize briefly. The final chapter of the most comprehensive book on ATI, by Cronbach and Snow (1977), is still probably the wisest summary. A slightly more recent summary by Snow and Yalow (1982) updates the ATI research but without a material change in conclusions.

First, the bad news. Cronbach and Snow (1977) concluded:

> We once hoped that instructional methods might be found whose outcomes correlate very little with general ability. This does not appear to be a viable hope. Outcomes from extended instruction almost always correlate with pretested ability, unless a ceiling is artificially imposed. The pervasive correlations of general ability with learning rate or outcomes in education limits the power of ATI findings to reduce individual differences. (500)

Now, the good news. The most general finding of ATI research is that the lower the aptitude, the more the learner will benefit from instruction that reduces the information load per unit of instruction time. Most of the examples of varied instructional methods that yield a beneficial interaction with individual differences in general ability seem to consist of variations on the following general principle: *Modify instruction for low-aptitude students in ways that will relieve them as much as possible of the burden of information-processing.*

The benefits of applying this principle are neither entirely consistent nor very dramatic, but the ATI literature indicates generally positive effects, so it deserves more serious consideration as one of the few promising outcomes of ATI research. The chief value of the approach is not just that it makes it easier for low-aptitude students to achieve what they would have achieved otherwise with a bit more difficulty; it also permits them to acquire essential knowledge and skills that many would not acquire at all if given only the kind of instruction that is most effective for students with higher aptitude. Too much frustration, failure, and the consequent "turn-off" of students' efforts in the early stages of learning a new subject blocks further learning of the given subject. That turn-off can be prevented to a large extent. Since initial performance on complex tasks is generally a poor predictor of final performance, getting all students successfully over the initial "hump" in a new subject is crucial. Instruction can be geared to that purpose for those who would ordinarily have undue difficulty. Many children who are given private music lessons on one of the more difficult musical instruments, for example, soon quit studying, because the first few weeks or months of effort are so unrewarding—the excruciating sound of a beginner's violin or oboe is a far cry from music. Similar frustration is experienced by some children early in their attempts to learn the three Rs, but the personal penalty for quitting is obviously far more serious than not learning to play a musical instrument.

The generalized prescription stated above translates into having highly structured and carefully sequenced instruction: simplifying or breaking down the task; minimizing nonessential elaboration, or maintaining a high degree of "figure-ground" contrast between essential and nonessential content; explicitly pointing out all the intellectual manipulations intrinsic to learning the given task, such as generalizations, logical inferences, deductions, and the like; supplementing verbal explanations with visual or pictorial displays; and substituting concrete or personalized examples for abstractions or generalizations.

These general principles for making learning less dependent on *g* are well exemplified in the recommended training procedures to facilitate learning and transfer in low-aptitude personnel, termed *functional context training*, derived from research on the training of category IV recruits in the armed services (Sticht et al., 1987). For many academically disinclined or moderately low-*g* youths, the systematic and conscientious application of these principles in specialized vocational training programs can inculcate certain high-demand occupational skills that permit entrée to productive employment in the economic mainstream.

High-aptitude students can often benefit from a more permissive instructional style that leaves them more to their own ingenuity. For them, inquiry and discovery learning can be challenging rather than defeating. More instruction time can be spent elaborating on the basic content of the lesson and it can often be embedded in a broader or more abstract conceptual context. The highly structured instruction that works well for low-aptitude students risks boring high-aptitude students with what is for them too much explanation and pointing out of the obvious. Well-structured lessons must be appropriately paced for high-aptitude students, who are able to grasp many more of the conceptual connections and inferences without explicit emphasis by the instructor.

ATI research is also full of puzzles and surprises. For example, a review by Clark (1982) of ATI studies that focused only on general ability found that high- and low-ability groups each enjoy a different style of instruction from which, it turns out, surprisingly, they actually benefit least in terms of achievement. High-ability subjects show higher levels of achievement under more permissive types of instruction, yet they generally claim to enjoy and prefer highly structured instruction, which results for them in lower achievement. They enjoy structured instruction more than permissive instruction presumably because they perceive the former as being easier, which it most probably is. Low-aptitude students were found to be just the opposite. They actually achieve more from highly structured instruction, yet they prefer the permissive approach, presumably because the task requirements and standards for performance are less clear-cut and their actual achievement more often escapes being closely monitored, either by themselves or by the instructor. The implication would seem to be that the student's own preference is at best a poor guide to the most appropriate instructional method for enhancing achievement.

Level I and Level II Abilities

The concept of Level I and Level II abilities was introduced about twenty years ago as a way of formulating some empirical generalizations from my research on children's learning (Jensen, 1968). These generalizations can now be recognized as potentially just one more kind of ATI. Because there already exist comprehensive reviews and bibliographies of my own and others' research on the Level I–II formulation (Vernon, 1981, 1987). I will not attempt here to review all the findings but will give only the gist of what seems important for our present purpose—finding ways to mitigate the effects of *g* differences in education and employment.

Three observations gave rise to Level I-Level II:

1. School-age children from poor socioeconomic (SES) backgrounds (especially low SES black children) with IQs below eighty-five or so and with correspondingly poor scholastic achievement performed very significantly better than middle or upper-middle SES white children (with the same IQs as the low SES children) on a number of experimental learning and memory tasks.
2. The correlation (with or without corrections for attenuation and restriction of range) between performance on the learning tasks and IQ was significantly smaller in low SES groups (especially if they were black) than in middle and upper SES white groups. (Middle SES black children, being scarce in the schools in which we first conducted these studies, were not represented.)
3. The dispersion (SD) of measures on simple learning and memory tasks (for example, forward digit span) in low IQ groups (that is, IQs 70–90) was greater than in higher IQ groups (IQs above 110); that is, the degree of scatter around the line of regression of learning or memory ability on IQ gradually decreased as IQ increased. This form of bivariate distribution suggested the hypothesis that a certain level of learning and memory ability is a necessary-but-not-sufficient condition for the development of a certain level of intelligence (as indexed by IQ). Hence, we see almost the full range of simple learning and memory abilities among persons of low IQ, while very low learning or memory abilities are rarely found in persons of high IQ.

These three observations led to the hypothesis that two separate classes of ability, termed Level I and Level II, were interacting with SES (or with race—black/white) in the performance of different types of tasks.

Level I was conceived as the ability for receiving information and recalling it in very much the same form in which it was presented, with a minimum of elaboration or transformation. The prime example (and test) of Level I ability is forward digit span (FDS), the number of digits that can be recalled entirely correctly in the order of input immediately upon hearing (or seeing) a series of digits presented at the rate of one digit per second. Backward digit span (BDS, the digits have to be recalled in the reverse order) is a less "pure" measure of Level I, because in BDS the subject has to mentally manipulate the input before responding with the output.

Tasks characterize Level I to the extent that a minimum of mental manipulation of the information input is required for correct response output. The nature of a task need not actually preclude a good deal of mental manipulation to be Level I; it is only necessary that it can be (and usually is) performed with a minimum of mental manipulation. One does not think in terms of an absolute zero of mental manipulation on this continuum but in terms of the rank-order of complexity of cognitive operations typically elicited by various kinds of tasks. Other Level I tasks besides FDS used in these studies were serial rote learning, paired-associates learning, and free recall of items during multiple presentations. In all of these learning tasks the information input consisted of simple words, variously colored geometric forms, pictures of familiar objects, or actual familiar objects. The role of knowledge per se was minimized in these tasks, so that individual differences would reflect mainly differences in proficiency of learning in the test situation itself. Generally, low SES/low IQ children performed better on these Level I tasks than did middle SES/low IQ children, and many of the former performed at about the same level as the average run of their white, middle-class age-mates.

Level II ability requires transformation and manipulation of the input information in order to arrive at the appropriate output response. This characterizes most of the items of conventional IQ tests, especially the kinds that are said to measure "fluid" intelligence, that is, problem-solving ability displayed in tasks for which the difficulty level depends much more on the complexity of the mental processing required on the given information than on the recall of specific knowledge acquired outside the test situation. The main measure of Level II in my studies was the Raven Progressive Matrices, a nonverbal test of reasoning based on figural material. It turns out that in numerous factor analyses, the Matrices is the most highly *g*-loaded of any single test we have found. It soon became apparent that Level II was indistinguishable from Spearman's *g*; it, therefore, reflected the same magnitudes of SES and race differences typically found with the usual IQ tests. And, of course, it predicts scholastic achievement the same as IQ, regardless of SES and race. In marked contrast, however, Level I tests do not predict scholastic achievement the same as IQ, and they predict differently for low and high SES (and for black and white) populations. Most low SES and black students showed lower scholastic achievement than would be predicted from the regression of achievement on Level I ability in the white, middle-class school population.

The key question then was: is Level I an ability that can be tapped by certain methods of instruction to achieve the basic aims of schooling for children who are below-average in Level II but are average or above in Level I, entirely without respect to SES or race? (Statistically, there is a larger proportion of Low Level II/High Level I individuals in low SES and black populations, so these groups would stand to benefit the most from instruction that capitalizes on Level I ability.)

The answer to this question can really be subsumed under the ATI generalization discussed in the previous section, because Level I can be described in terms of tasks or ways of learning that do not make heavy demands on complex information-processing. To engage pupils' Level I ability, instructional methods would have to be specifically designed to reduce the role of Level II ability, or *g*, for successful performance.

Because many black and low SES/low IQ children could perform well on Level I experimental learning tasks, and because these tasks were usually forms of "rote learning," the Level I idea became popularly interpreted as the advocacy of teaching by rote, that is, by sheer repetition of stimulus-response associations elicited by the instructor. A more sophisticated notion was that the same ability (Level I) that made it possible for some low IQ children to do well in rote learning tasks could somehow be brought to bear on learning school subjects, but not necessarily by casting them in the form of rote learning. The kinds of instruction besides rote learning that would do the trick were not specified beyond the general ATI principle of structuring lessons in ways that would reduce the burden of information-processing for the learners.

So what has become of these Level I-Level II notions in the last twenty years? In retrospect, it seems they had at least two strikes against them from the beginning: (1) educationists' scorn for anything that hints of rote learning and (2) the unfortunate but unavoidable association of the Level I-Level II theory with *jensenism*—a term coined by the popular media in their sensationalizing of my article "How Much Can We Boost IQ and Scholastic Achievement?" in the *Harvard Educational Review* (Spring, 1969), in which the Levels theory was propounded alongside other controversial topics that made the whole thing anathema to many educators and social scientists.

Although in the following years there were a great many one-shot laboratory studies aimed at testing Level I-Level II strictly as a psychological theory, usually confirming the basic observations that gave rise to the theory (studies reviewed by Vernon, 1981, 1987), I have not found examples of actual long-term classroom instruction explicitly based on Level I abilities and reported in the literature as such.

Yet, over the past twenty years, I have heard comments from teachers in predominantly black schools that their own teaching style has been shaped to a large extent by the particular style of instruction that seemed to work best with their pupils. The teaching style these teachers described is about what one would expect if one wished to promote the learning of scholastic material by pupils whose Level I ability is notably stronger than their Level II ability. The teaching style is shaped in the direction of greater emphasis on rote learning of basic information, frequent rehearsal of immediately past learned material, verbal repetition, and memorization. And the instruction tends to deemphasize intellectualized explanations involving abstractions, generalizations, concepts, and principles—in short, the very cognitive activities that most characterize Level II. These kinds of observations have come from experienced teachers who had never heard of the Levels theory or of Spearman's *g*. It is also worth noting that children's educational television programs, such as "Sesame Street" and "The Electric Company," include elementary scholastic content (ABCs, simple words and numbers, and so on) presented in the form of brief "lessons" that perfectly exemplify a Level I teaching style.

Obviously there are some problems with this style of teaching. It may work well in the early phase of acquiring simple scholastic knowledge, but success in Level I learning may also reinforce a mental set that encourages pupils to approach all new learning in a rote fashion, which may hinder development of meaningful learning involving conceptual and abstract thinking essential for advanced levels of achievement. As Cronbach and Snow (1977) have cautioned in their discussion of the Level I-Level II theory, "Any attempt to evaluate instruction that *is* related to memory abilities will have to give as much attention to transfer outcomes (including the growth of Level II abilities) as to the responses directly taught" (485).

One answer to this criticism of instruction that caters to Level I abilities is that there are some children low in Level II who would likely fail at the more advanced levels of conceptual academic learning in any case, and it is better to teach them the things that they can learn with the abilities they possess than to try to teach them these things in ways that virtually guarantee failure. There are many useful and necessary kinds of knowledge and skills that can be successfully learned by those with average Level I ability despite their being well below-average in Level II ability. But for such students, a Level II approach hinders or prevents the learning of these things.

It is a matter of degree as to how much of the conceptual underpinning of knowledge and skills one must acquire for them to be useful for the individual's particular purpose. Russell and Whitehead in their *Principia Mathematica*, for example, take some one hundred pages to establish an understanding of the logical basis for the concept of the number one, without which presumably no other number can be truly understood. Yet most of us have been using numbers in arithmetic and mathematics in our work and daily lives without the least idea of what Russell and Whitehead expounded. Rote learning has indeed been very underrated in education circles, even though it in fact plays a big part in everyone's life, regardless of one's status on Level II. I am here typing away on a computer key pad. Yet, I have almost zero knowledge of how my computer works, knowing only how to make it do what I want (most of the time) by hitting various keys, which I had to learn completely by rote (or Level I processes), since there is no intrinsic logic to the labels and locations of the keys. True, one cannot understand something like theoretical chemistry in Level I terms. But even if a Linus Pauling wanted to tend bar, he would have to learn how to make all the popular mixed drinks by rote memorization. One can easily think of countless other examples of the essential role in one's life and work played by Level I types of learning.

So, I still believe that the main idea of the Levels theory may have promising applications in schooling and in specialized forms of job training, although there is scant evidence that it has yet been systematically tried and evaluated. In their review of the Levels theory, Cronbach and Snow (1977) present what seems to me a wise evaluation:

> Some educational objectives could perhaps be better-attained, for the student averaging high in Level I and low in Level II, by making more use of rote methods in the classroom. No evidence regarding the relevance of memory tests to school learning under alternative procedures is now available, however. (485)

> This suggestion [that basic skills be taught by rote or drill to those who are comparatively weak in reasoning] is not to be rejected out of hand merely because the word "rote" is distasteful. Better that primary pupils attain literacy and numeracy by whatever means than that they should fail; perhaps a similar case can be made at later grades. But. . .if beginners are not shown meaningful connections, they will not learn from logically coherent instruction. Hence, purely rote learning leaves them permanently unfit for meaningful instruction. Assuming that the ATI for rote vs. meaningful instruction does become solidly established in some subject at some grade level, it would then be defensible to

> make rote the main vehicle for teaching that subject to certain students. But alongside this teaching there must be an effort to promote skill in the kind of learning at which these students are deficient. A coordinated attack could capitalize on strengths while repairing weaknesses. [521]

At about the time of the Cronbach and Snow review, the U.S. Navy took cognizance of the Level I-Level II idea in connection with their research on the success of category IV recruits (those between the tenth and thirtieth population percentiles on the AFQT, a general ability test) in apprenticeship and technical training programs. The navy researchers included a Level I test, auditory "Memory for Numbers," in their battery of prediction tests. In studies (Cory et al., 1980) based on very large samples, it is interesting that although this relatively pure Level I test showed nearly zero validity for predicting overall grades in navy schools, it had significant validity (+.18; corrected for attenuation, +.30) for category IV trainees—a slightly (but not significantly) higher validity than the AFQT. The validity of the Level I test (Memory for Numbers) was even higher (+.25; corrected for attenuation, +.44) in predicting *advancement* of category IV recruits into technical jobs; and again, it had better validity than ten other psychometric tests, including the highly *g*-loaded AFQT. This supports the hypothesis that success in the training of persons who are relatively low in *g* depends in part on their status on Level I abilities. The low-*g* category IV recruits who were highest in Level I were the most successful in terms of advancement from apprentice positions to technical training and subsequent technical jobs and in terms of global performance marks.

Hence, the Level I-Level II notion should not be abandoned but should be afforded a true experimental test on a large enough scale to inspire confidence in the conclusions concerning its practical efficacy. What is really needed is a clear-cut ATI design which would yield a statistical assessment of training outcomes in terms of the main effects and interactions of all combinations of both high-/low-ability groups on both Level I/Level II abilities, under both rote/meaningful instruction treatments (that is, a three-way analysis of variance).

To determine if some of the educational psychologists who are knowledgeable in the field of instructional psychology knew of any informal or unpublished studies that might throw more light on the potential merits or demerits of Level I-Level II, I wrote and spoke to a number of key researchers. Although they could point to no ideal studies, they all expressed positive but qualified opinions of the potential value of Level I types of instruction, much like the previously quoted views of Cronbach and Snow (1977). One prominent correspondent (who wishes not to be quoted by name) added the following observation to his comments on Level I-Level II:

> I believe your earlier [Level I-Level II] idea still has merit, but when it comes to achievement, the black underclass has huge deficits which are not at all mandated by their lower Level II, but by their behaviors in other ways not nearly so evident in majority youth of equivalent ability, and behaviors that are indeed susceptible to molding, channeling, and reinforcement. This leads me to a belief that the place to concentrate is on family life, and the re-design of welfare and other policies to encourage fathers to stick—not discourage them as we do now; and to encourage fiscal independence, not pauperism or the rackets. My own research lead me to believe in the social benefit from the dogmas and activities of religion—these variables appear to be about zero correlated with IQ, but indeed are correlated with achievement. (Anon., March 1, 1988)

The Carnegie Corporation is presently sponsoring several large-scale projects in inner-city schools based on these very ideas, in which churches in the black community have organized programs for parents and children expressly to promote the kinds of morale, social attitudes, and personal lifestyle that favor scholastic endeavor (*Carnegie Quarterly*, 1987–88).

The Demise of Level I-Level II as a Theory. This has no bearing on the possible educational applications of the Level I-Level II concept, but I never viewed this formulation as a theory so much as merely a set of fairly well-established empirical generalizations. Some ten years ago, I quit using the terminology of Level I-Level II and also gave up all thoughts of trying to develop it as a theory. I abandoned Level I-Level II as a potential theory for three reasons:

1. There were futile arguments among researchers as to whether certain tasks should be classified as Level I or Level II, and there were no objective means for resolving these arguments. Many of the purported disproofs of the "theory" were based on flagrant misconceptions of the meaning of the two levels and hence a lack of agreement among different investigators in the classification of experimental tests as Level I or II. I have little use for theories that cannot be empirically falsified or cannot compel agreement among reasonable persons on the basis of empirical evidence. The ambiguities of definition that foiled the rigorous testing of the theory undermined its attractiveness from a scientific viewpoint.

2. The application of factor analysis to the correlations among a number of different Level I tests and Level II tests revealed that (a) Level II is indistinguishable from Spearman's *g* and (b) Level I is not a unitary factor. Various Level I tests (memory span, serial learning, paired-associates

learning, and free recall) largely part company in a factor analysis, and the relatively small general factor among Level I tests is nothing other than *g*. Hence, it was necessary to speak of Level I abilities in the plural, as being whatever different *non-g* factors were involved in tasks that could be classified together as either requiring rote learning or short-term memory. Since such tasks did not necessarily cluster in a factor analysis, there was no objective means for settling arguments as to which particular tasks could be legitimately classified as valid measures of Level I. In fact, just about any task with a very low *g* loading and a large specificity would behave with respect to race and SES differences much like the best Level I tasks, although they would not necessarily be memory or learning tasks. In other words, Level I had a big problem with construct validity. And since Level I measures behaved just like any other non-*g* (or very low *g*) ability measures, it also lacked divergent validity. In essence, Level I and Level II abilities boiled down to *g* and non-*g* factors.

3. Hence the most compelling reason for abandoning the Level I-Level II "theory" is that it turns out to be unnecessary. The Law of Parsimony requires that it be dropped. I came to realize that the Level I-Level II theory was essentially a special case of what I have termed *Spearman's hypothesis* (that is, the magnitude of the mean black/white differences on various tests is a function of their *g* loadings).

Also, the fact that the Level I-Level II generalizations held up much better when tested in terms of black and white groups than when tested in terms of high and low SES white groups reinforced the belief that they were actually a demonstration of Spearman's hypothesis. In the course of my empirical investigations of Spearman's hypothesis, however, I substantiated a subsidiary hypothesis (unknown to Spearman) that is an important supplement to Spearman's hypothesis for fully comprehending the observations that gave rise to Level I-Level II. The subsidiary hypothesis is this: When representative black and white groups are matched on *g* (or *g* is statistically controlled), blacks, on average, outscore whites (and, it turns out, Asians and Hispanics as well) on a memory factor (mainly loaded on digit span). Hence all the descriptive aspects of black/white differences on various kinds of psychometric tests can be comprehended strictly within a completely objective factor-analytic framework in terms of Spearman's hypothesis plus the subsidiary hypothesis that blacks, on average, outperform whites on tests of memory independent of *g*. Unlike the Level I-Level II theory, the modified Spearman's hypothesis can be (and already has been) put to completely objective and statistically rigorous tests and is strongly borne out (Jensen, 1985a, 1985b, 1987c; Jensen and Reynolds, 1982; Naglieri and Jensen, 1987).

Mastery Learning

Individual differences in *g* (or IQ) are highly correlated with the time to learn a new lesson up to a given criterion of mastery (Gettinger 1984). In typical classrooms, the slowest pupils take 500 to 600 percent more time than the fastest pupils to learn a given amount of material to the same level of mastery. Conversely, given a uniform amount of time, the fastest pupils should be able to learn five to six times more than the slowest. The pacing of instruction in typical heterogeneous classes usually does not allow either of these extremes. Slow learners attain low levels of mastery in the knowledge or skill content of a given lesson before having to move on to the next, while fast learners often attain a high level of mastery and are ready to move ahead well before they are presented the next lesson.

The idea of *mastery learning* is to keep pupils working at a given lesson, with the teacher's help, until they reach a uniformly high level of mastery, as indicated by a test of the lesson's content. All pupils are required to attain the same high level of mastery (say, 90 percent correct on a test designed to sample the lesson's contents), even if there is a wide range of individual differences in the total time needed to reach the required degree of mastery. This procedure obviously demands frequent and specific monitoring of pupil performance and hence can be administered more effectively with individual than with group instruction, where teachers may get bogged down in testing and record keeping. Also, it has been found that in group instruction teachers tend to allow the average pace to be set by the slower pupils (generally those in the tenth to twenty-fifth percentile of learning rates), thereby limiting the achievement of the faster learners—a condition that educators have termed the "Robin Hood effect." Obviously, computer-assisted instruction, which permits individuals to learn at their own pace, is a decided boon to mastery learning.

Contrary to some of the exaggerated claims made about the benefits of mastery learning, there is no getting around the fact that it amounts to a trade-off between the level of mastery achieved and the amount of material covered. If we decrease the range of individual differences in level of mastery attained on any unit of instruction by providing all pupils sufficient time to attain the same level, we correspondingly increase the range of individual differences in the number of units that can be covered in a school term, assuming that the pacing of instruction is not drastically slowed down for the faster learners. Mastery learning appears to decrease the range of individual differences in achievement only when the outcomes are assessed by means of specially designed achievement tests that test only for the information that was directly taught. Hence, under mastery learning of a limited and clearly specified content of information which is exclusively the basis of the outcome measure, a performance ceiling is imposed, which theoretically should reduce individual differences in performance to near zero.

In practice, however, comparisons of mastery learning classes with control classes show achievement gains in the range of one-half to one SD, but only on tests of explicitly taught material. The effect size is small indeed when achievement is measured by conventional standardized tests. Besides assessing performance on what was directly taught, standardized achievement tests also assess related incidental learning, inferences drawn from the explicitly taught material and generally a broader range of information in the given domain of subject matter than is tapped by the tests specifically designed to assess mastery learning outcomes. These broader aspects of achievement, which reflect general transfer of training, tend to be highly *g*-loaded. Consequently, mastery learning only slightly reduces the range of individual differences in scores on standard achievement tests.

Much of the research on mastery learning outcomes and their problems is impressively reviewed by Slavin (1987a), along with a number of critical commentaries (Anderson and Burns, 1987; Guskey, 1987; Bloom, 1987). Slavin (1987b) concluded, "To value the results of mastery learning research, it helps to hold a philosophy that reducing the variance in student performance is more important than increasing the mean for all students" (234).

There is no reason in theory, however, that mastery learning should necessarily have this undesirable effect. Properly applied, with sufficient attention given to individual differences in learning rates, it should be able to increase the level of achievement for all students without reducing variance in performance. One of the main problems with the mastery learning approach is the trade-off between level of mastery and breadth of coverage of subject matter. It is necessarily a trade-off, because children can spend only a limited amount of time in school. (At a conference I attended some years ago, a mastery learning enthusiast suggested it should be theoretically possible, if given a sufficient amount of time, to bring a retarded child up to the level of Bertrand Russell in mathematics. Asked "How *much* time?" his answer was "Perhaps two hundred years.") Obviously, there have to be choices of precisely which scholastic material everyone should be required to master to a high criterion and which subjects we can allow to have a much wider range of variation in degree of mastery. Reading text and reading music are contrasting examples.

A limited use of mastery learning for just those elements in the curriculum that constitute the most basic declarative and procedural knowledge that are essential tools for further scholastic progress is feasible and theoretically should raise the achievement levels of many pupils who ordinarily would "top out" at a socially unacceptable level. Failure to attain a high level of mastery of the essential prerequisites for learning a given subject increasingly hinders further learning and imposes a low ceiling on the student's eventual level of ahievement in that subject area. Some students, for example, are unable to learn the kinds of arithmetic normally taught in the fifth and sixth grades because they have not sufficiently mastered the more elementary arithmetic taught in the third and fourth grades. Some elements must be mastered, or overlearned, to the point of being "automatized," if they are to benefit students when they are confronted with more advanced material. The reasons for this are understandable in terms of recently researched models of information-processing, which are discussed in the final section of this chapter. The potential efficacy of mastery learning can best be understood in the context of information-processing theories.

Teaching Thinking Skills

The mental activity of thinking is certainly not the same thing as *g*. Thinking is actually a form of behavior, and so, like any other behavior, is subject to the principles of learning. Hence it can be taught, reinforced, shaped, and honed, much as any other skill. Thinking is essentially talking to oneself, overtly or covertly, in ways that interrelate and organize certain items of knowledge or experience to construct a coherent and consistent model of some phenomenon. It is also a way of asking questions of one's experience, recognizing problems, and discovering what is needed to solve them. Thinking skills are fairly generalizable processes or strategies, such as simple classification and hierarchical classification, sequential ordering of things along some dimension, reasoning by analogy, breaking down complex concepts into simpler components, reducing a complex problem to its essential elements, and the like. Study skills are really just the application of the appropriate thinking skills to the learning of a given subject matter.

Everyone agrees that a primary goal of education is to teach students how to think. Yet, in recent years, the schools have been accused of falling short in this endeavor. In an era of rapidly developing and changing information-intensive occupations, it is thought that learning and thinking skills are more called for than a fund of specific subject knowledge per se, beyond the basic tools for educational advancement—the three Rs. But these skills are best acquired in the context of some real, relevant, and culturally recognized subject matter. Education critics, and many educators themselves, have argued that more explicit teaching of thinking skills must be infused into this subject matter context. Some even advocate separate courses for the teaching of thinking skills.

This is one of the presently debated issues in this field—whether thinking skills can be taught separately or whether they must be an adjunct to some conventional content area, such as science, social studies, or literature. Thinking obviously requires content—it does not take place in a vacuum. It has long been argued that thinking skills, although they are transferable to a wider context than that in which they were specifically taught, show drastically diminished transfer to other domains that have few elements in common with the training context. But I doubt that any broad generalization is warranted on this issue. The extent of transfer of training of thinking skills depends on so many variables that there are virtually no principles that will reliably predict the transfer outcome of any given training procedure, and so arguments about the effects of any given program of training must be answered empirically (Nisbett et al., 1987). One argument made for teaching thinking skills in separate courses rather than exclusively in the context of a specialized subject domain is that the instruction does not have to contend with the usual wide range of individual differences in the knowledge of a particular subject.

In recent years, we have seen an explosion of programs for training thinking skills. It has become a growth industry in education. Some of the better known programs are Instrumental Enrichment, Philosophy for Children, Structure of Intellect (SOI), Problem Solving and Comprehension: A Short Course in Analytical Reasoning, and Odyssey. They are all diverse in their methods and the types of students for which they are intended. The research that I have seen on most of these programs is methodologically so far below the normal standards of referred psychological and educational journals as to afford little basis for appraisals or comparisons of their efficacy.

The most notable exception is the Odyssey program, which was developed and tried out in connection with Venezuela's Ministry for the Development of Intelligence. The study by Herrnstein et al. (1986), based on the Odyssey program, could well serve as a model for research in this field. It is one of the few large-scale studies that uses proper control groups and assesses transfer outcomes with a variety of criterion measures that do not overlap the specific training tasks, and the results are encouraging. Some four hundred Venezuelan seventh-grade students were given, during one school year, one hundred lessons of about forty-five minutes each in various thinking skills involving language, reasoning, problem solving, decision making, and inventive thinking. Training outcomes were measured with a target based on the instructed material and three standard tests of general abilities (Otis-Lennon, Cattell, and GAT) that had been designed without reference to the instructional program. The gains over an untrained control group were of the order of 0.3 SD to 0.4 SD on the standard tests and about twice that amount on the target tests. These well-substantiated effects seem especially remarkable considering the brevity of the training. There is not yet evidence of the degree of persistence of these training effects or of their transfer to subsequent scholastic achievement, but the initial effects certainly warrant further studies of the Odyssey program in American schools.

The training of thinking skills thus appears to be one of the most promising avenues for improving students' competence, especially in ways that should be beneficial beyond their formal education. But few of the thinking skills programs are specifically designed for the kinds of students who have the greatest difficulty in school. We will need more demonstrations of the effectiveness of such training for this group. It is likely that we will need to develop special programs to be optimal for different levels of general ability. It is most encouraging, in this respect, to note that there were very similar *percentage* gains for students across the entire spectrum of general ability in the study by Herrnstein et al. (1986).

But there are so many complex theoretical and empirical issues in this field, and so many different approaches and empirical findings around which the current debates revolve, that it would be quite impossible to do them justice in this brief introduction. Fortunately, there are now some superbly thoughtful and critical reviews of the main currents in this field (Adams, 1989; Nickerson, 1988; Sternberg and Bhana, 1986)—essential reading for those who would venture into this field.

An Information-processing Model of Psychometric *g*

The phenomenon of psychometric *g* was discovered eighty-five years ago (Spearman, 1904), and it has been a classic "black box" in psychology ever since. Psychologists are still trying to discover the nature of the mechanisms or processes that can explain *g* and its relation to other phenomena. The experimental psychologist's endeavor is not unlike the physicist's effort to fathom the basis of matter. Various behavioral (and a very few physiological) correlates of psychometric *g* are measured under specially devised laboratory conditions, and the data are used as probes to develop and test hypotheses or models of what goes on in the black box.

The cutting edge of this research today is allied with recent work in experimental cognitive psychology. It has largely adopted the terminology, concepts, and metaphors of the information-processing models that originated with the development of computers and work on artificial intelligence (Newell and Simon, 1972).

Rather than explicating any specific model that has been the focus of theoretical investigation to any particular school of thought in this field, I will try to indicate some of the main research aims in terms of a simple generic model that incorporates the essential features of many other models. To begin, a few definitions will help.

An *elementary cognitive process* (ECP) is a hypothetical construct that plays a crucial role in information-processing. A specific mental content (that is, sensation, percept, image, memory) is acted upon (discriminated, encoded, repeated or rehearsed, transformed, stored, retrieved) in some singular way by a particular ECP. The information-processing system has some limited (but as yet undetermined) number of ECPs. It is assumed (and empirically demonstrated in some ECPs) that there are individual differences in ECPs, but it is not yet settled how independent (uncorrelated) the ECPs are. My present surmise from the available evidence is that various ECPs are correlated, but far from perfectly. There is probably a common factor (most likely at the neural level) in all ECPs, but this common factor in ECPs constitutes only some fraction (probably less than one-fourth) of the total variance of psychometric *g*. Hence, the ECPs would have considerable independence, and a great many different patterns of individual differences would exist that could not be explained by any single factor common to all of the ECPs. It seems most likely that variance in psychometric *g* comprises both the common factor variance of all the ECPs and the specific variance of each of them. Getting a solid answer to this conjecture is one of the major aims of investigation.

An *elementary cognitive task* (ECT) is (usually) a laboratory contrivance for measuring an ECP in terms of response time, which affords an absolute scale. Most ECTs unavoidably measure two or more ECPs, but the separate ECPs can often be measured indirectly by comparing measurements derived from two or more different ECTs that are hypothesized to reflect different ECPs. For example, one ECT measure reflects the operations of ECPs $a + b$, while another ECT measure reflects the operations of $a + b + c$. By subtraction we can obtain a measure of c. (There are also other more complex methods than simple subtraction that we need not go into here [see Jensen, 1985c].) ECTs measure elementary processes such as stimulus apprehension, encoding of sensations, discrimination, choice or decision, and retrieval of information short-term or long-term memory.

Because ECTs are necessarily exceedingly simple tasks, individual differences cannot be measured in terms of number of correct or incorrect answers, as in conventional psychometric tests. It is usually an essential requirement of the experiment that all subjects be able to perform the ECT with a very high level of accuracy. Error rates are usually kept so low that individual differences in errors are almost random and hence have low reliability. Reliable measurement of individual differences, therefore, depends on the use of chronometric techniques (Jensen, 1985c). The main measures of interest, then, are (1) speed of response, or the median reaction time (RT) over *n* trials, and (2) the consistency of response speed from trial to trial, measured as the SD of the RTs over the *n* trials.

Every component of information-processing occurs in time and the amounts of time for various components can be measured with great precision. The fact that time is measured on a true ratio scale with internationally standardized units is one of the great scientific advantages that experimental research and theoretical development based on mental chronometry has over conventional psychometry.

In normal young adults, the total time required for the performance of most ECTs is very short, usually less than one second between stimulus and response, and many ECTs are in the range of two hundred to six hundred milliseconds. Some part of this time (something between one hundred and two hundred milliseconds) consists of sensory lag plus afferent and efferent neural conduction time. The rest is central processing time. Now, it is an important empirical fact, quite apart from any theoretical interpretation, that virtually every ECT in which individual differences have been tested chronometrically has shown a significant correlation with psychometric *g*. Individual differences in median RT and in trial-to-trial consistency (that is, intra-individual variability) of RT are both correlated with *g*, each somewhat independently of the other. The correlations vary, depending on the particular ECT—its degree of complexity, or the number and types of different ECPs it is hypothesized to reflect. The correlations are generally in the .2 to .4 range but are seldom higher than .5 or .6, even after corrections for attenuation and restriction of the range of talent.

Multiple correlations based on a number of different ECTs, however, can be considerably higher. This suggests that psychometric *g* reflects the operations of a number of at least partially independent cognitive processes, no single one of which can account for more than about 10 percent to 15 percent of the true variance in *g*. But there is reason to believe that some processes probably contribute more than others to the variance in *g*. The simplest or most elemental process we have yet found to be significantly correlated with *g* (recently demonstrated in my lab in collaboration with Professor T. E. Reed) is the time for a visual stimulus to arrive at the visual cortex. It is less than about one-fourth of the time required for conscious recognition of the stimulus. So this represents just the preconscious phase of stimulus apprehension, reflected by the brain-evoked potential, occurring on average about one hundred milliseconds following onset of the visual stimulus. Yet, amazingly, it is very significantly correlated with nonverbal IQ in a restricted range of college students. I mention these surprising findings to emphasize that even the most basic and elemental information processes contribute some part of the variance in *g*. Many examples of correlations between ECTs and *g* have been reviewed by Snow and Lohman (1988) and in a recent book edited by Vernon (1987).

The term *metaprocess* in this domain refers to executive processes, or processes which govern the deployment and organization of ECPs for problem-solving routines, planning a course of action, and monitoiing performance. Its metaphoric overtones of a homunculus acting somewhere in the brain like a traffic cop or orchestra conductor makes it seem a rather unappealing construct. It is acceptable, however, if metaprocess is used generically to mean any kind of learned strategy, complex routine, or integrated set of covert or overt responses that play a part in purposive behavior. Some metaprocesses, by becoming "automatic," theoretically explain marked variation in the efficiency of complex information-processing.

A simple schematic representation of the hypothetical information-processing system is shown in figure 4–2. No special virtue is claimed for this particular schema, but it includes the main elements of many other proposed models. Research in mental chronometry shows that each of these processing stations (represented as rectangles) takes some amount of time, as does the transfer of information between them (represented as arrows). The total processing time between stimulus input and response output will depend on:

1. the complexity of the input,
2. how many processing elements are involved,
3. how many paths are traversed,
4. how many transformations are made, and

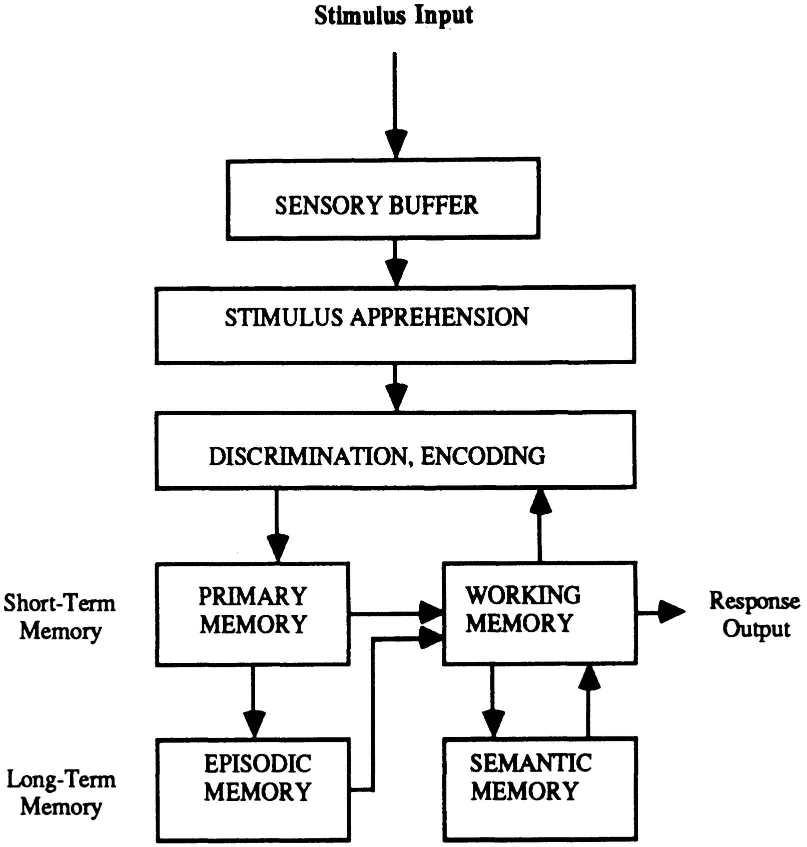


Figure 4–2. Hypothetical Schema of Information-Processing Components, with Arrows Indicating the Direction of Information Flow

5. how efficiently all these processes occur, both separately and as an integrated system.

All of these variables seem to be positively correlated to different degrees with *g*.

The largest source of *g* variance is the short-term memory (STM) system. A theoretical interpretation of the research findings in this system makes it necessary to posit two distinct aspects of STM, termed *primary memory* and *working memory*. Primary memory is a passive, limited-capacity, short-term storage system. It is passive in the sense that it does not perform transformations or other manipulations on the information contained in it. It has limited capacity in the sense that it can "hold" only a limited amount of information at one time; further input beyond that limited amount interferes with and "erases" the previous information held in primary memory. And it is short-term in the sense that the most recently input information is rapidly lost as a function of time.

Working memory can be thought of as the focal point of *g* in the whole system. It has been referred to as the mind's scratch pad, but it is better likened to a computer's central processing unit. The working memory transforms and manipulates information received from primary memory or retrieved from semantic memory. In most learning and problem solving, the working memory retrieves from long-term semantic memory whatever information is needed to interpret the recently input information in primary memory. Like primary memory, the working memory has a limited capacity and the information it operates on is also subject to rapid loss over time. The working memory can transform information received from primary memory and return it to primary memory for later use, or it can further encode and rehearse the information to get it into long-term memory (LTM) from which it can be retrieved for later use.

Two properties of working memory are hypothesized to account for what experimental evidence leads us to surmise about its functions: (1) the *speed* or efficiency of performing its operations and (2) its *capacity*, or the amount of information or number of different operations it can deal with at one time. These two attributes seem to be closely related, but the exact nature of the relationship is still obscure. Yet both seem necessary for understanding some of the phenomena of individual differences. We know that when the capacity of working memory is strained by requiring the subject to perform a dual task (some mental manipulation on one task while holding immediately previous input information in STM), the RT on the target task is increased over what it would have been if the same task had been presented as a single task. (Intra-individual variability in RT is also increased in the dual task condition.) And the *g* loadings of both variables are larger in the dual than in the single task condition. If there were individual differences in only one underlying process (either speed or capacity), single tasks and dual tasks should correlate equally with *g* when corrected for attenuation. There are also more complicated findings (which would require too much explication for inclusion here) that cannot be explained if we hypothesize a single basic individual-differences parameter in working memory rather than at least two. I hypothesize a model in which an individual's RT to a given ECT is a function of the total time (T_0) in seconds taken by all the processing of the input and output that occurs outside the working memory, the processing speed (S) of working memory (in bits per second), the *capacity* (C) of working memory (in bits), and the *number* (N) of bits of information that have to be processed. Thus $\mathrm{RT} = \mathrm{f}(T_0, N, S, C)$. But the exact form of the equation has not yet been worked out.

If $N > C$, either processing must occur in stages, if possible, or there is a "breakdown" in performance. The variable of capacity (C) can be measured indirectly and expressed in real time units by the size of its effect on RT. When a given amount of information x is being processed in working memory and then is increased by the addition of information y to the STM system, the RT to x is increased, and the amount of increase in RT is inversely related to the capacity of the individual's working memory.

Studies of memory span for digits probably illustrate the capacity hypothesis most simply. For example, I have found that Berkeley students have a memory span of about eight digits on average, but only if they are presented no more than eight digits. If as many as twelve digits are presented, students typically can recall only the first three. But individuals who are perfectly matched on a digit span of, say, eight will show significantly different amounts of interference loss as a result of having to pay attention to the additional four digits in the twelve-item digit series (Jensen, 1965). It seems impossible to account for this phenomenon in terms of a single underlying process. An even more striking example is the difference between forward and backward digit span (FDS and BDS). The fact that FDS and BDS are not highly correlated even when corrected for attenuation indicates that at least two different processes are involved. The sheer storage requirements of the FDS and BDS tasks are identical. But because in BDS the set of digits has to be perfectly reversed between input and output, BDS makes a considerably greater processing demand on working memory than does FDS. BDS is always smaller and has a longer latency of response than FDS. Hence, both the capacity and speed of working memory are better measured by BDS than by FDS, and one index of the overall efficiency of working memory is the difference between BDS and FDS. As should be expected from our theory of the central role of working memory in information-processing, the *g* loading of BDS has been found to be about double that of FDS (Jensen and Figueroa, 1975).

The importance of processing speed in the operation of working memory stems directly from the capacity limitation and the rapid decay of information in STM. The limited capacity of the 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 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 primary memory, 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 on it that are required, 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 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 most discriminating test items are those that "threaten" the information-processing system at the threshold of overload or "breakdown"—the point that most fully reflects both the speed of operations and the capacity of working memory. In a series of items of graded complexity, this breakdown would occur at different points for various individuals. Processing time and probability of breakdown are closely related. For example, we (Jensen, Larson, and Paul, 1988) have found that in a set of extremely easy test items suitable for third-graders whose average failure rate on the items under an untimed testing condition is only about 17 percent, the different failure rates (that is, breakdown) on the various items are almost perfectly correlated with the response latencies on the same items obtained from young adults who were given each item as a reaction-time test. (The adults' error rates were almost nil.) Also, the adults' RTs on this test were significantly correlated with psychometric *g*.

Measurements of individual differences in the speed of elemental components can be obtained on tasks that are so simple as to make breakdown failure very unlikely, as in the various ECTs based on chronometric techniques in which RT is found to be correlated with scores on complex psychometric tests, such as the Wechsler Scales and the Raven Matrices. A faster rate of information-processing permits more information to be processed per unit of time, and since all knowledge and skills acquisition depends on information-processing, persons who process information faster acquire more knowledge and skill from a given amount of experience. One of the best researched ECTs measures an individual's speed of mentally scanning very recently input information in STM (Sternberg, 1966). These speeds are very fast, averaging in young adults about seventy-five milliseconds per each digit scanned, and these speeds are substantially correlated with *g* (Jensen, 1987d). Although individual differences in the extremely short RTs to ECTs seem very slight, often amounting to no more than a few milliseconds, they become of considerable consequence when multiplied over an extended period of time. For example, the seemingly slight but reliable differences in RTs to very simple ECTs between average and gifted children (even when the average and gifted are full siblings reared together) are correlated (at about age twelve) with quite large differences in amounts of general knowledge, vocabulary, and academic skills (Cohn, Carlson, and Jensen, 1985; Jensen, Cohn, and Cohn, 1989). These findings are consistent with the well-substantiated fact that the time required to learn scholastic subjects to a uniform criterion of mastery is highly correlated with IQ, that is, mainly with *g*.

Semantic memory (returning to figure 4–2) has practically unlimited capacity and is a reservoir of verbally or otherwise symbolically encoded information, including past-learned meanings, relationships, rules, and strategies for operating on certain classes of information (such as words, syntax, numbers and arithmetic operations, musical notation, chess combinations, and the like). Specific contents of semantic memory may be crucial for solving certain problems. The amount of information in semantic memory is a function of experience, interest, motivation, and learning opportunities, but mostly it is a function of the efficiency of working memory, which is the agency through which information becomes stored in semantic memory. Its usefulness to the individual, aside from the amount of information stored, depends on its accessibility to working memory. The speed of access to information in LTM is a joint function of the processing speed of working memory, the manner in which information is encoded and organized in semantic LTM, and the amount and recency of practice in retrieving the information. A number of ECTs, all variations on the well-known "Posner paradigm" (Posner, Boies, Eichelman, and Taylor, 1969), have been used to measure the speed with which individuals can access various kinds of information in semantic LTM. The speed of access even to extremely simple and highly over-learned verbal codes in LTM, such as the letters of the alphabet, is correlated with *g* and with verbal ability (independently of *g*) in university students (Hunt, 1976).

In any learning situation, working memory brings the products of past learning from LTM into conjunction with novel inputs to arrive at problem solutions or to encode and rehearse the perceived relationships of the "new" information to the "old" information in preparation for storing it in semantic LTM. Hence, the information content of LTM is cumulative, but the degree and system of organization of the stored information is an important determinant of its later accessibility and usability. This is obviously an aspect of the information-processing system in which instructional methods could play an important role.

Episodic LTM is of lesser interest in the present context. It is a store of non-semantically encoded spatial-temporal experiences. "Recognition memory" of faces and places and memories of sensory and emotional experiences and specific events and their contexts are classed as episodic memory. Some episodic memories can be tagged, encoded, or transformed by the working memory for storage in the semantic LTM.

The efficiency of the operations performed in a given individual's working memory is not a constant but can vary markedly according to the processing strategies adopted and the amount of practice the individual has had in processing a particular kind of information. Single bits of information can be organized or "chunked" into larger units, which can then be dealt with as single bits by the working memory. The net effect is equivalent to increasing the capacity of the individual's working memory. But this strategy seems to be less general in enhancing the total economy of information-processing than the phenomenon termed *automatic processing*, or *automatization*.

Cognitive theorists have made the important distinction between *controlled* and *automatic* processing, which some theorists regard not just as quantitatively different aspects of processing (for example, slow versus fast and hard versus easy), but as qualitatively different kinds of processing (Shiffrin and Schneider, 1977). The development of automatic processing through learning and practice, however, is necessarily always preceded by controlled processing (Shiffrin and Dumais, 1981).

Because automatic processing is mainly a function of learning and practice, it probably has greater implications for education than any other single aspect of cognitive theory.

Controlled Processing. Controlled processing of information demands the individual's focused attention, requires conscious mental effort, is relatively slow, and deals with information input sequentially, being able to deal with only very limited amounts of information at one time and being unable to execute different operations simultaneously. These of course are all recognized as the characteristics associated with working memory. In some circumstances the input demand on controlled processing may crowd the full capacity of working memory; any faster rate of input overloads the system and results in “breakdown” or “turn-off.” Solving novel problems, learning new knowledge or skills, and consciously monitoring an unpredictably changing situation that calls for varied responses all involve controlled processing. If everything we did mentally had to depend entirely on controlled processing, life would be intolerably burdensome indeed, and our efficiency would be greatly impaired. Fortunately, the evolution of the human brain has provided it with the means for escape from such a fate, namely, the development of automatic processing.

Automatic Processing. Practice, if sufficiently long-term, can automatize certain information-processing routines, which frees working memory for the controlled processing of other information. In contrast to controlled processing, automatic processing does not demand one’s entire attention; it is relatively effortless and can deal with large amounts of information and perform different operations on it simultaneously.

The degree to which task performance can become automatized depends on how consistent, predictable, or routine the information-processing demands of the task are. Automatization is easier the more consistent the required sequence of operations between input and output. In learning to send Morse code, for example, there is an invariant relationship between letters of the alphabet and their corresponding dot-and-dash codes. The act of sending and receiving messages becomes completely automatized for expert telegraphers.

Most skills, however, involve both controlled and automatic processing. Driving a car is a good example. In the early stage of learning to drive, controlled processing predominates. To minimize external distractions, the learner must practice in a quiet street. The learner’s full and undivided attention is required to execute smoothly the simultaneous operations of the clutch, the gear shift, the gas pedal, the steering wheel, and the brake, and also to remember to make the appropriate hand signals at the right times. While doing all this the learner finds it impossible to converse, listen to the radio, or think about other things, without risk of grinding gears, killing the engine, running off the road, or worse.

With more practice, driving skill becomes increasingly automatic. The seasoned driver performs all these operations without having to think about them at all. Controlled processing is still necessary, however, to deal with constantly changing traffic conditions. We have to relinquish conversation or other attention-demanding activity momentarily when traffic conditions change suddenly or look complicated and unpredictable. The working memory is briefly occupied to full capacity. That is controlled processing. If all of the driver's operational skills were not fully automatic, they would encroach on the capacity of working memory and thereby impair the efficiency of the controlled processing needed to get through the traffic crisis without a mishap.

A perfect example of the combined roles of controlled and automatic processing is sight-reading an unfamiliar piece of music—an essential requirement of professional orchestra players. The controlled processing aspect of this feat occupies a considerable part of the capacity of working memory, especially if the performer must play up-to-tempo and at the same time be highly responsive to the conductor's expressive signals. Yet it would be utterly impossible for controlled processing to accomplish this kind of performance were it not for the fact that in professional musicians both the reading of musical notation and its execution on their instruments are about 99 percent automatized. Scarcely any thought at all need be given to those aspects of the musical notation per se that normally demand so much of the novice's attention, or to the incredibly complex combinations of perfectly coordinated muscular movements required to produce the correct sequences of notes on a musical instrument.

Indeed, many complex skills can never be mastered at all without a high degree of automatization of many of its components, because just the absolutely irreducible demand on controlled processing alone takes up the full capacity of the working memory, making it necessary that other components of the skill occur automatically. A high degree of automatic processing is not just a greatly speeded-up form of controlled processing. It is most characterized by simultaneity of different processes and "pattern thinking." Duffers at chess, for example, think only one move ahead at most. Excellent chess players often think several moves ahead. But world-class chess masters work quite differently. Research on the nature of their skill in chess has discovered that they seldom think ahead at all. They instantly perceive a whole pattern on the chessboard, and the properties of the pattern largely dictate the optimal move in light of the player's particular strategy. Chess masters easily memorize entire chess games in terms of such patterns, much as we can recall a sentence we have just read, without any conscious attention to the sequence of all the individual letters it contains. Yet studies have shown that chess masters do not have an exceptional memory in general. Given various memory tests unrelated to chess, they perform on a par with most college students. The difference is that the chess master's LTM is extraordinarily well-stocked with chess rules, strategies, positions, combinations, and the like, which are automatically accessed the moment the chess master looks at a particular configuration of pieces on the chessboard. The phenomenon is akin to literacy in one's native language.

Although the speed of controlled processing and the capacity of working memory are of great importance because of their heavy contribution to variance in *g*, recent research on persons who show truly exceptional performance in any field indicates that the critical difference between them and the average run is not raw *g*, but depends essentially on a much greater than ordinary amount of automatization of certain knowledge and skills in the person's field of achievement (Ericsson and Crutcher, in press).

The road to automatization—apparently the only road—is practice, and plenty of it, accompanied by conscious effort to improve performance. Few people realize the exceeding amount and consistency of practice that recent studies have revealed to be the indispensable precursors of surpassing skill or expertise in any field. Paderewski, who routinely practiced the piano ten hours a day in his youth, when later acclaimed a genius, remarked, "Yes, and before I was a genius I was a drudge." (See references to this research literature in Ericsson and Crutcher, 1990.)

Automatization and *g*. Individual differences in the development of automatization are also probably related to *g*, or at least to that part of its variance associated with the processing speed and capacity of working memory. However, the most thorough discussions of the research on this topic I have found in the literature (Ackerman, 1986, 1987; Ackerman and Schneider, 1985) give no definitive evidence that automatization is related to *g*. But it seems to me almost inevitable that the development of automatization would be related to the efficiency of controlled processing in working memory, because that would govern the amount of information that could be processed per unit of practice time. It is common knowledge among music teachers, for example, that high IQ children acquire skill on a musical instrument faster than low IQ children. The nominally same amount of practice generally results in greater improvement in the brighter children, quite apart from a musical talent factor. So when future research eventually answers this question, it would be most surprising if no correlation were found between individual differences in automatization and psychometric *g*. But this in no way rules out the potential value of increasing automatization in some people who, for whatever reason, have not adequately automatized certain essential skills.

Motivation and Automatization. The development of automatic processing in a particular area is strongly related to how motivated the person is in that area. Attempting to improve students' motivation is always problematic, but it is especially so in connection with low ability. The reason is that the behavioral characteristics we recognize as indicating a high level of motivation for learning and practice in a particular subject are really just the predictable result of positive reinforcement. And one of the most fundamental "laws" in psychology is that positive reinforcement increases both the strength and the frequency of the reinforced behavior.

Generally, the most effective reinforcements in human learning result from the learner's immediate perception of the successes and failures in the learner's own performance, as the learner assesses it, in terms of its external consequences or by comparison with the performance of others or by subjective comparison with an internalized standard. One of the most important functions of a teacher is to set the standard or model of performance for the student to internalize. Both the student's performance and the effort that accompanies it are positively reinforced every time the student perceives he or she has made a closer approximation to the internalized standard. A fairly optimal schedule of such reinforcements results in the characteristics we recognize as a motivated student. The ablest individuals in any domain usually get a more optimal schedule of reinforcement and therefore become more motivated in the successful activity. It is a positive feedback loop. Just the opposite occurs for the least able learners. If one's attempts to learn something are insufficiently reinforced, with the consequent decrease in motivation, the person usually drops the learning activity with impunity, unless it involves an educationally or economically critical skill.

There are also individual differences in energy level. It is a general factor probably with a strong biological basis. Energy level interacts with motivation. But motivation itself is not a general factor, it is acquired in a particular context and remains connected with specific activities. Highly motivated performance sustains the most practice and therefore is more subject to automatization of those aspects for which controlled processing is not essential. Such automatization, like motivation, also creates a positive feedback in the person's further progress in the practiced domain. Thus, we see the magnifying effect of experience on individual differences in every kind of achievement that does not impose a low ceiling on performance.

Assessment of Automatization. Before explaining the importance of automatization in school learning, something must be said about the assessment aspects of it. Automatization is difficult to study experimentally because it develops only over extended periods of practice. For practical educational and diagnostic purposes, however, it is feasible to study the automatization of certain scholastic skills in which students are known to have received a given number of months or semesters of instruction. Little is known about individual differences in automatization tendency when practice conditions are held constant. But we do know that amount of experience, or practice, is a major determinant of automatization. So when we assess individual differences in the automatization of material learned in natural settings, such as school learning, we are probably measuring an amalgam of both intrinsic differences in tendency to automatization of learned material and differences in the amount of practice that has been devoted to it, whatever the cause of these differences may be.

As already explained, the automatization of knowledge and skill has mainly two results: (1) it greatly speeds up access to the needed information in LTM and (2) it frees some of the capacity of working memory. The assessment of automatization, therefore, must reflect both of these aspects. To make the discussion realistic, say we are concerned with the automatization of the simple "number facts" in arithmetic, for example, the multiplication of all the single digits one through nine with each other. (The importance of automatizing such simple skills is discussed in the following section.)

Automatization is meaningless unless some content to be automatized already exists. The first thing we must do, then, is to determine if the student actually knows the times tables, which can be revealed by a non-speeded paper-and-pencil test. Assuming we find that the student knows all the times tables, we then have to determine to what extent this knowledge has been automatized. Like much of medical diagnosis, the task involves interpreting a number of different clues. Chronometric techniques are essential for doing this. A baseline measurement is the student's median RT and intra-individual variability (SD of RT over trials) for the student's binary responses (True/False) to number discriminations that involve approximately the same amount of visual information as the multiplication facts, for example, $5 = 5 = 5$ (T) or $3 = 5 = 3$ (F). The very same RT variables are then obtained on the multiplication facts, for example, $3 \times 5 = 15$ (T) or $3 \times 5 = 18$ (F).

We have determined in our laboratory that median RTs on these kinds of tasks can be obtained with reliability coefficients above .90 in about twenty minutes of testing. In a group of age-matched students, there will be a moderate correlation between the median RTs on the two tasks just described. The regression of RT for multiplication on RT for simple number discrimination, then, affords one index of automatization. Students who place significantly above the regression line are suspected of not having automatized access to the multiplication facts as well as those who fall below the regression line. The slope of the regression line should decrease as a function of age and increasing skill in arithmetic, and a person's RTs can be viewed in relation to the regression line in various normative age and ability groups. The possibility that simple number recognition itself has not been well automatized can be investigated in the same way, from the regression of (1) the median RT in the simple number discrimination task on (2) the median RT in a binary choice RT task using visual stimuli that have no scholastic content, such as red versus green lights.

Intra-individual variability can be analyzed in the same way. Some students are very uneven in the degree of automatization of number facts, and this shows up in their much greater variability in RTs across different items.

The degree to which automatization of one skill frees the capacity of working memory for dealing with some other input can be assessed by means of a *dual task* procedure, which also depends on the measurement of RTs. It requires one task that inescapably demands controlled processing and occupies the working memory. A good example of such a task is the Semantic Verification Test (SVT). In the SVT a simple statement of the following type appears on the screen:

A before B

(The negative counterpart, A not before B, appears on half of the trials; also A after B, and A not after B are used.) After two seconds, the statement goes off, and after a one-second blank interval, the two letters appear side-by-side, A B (or B A), and the subject responds either True or False (by pressing buttons labeled T of F), indicating whether the order of the paired letters agrees or disagrees with the stem statement. There are considerable individual differences in the median RT. University students were found to have a 0.67 SD faster RT than age-matched navy recruits on a three-letter SVT for which median RT was correlated −.45 with the highly *g*-loaded Raven Matrices scores in university students (Jensen, Larson, and Paul, 1988). The SVT obviously makes a considerable demand on the working memory. The target task, say, is multiplication facts, for example, 5 × 4 = 20 (T). Median RTs (and SD of RT) are obtained on both the SVT and the target task, each administered separately. But then the two tasks are also combined as a dual task, in which the sequence of presentation is as follows:

A before B	(two seconds)
5 × 4 = 20 (T)	(RT in milliseconds)
B A (F)	(RT in milliseconds)

Interposing the target task between the stem and the reaction stimulus of the SVT has the effect of increasing the RT on both tasks, as compared to the RT when the tasks are given separately in the single-task conditions. The measured increment in the RTs is inversely related to the degree to which the retrieval of the information called for by the interposed task has been automatized.

These kinds of procedures and variations of them can be used to study individual differences in the elemental components of controlled and automatic processing of the kinds of information that constitute the basic knowledge and skills on which scholastic achievement beyond the elementary grades critically depends.

Controlled and Automatic Processing in Scholastic Skills

Proficiency in the three Rs depends on a high degree of automatization of numerous subskills, as does the learning of most other school subjects. At every step in learning complex subject matter, the consistent or routine aspects of the task must become automatic in order to free the working memory to cope with the novel aspects of the task, or to process new information coming from the teacher or the text. Much that is learned without becoming automatic turns out to be functionally inadequate when the learned material must serve as a prerequisite for more advanced learning. The failure to automatize certain subskills at one stage of learning can prove a severe handicap when the learner reaches a more advanced level. In some cases, more advanced learning is even impossible, because the task demands on controlled processing greatly exceed the capacity of working memory. Hence, progress is possible only if most of the task demands have already been automatized, leaving the working memory free to deal with only those novel aspects that can never become automatic. Experts differ from novices in (1) the extent of their task-relevant knowledge in LTM, (2) the way it is encoded and organized in LTM, and (3) the speed of its retrieval and use through automatic processing. It is mainly these three factors that account for experts' superior problem-solving skill and their ability to learn something new in their field of expertise much easier and faster than is possible for novices, regardless of their IQ.

Reading Comprehension. A prime example of these concepts of information-processing and automatization is reading comprehension. The act of reading is an incredibly complex process that depends on virtually every element in the entire information-processing system as well as the automatization of many subskills. Reading, in fact, is so complex and its information-processing demands are so great that it would be impossible if so much of it were not automatized. In good readers, the decoding (that is, letter and word recognition) processes are completely automatic. If decoding depended on controlled processing and usurped the reader's working memory, reading *comprehension* would be virtually impossible.

Most people fail to realize the degree of complexity involved in reading, such as the fact that a number of specific subskills have to be coordinated, because these subskills have become so automatized that good readers have lost all awareness of them—letter recognition, name retrieval, word decoding, and semantic access (that is, the automatic search and elicitation of meanings stored in LTM). In recent years, research on reading has become highly sophisticated. Information-processing conceptions and the use of chronometric techniques make it possible to observe and measure the different components of reading skill at an extremely fine-grained level (see, Carpenter and Just, 1981; Jackson and McClelland, 1979; LaBerge and Samuels, 1974; Lesgold and Perfetti, 1981; Perfetti, 1983; Waldron, 1987). Automatic processing, which is unconscious, is found to be much faster than controlled processing, and even automatic processes in reading that occur within 1/100th of a second can be specifically identified.

Some of these automatic processes can be isolated from the context of actual reading, to be subjected to analytical study of their general properties in the experimental laboratory. It is a telling fact that when a number of the subskills of reading are measured separately, the correlations between them are higher in a group of good readers (that is, high scorers on standardized reading tests) than in a group of poor readers. This suggests that the good readers have mastered each of the subskills at the automatic level, whereas the poor readers have automatized some of the subskills but not others. Chronometric methods have also revealed many other ways that good and poor readers differ. Poor readers show a slower speed of retrieving well-learned letter or name codes from LTM, and they are slower in semantic matching, that is, responding either same or different to a pair of familiar words depending on whether they are synonyms or antonyms—a test that reflects speed of access to the meanings of words and phrases stored in LTM. Poor readers also show a slower speed of initiating the pronouncing of pseudo-words; individual differences in this simple task were found to be correlated .68 with scores on a standard test of reading comprehension (Fredericksen, 1980).

Word-span is the number of unrelated words that can be perfectly recalled *in toto* immediately following a single presentation of a set of words at a given rate. Word-span shows only low to moderate correlations with reading comprehension scores on standard tests. The correlations are not higher evidently because passive memory span requires little controlled processing and involves the primary STM more than the working memory.

This conjecture led Daneman (1982) to invent an exceedingly simple but clever test, known as the *Reading Span Test.* The subject has to read aloud sets of two to five unrelated sentences, with the instruction that the last word in each sentence would have to be recalled after the set of sentences had been completed. The task is surprisingly difficult. Scores among university students range from two to five final words recallled, with a mean of about three. Obviously the controlled processing demands of reading these sentences with comprehension takes up much of the capacity of working memory, unlike the passive word-span test. But the most striking finding is that scores on the Reading Span Test are typically correlated about .7 with scores on standardized tests of reading comprehension. This indicates that even when the decoding process is highly automatized, reading comprehension depends heavily on the efficiency of working memory, which helps explain why tests of reading comprehension are just about as highly *g*-loaded as either verbal or nonverbal IQ tests in groups that have no trouble with the decoding aspect of reading.

Not surprisingly, the Reading Span Test is correlated almost as highly with tests of listening comprehension, because verbal comprehension makes much the same processing demands on working memory, regardless of whether the subject actively reads or simply listens to the material. Individual differences in either reading comprehension or listening comprehension arise largely from individual differences in the efficiency of working memory.

All these points (and others) are well illustrated in some excellent research in the armed sevices that separately measures the decoding and comprehension aspects of reading (Sticht, Hooke, and Caylor, 1981). But this research also reveals some important points that would be almost impossible to demonstrate in studies of reading based exclusively on university students. In 1976, some 50 percent of enlistees at an army base had reading levels at the fifth grade and below. In such groups with borderline literacy, measures of decoding skills measured independently of comprehension account for even more of the variance in scores on standard reading comprehension tests that does the efficiency of working memory. Thus, considerable improvement in the absolute level of reading comprehension could be achieved by training aimed at the automatization of decoding skills in such groups of poor readers. Persons who are poor readers because they have not fully automatized all of the decoding processes can be identified by means of a clever test of word decoding used by Sticht et al. (1981) and also by their showing significantly better listening comprehension than reading comprehension. Poor readers who show good listening comprehension are the most apt to benefit from training in decoding.

Arithmetic and Mathematics. These subjects have as yet received much less empirical study from an information-processing perspective than reading. Yet, they are nearly as important in the school curriculum, and they lend themselves ideally to conceptualization in information-processing terms and to analysis by chronometric techniques. (A good overview of current thinking in this area is provided by Briars [1983].) The learning of arithmetic and mathematics is hierarchical, that is, ease of learning at each higher level of complexity depends on prior mastery of more elementary skills. Such mastery depends heavily on the development of automatic processing if it is to promote more advanced learning.

Many pupils begin to experience unusual difficulty in learning arithmetic when they are in fourth to sixth grade, even when they have not evinced any real difficulty in earlier grades. Usually they can obtain perfect scores on non-speeded paper-and-pencil tests of elementary arithmetic requiring knowledge of simple number facts, such as addition, subtraction, and multiplication of single-digit numbers. Then, quite suddenly, in grades four to six, when more complex arithmetic operations and applications are introduced, such as short and long division, fractions, decimals, percentages, powers and roots, and words problems requiring these operations, some pupils experience inordinate difficulty. As a result, many are completely "turned off" to math and swell the ranks of adult innumerates. Such problems can be studied most fruitfully in the context of information-processing by means of chronometric techniques.

An obvious hypothesis from an information-processing standpoint is that the elementary skills may have been learned sufficiently to pass the ordinary tests of these skills, but they have not been sufficiently "overlearned" to be automatized. They, therefore, require too much controlled processing and usurp too much of the capacity of working memory whenever these elementary skills are needed to deal with more complex kinds of problems that make heavy demands on controlled processing. Without automatization of basic skills, the presentation of more advanced material that depends on them simply overloads the student's processing capacity, causing a breakdown in learning.

This probably holds true at every level of learning mathematics. The importance of rule automation for learning and transfer in algebra transformations and word problems is shown in a series of studies by Cooper and Sweller (1987). Their findings suggested training techniques that facilitated the development of automation (as they call it). Most interesting is their finding that the strongest effect of automation is its facilitation of *transfer* of the learning of one type of problem to different types of problems. Breadth of transfer is a well-known correlate of *g*. Individuals with low IQ show little transfer of specific skills they have learned to novel conditions. But automation of a skill is a relatively slow and prolonged affair compared to just learning the skill. What seems to have happened in the study by Cooper and Sweller is that as students automatized certain aspects of algebra, it freed their working memory and, in effect, conserved their *g*, which made for greater transfer. Transfer of training depends on the subject's analysis of the relatively novel transfer problem to find familiar features, and that requires controlled processing. This analytic process is hindered if the algebraic subskills relevant to the familiar features in the transfer problem have not been automatized.

Aims of Future Research

A disturbing feature of the contemporary research scene in educational psychology has been its conspicuous retreat from large-scale programmatic research on the cognitive aspects of children who are "at risk" for unacceptably low levels of achievement.

The tidal wave of studies of the educationally disadvantaged in the 1950s and 1960s was based mainly on sociological and behaviorist theories. The prevailing thoughts about educational deficit and compensatory education during that period largely ignored the body of knowledge and methods of differential psychology. Since the 1940s, this field had become a stagnant backwater in psychology. Its methodological legacy from such past luminaries as Galton, Spearman, Thorndike, and Thurstone became submerged in the more thriving field of psychometrics. With education's burgeoning research on the disadvantaged, fostered by abundant funding in the Kennedy-Johnson era, psychometrics unfortunately became a disfavored discipline and assumed a defensive and apologetic posture.

Information-processing theory and the revival of mental chronometry were scarcely more than mere seedlings at that time, so they could hardly have made an impact on educational research. At about the same time that these newer approaches to the study of cognitive abilities came fully into their own, in the 1970s and 1980s, there was a striking revival of resarch on most of the traditional problems of differential psychology. New journals and scientific societies sprung up devoted entirely to research on intelligence and behavioral genetics. These fields acquired a new look—more distanced from psychometrics and education and increasingly allied with experimental cognitive psychology, information-processing concepts, the methods of mental chronometry, and the neurosciences. It is hardly an exaggeration to say that there has been greater scientific ferment and progress in theory and research on the nature of human abilities in just the past decade than in all of the preceding half-century. And progress continues apace.

A peculiar thing has happened along the way, however—the earlier intense interest in the educationally disadvantaged rapidly dwindled and all but vanished. It was not taken up—or perhaps it was intentionally avoided—by the new school of cognitive researchers. Their studies are based largely on college students, scholastically mainstream students, and the gifted. Children with specific learning disabilities and the clinically retarded have also figured in some of this research. With the exception of studies of low-ability personnel in the armed services, however, present-day cognitive researchers have largely shunned the educationally "high-risk" schoolchildren who were the focus of so much research in the 1960s and 1970s. There are some scientifically legitimate reasons for this seeming avoidance, such as the need for experimental studies of cognitive processes per se, unencumbered by the social complications and controversy involved in studying population differences in scholastic performance.

The popular sociological and anthropological theories of educational disparity are too indirect and overarching to explain precisely the nature of children's achieving or failing to achieve in school. A quite different order of research and analysis are required to discover the mechanisms through which cultural and social factors, to whatever extent they may be involved, actually exert their effects on scholastic performance. The broadbrush concepts of sociology and anthropology seem unsuited for the level of analysis required.

If we are concerned with the fine grain of such questions, psychologists can probably best contribute by bringing the concepts and methods of information-processing research to bear on the study of children who sooner or later fail to benefit from schooling as we know it today.

Such research would not be just more of the same, which has been so discouraging in the past. Achievement is a complex product of different cognitive processes, each of which makes its contribution. A failure to learn what has been taught, or inordinate difficulty with some subject, or poor retention, or poor conceptual grasp—any of these deficits may be traced to one or more specific deficiencies in the information-processing system. For example, although studies have shown that the efficiency of the working-memory component accounts for most of the variance in reading comprehension among young adults, cognitive processing research on persons who were born deaf indicates that their mediocre verbal IQ and poor reading comprehension are not at all connected with the efficiency or capacity of their STM or working memory, but with a deficit of semantic codes in LTM—a remediable condition. Thus deficits in cognitive performance have to be understood at the process level in order to discover precisely what we can or cannot feasibly do about them.

Making such discoveries is probably possible with presently available methods. But if it is to be accomplished in the foreseeable future, it will require a concerted research effort by a great many of the most experienced investigators in this field, with financial support on a par with that of other major scientific missions, like discovering a cure for AIDS or deciphering the human genome.

The three classes of phenomena we still need to know much more about if we are to attempt educational innovation with good chances of success are the following:

1. Since psychometric *g* accounts for such a large part of the variance in scholastic performance, we need to know specifically which processing components are involved in *g* and the relative contributions each of these processes makes to the total variance in *g*. Going beyond mere correlation coefficients, we need to know how psychometric *g* is causally related to various kinds of achievement. Also, the design of research in this area must take account of the distinct possibility that the answers may not prove to be the same for different age brackets, for culturally different populations, or between the sexes. Scientifically compelling answers are not yet in our grasp.

2. We also need to develop a science and technology of cognitive process analysis for different types of achievement (for example, academic learning and vocational skills). This is an extension of traditional task analysis, which focuses on overt skills, to the underlying mental processes. The application of such analytical techniques to the study of reading, as indicated earlier, is a model for the study of many other elements of the curriculum at every level of education.

3. Finally, the presently most neglected subject of research: Focusing the process analysis of *g* and achievement directly on those segments of the school population that are now most predictably "at risk" for scholastic failure—mainly blacks and Hispanics in our inner-city schools. To coordinate such research with innovative educational experiments, it may be

necessary to bring a number of such "high-risk" schools under the auspices of university research departments organized for this purpose.

The cognitive process analysis of *g* and achievement, with their seemingly intractable variance in the school population, would yield knowledge that, if scientifically valid, could only turn out to be good news for education.

References

Ackerman, P. L. (1986) Individual differences in information processing: An investigation of intellectual abilities and task performance during practice, *Intelligence*, 10, 101–39.

Ackerman, P. L. (1987) Individual differences in skill learning: An integration of psychometric and information processing perspectives, *Psychological Bulletin*, 102, 3–27.

Ackerman, P. L. and W. Schneider. (1985) Individual differences in automatic and controlled information processing. In *Individual differences in cognition*, Vol. 2, ed. R. F. Dillon. New York: Academic Press.

Adams, M. J. (1989) Thinking skills curricula: Their promise and progress, *Educational Psychologist*, 24, 25–77.

Anderson, L. W. and R. B. Burns. (1987) Values, evidence, and mastery learning, *Review of Educational Research*, 57, 215–23.

Arvey, R. D. (1979) *Fairness in selecting employees*. Reading, MA: Addison-Wesley.

Atkinson, R. C. (1974) Teaching children to read using a computer, *American Psychologist*, 29, 169–78.

Baldwin, A. Y. (1985) Programs for the gifted and talented: Issues concerning minority populations. In *The gifted and talented: Developmental perspectives*, ed. F. D. Horowitz and M. O'Brien. Washington, DC: American Psychological Association.

Baldwin, A. Y. (1987) I'm black but look at me, I am also gifted, *Gifted Child Quarterly*, 31, 180–85.

Benbow, C. P. (1988) Neuropsychological perspectives on mathematical talent. In *The exceptional brain*, ed. L. K. Obler and D. Fein. New York: Guilford Press.

Benton, D. and G. Roberts. (1988) Effect of vitamin and mineral supplementation on intelligence of a sample of school children. *Lancet* No. 8578:140–43.

Bereiter, C. (1987) Jensen and educational differences. In *Arthur Jensen: Consensus and controversy*, ed. S. Modgil and C. Modgil. New York: The Falmer Press.

Bloom, B. S. (1987) A response to Slavin's mastery learning reconsidered, *Review of Educational Research*, 57, 507–8.

Brebner, J. and T. Nettelbeck. (1986) Intelligence and inspection time. Special issue, *Personality and Individual Differences*, 7, 603–729.

Briars, D. J. (1983) An information processing analysis of mathematical ability. In *Individual differences in cognition*, Vol. 1, ed. R. F. Dillon and R. R. Schmech. New York: Academic Press.

Brogden, H. E. (1946) On the interpretation of the correlation coefficient as a measure of predictive efficiency, *Journal of Educational Psychology*, 37, 65–76.

Brown, A. L. and J. C. Campione. (1982) Modifying intelligence or modifying cognitive skills: More than a semantic quibble? In *How and how much can intelligence be increased?*, ed. D. K. Detterman and R. J. Sternberg. Norwood: Ablex.

Campbell, J. T., Crooks, L. A., Mahoney, M. H. and D. A. Rock. (1973) *An investigation of sources of bias in the prediction of job performance, a six-year study*. Final Project Report PR-73–37. Princeton: Educational Testing Service.

Carnegie-Quarterly, (1987–1988) Black churches: Can they strengthen the black family? *CQ*, 33 (1), 1–9.

Carpenter, P. A. and M. A. Just. (1981) Cognitive processes in reading: Models based on readers' eye fixations. In *Interactive processes in reading*, ed. A. M. Lesgold and C. A. Perfetti. Hillsdale: Erlbaum.

Clark, R. E. (1982) Antagonism between achievement and enjoyment in ATI studies, *Educational Psychologist*, 13, 19–101.

Cohn, S. J., Carlson, J. S. and A. R. Jensen. (1985) Speed of information processing in academically gifted youths, *Personality and Individual Differences*, 6, 621–29.

Cohn, S. J., Cohn, C. M. G. and A. R. Jensen. (1988) Myopia and intelligence: A pleiotropic relationship? *Human Genetics*, 80, 53–58.

Cole, N. (1982) The implications of coaching for ability testing. In *Ability testing: Uses, consequences, controversies*, Part II, ed. A. K. Wigdor and W. R. Garner. Washington, DC: National Academy Press.

Cooper, G. and J. Sweller. (1987) Effects of schema acquisition and rule automation on mathematical problem-solving transfer, *Journal of Educational Psychology*, 79, 347–62.

Cory, C. H., Neffson, N. E. and B. Rimland. (1980) *Validity of a battery of experimental tests in predicting performance of Navy Project 100,000 personnel*. San Diego: Navy Personnel Research and Development Center.

Cronbach, L. J. and G. C. Gleser. (1965) *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L. J. and R. E. Snow. (1977) *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Crouse, J. and D. Trusheim. (1988) *The case against the SAT*. Chicago: University of Chicago Press.

Daneman, M. (1982) The measurement of reading comprehension: How not to trade construct validity for predictive power, *Intelligence*, 6, 331–45.

Detterman, D. K. (1987) What does reaction time tell us about intelligence? In *Speed of information-processing and intelligence*, ed. P. A. Vernon. Norwood:

Ablex.

Detterman, D. K. and R. J. Sternberg, (eds.) (1982) *How and how much can intelligence be increased*. Norwood: Ablex.

Ericsson, K. A. (1988) Analysis of memory performance in terms of memory skills. In *Advances in the psychology of human intelligence*, Vol. 4, ed. R. J. Sternberg. Hillsdale: Erlbaum.

Ericsson, K. A. and R. J. Crutcher. (1990) The nature of exceptional performance. In *Life-span development and behavior*, Vol. 10, ed. P. B. Baltes, D. L. Featherman, and R. M. Learner. Hillsdale, NJ: Lawrence Erlbaum.

Flynn, J. R. (1984) The mean IQ of Americans: Massive gains 1932 to 1978, *Psychological Bulletin*, 95, 29–51.

Flynn, J. R. (1987) The ontology of intelligence. In *Measurement, realism and objectivity*, ed. J. Forge. New York: Reidel.

Frederiksen, J. R. (1980) Component skills in reading: Measurement of individual differences through chronometric analysis. In *Aptitude, learning, and instruction*, ed. R. E. Snow, P. Federico, and W. Montagu. Hillsdale: Erlbaum.

Gates, A. I. and G. A. Taylor. (1925) An experimental study of the nature of improvement resulting from practice in mental function, *Journal of Educational Psychology*, 16, 583–93.

Gedye, C. A. (1981) *Longitudinal study (grades 1 through 10) of school achievement, self-confidence, and selected parental characteristics*. Doctoral dissertation, Dept. of Education University of California, Berkeley.

Gettinger, M. (1984) Individual differences in time needed for learning: A review of literature, *Educational Psychologist*, 19, 15–29.

Goldman, R. D. and R. E. Slaughter. (1976) Why college grade point average is difficult to predict, *Journal of Educational Psychology*, 68, 9–14.

Goldman, R. D., Schmidt, D. E., Hewitt, B. N. and R. Fisher. (1974) Grading practices in different major fields, *American Educational Research Journal*, 11, 343–57.

Gorsuch, R. L. (1983) *Factor analysis*. (2d ed). Hillsdale: NJ: Erlbaum.

Gottfredson, L. S. and J. Crouse. (1986) Validity versus utility of mental tests: Example of the SAT, *Journal of Vocational Behavior*, 29, 363–78.

Guilford, J. P. (1964) Zero correlations among tests of intellectual abilities, *Psychological Bulletin*, 61, 401–4.

Guskey, T. R. (1987) Rethinking mastery learning reconsidered, *Review of Educational Research*, 57, 225–29.

Gustafsson, J.-E. (1988) Hierarchical models of individual differences in cognitive abilities. In *Advances in the psychology of human intelligence*, Vol. 4, ed. R. J. Sternberg. Hillsdale: Erlbaum.

Haier, R. J., Robinson, D. L., Braden, W. and D. Williams. (1983) Electrical potentials of the cerebral cortex and psychometric intelligence, *Personality and Individual Differences*, 4, 591–99.

Hernstein, R. J., Nickerson, R. S., deSanchez, M. and J. A. Swets. (1986) Teaching thinking skills, *American Psychologist*, 41, 1279–89.

Horn, J. M., Loehlin, J. C. and L. Willerman. (1979) Intellectual resemblance among adoptive and biological relatives: The Texas Adoption Project, *Behavior Genetics*, 9, 177–207.

House, E. R., Glass, G. V., McLean, L. D. and D. F. Walker (1978) No single answer: Critique of the follow-through evaluation, *Harvard Educational Review*, 48, 128–60.

Humphreys, L. G. (1968) The fleeting nature of the prediction of college academic success, *Journal of Educational Psychology*, 59, 375–80.

Humphreys, L. G. (1973) Predictability of academic grades for students of high and low academic promise, *Educational and Psychological Measurement*, 33, 385–92.

Hunt, E. (1976) Varieties of cognitive power. In *The nature of intelligence*, ed. L. B. Resnick. Hillsdale: Erlbaum.

Hunter, J. E. (1986) Cognitive ability, cognitive aptitudes, job knowledge, and job performance, *Journal of Vocational Behavior*, 29, 340–62.

Hunter, J. E., Schmidt, F. L. and J. Rauschenberger. (1984) Methodological, statistical, and ethical issues in the study of bias in psychological tests. In *Perspectives on bias in mental testing*. C. R. Reynolds and R. T. Brown (eds.). New York: Plenum.

Ilg, F.L. and L. B. Ames. (1964) *School readiness*. New York: Harper and Row.

Jackson, M. D. and J. L. McClelland. (1979) Processing determinants of reading speed, *Journal of Experimental Psychology: General*, 108, 151–81.

Jensen, A. R. (1965) *Individual differences in learning: Interference factor*. Cooperative Research Project No. 1867. Washington, DC: U.S. Office of Education.

——. (1968) Patterns of mental ability and socioeconomic status, *Proceedings of the National Academy of Sciences*, 60, 1330–37.

——. (1975) The price of inequality, *Oxford Review of Education*, 1, 59–71.

——. (1980) *Bias in mental testing*. New York: Free Press.

——. (1982) Reaction time and psychometric *g*. In *A model for intelligence*, ed. H. J. Eysenck. Heidelberg: Springer-Verlag.

——. (1984) Test validity: *g* versus the specificity doctrine, *Journal of Social and Biological Structures*, 7, 93–118.

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

——. (1985b) The black-white difference in *g*: A phenomenon in search of a theory, *Behavioral and Brain Sciences*, 8, 246–63.

——. (1985c) Methodological and statistical techniques for the chronometric study of mental abilities. In *Methodological and statistical advances in the study of individual differences*, ed. C. R. Reynolds and V. L. Willson. New York: Plenum.

——. (1986) *g*: Artifact or reality?, *Journal of Vocational Behavior*, 29, 301–31.

——. (1987a) Psychometric *g* as a focus of concerted research effort, *Intelligence*, 11, 193–98.

——. (1987b) The *g* beyond factor analysis. In *The influence of cognitive psychology on testing and measurement*, ed. J. C. Conoley, J. A. Glover, and R.

R. Ronning. Hillsdale: Erlbaum.

——. (1987c) Further evidence for Spearman's hypothesis concerning black-white differences on psychometric tests, *Behavioral and Brain Sciences*, 10, 512–19.

——. (1987d) Individual differences in the Hick reaction time paradigm. In *Speed of information-processing and intelligence*, ed. P. A. Vernon. Norwood: Ablex.

——. (1987e) Process differences and individual differences in some cognitive tasks, *Intelligence*, 11, 107–36.

——. (1989a) The relationship between learning and intelligence, *Learning and Individual Differences*, 1, 37–62.

——. (1989b) Raising IQ without increasing *g*? A review of "The Milwaukee Project: Preventing mental retardation in children at risk," *Developmental Review*, 9 (3), 234–58.

Jensen, A. R. and R. A. Figueroa. (1975) Forward and backward digit-span interaction with race and IQ: Predictions from Jensen's theory, *Journal of Educational Psychology*, 67, 882–93.

Jensen, A. R. and C. R. Reynolds. (1982) Race, social class, and ability patterns on the WISC-R, *Personality and Individual Differences*, 3, 423–38.

Jensen, A. R., Cohn, S. J. and C. M. G. Cohn. (1988) Speed of information processing in academically gifted youths and their siblings, *Personality and Individual Differences*, 10, 29–34.

Jensen, A. R., Larson, G. E. and S. Paul. (1988) Psychometric *g* and mental processing speed on a semantic verification test, *Personality and Individual Differences*, 9, 243–55.

Kolata, G. (1987) Early signs of school age IQ, *Science*, 236, 774–75.

Kyllonen, P. C. (1986) *Theory-based cognitive assessment.* AFH Technical Paper 85–30. San Antonio: Manpower and Personnel Division, Brooks Air Force Base.

LaBerge, D. and S. J. Samuels. (1974) Toward a theory of automatic information processing in reading, *Cognitive Psychology*, 6, 293–323.

Lesgold, A. M. and C. A. Perfetti, (eds.) (1981) *Interactive processes in reading.* Hillsdale: Erlbaum.

Linn, R. (1982) Ability testing: Individual differences, prediction, and differential prediction. In *Ability testing: Uses, consequences, and controversies*, Part II, ed. A. K. Wigdor and W. R. Garner. Washington, DC: National Academy Press.

Loevinger, J. (1947) A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs*, 61, 1–49.

Maeroff, G. E. (1988) Withered hopes, stillborn dreams: The dismal panorama of urban school, *Phi Delta Kappan*, 69, 633–38.

Manning, W. H. and R. Jackson. (1984) College entrance examinations: Objective selection or gatekeeping for the economically priviledged. In *Perspectives on bias in mental testing*, ed. C. R. Reynolds and R. T. Brown. New York: Plenum.

Messick, S. (1982) Issues of effectiveness and equity in the coaching controversy:

Implications for educational and testing practice, *Educational Psychologist*, 17, 67–91.

Modgil, S. and C. Modgil. (1987) *Arthur Jensen: Consensus and controversy.* New York: The Falmer Press.

Naglieri, J. A. and A. R. Jensen. (1987) Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis, *Intelligence*, 11, 21–43.

National Commission on Excellence in Education. (1983) *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

Newell, A. and H. A. Simon. (1972) *Human problem solving.* Englewood Cliffs: Prentice-Hall.

Nickerson, R. S. (1988) On improving thinking through instruction, *Review of Research in Education*, 15, 3–57.

Nisbett, R. E., Fong, G. T., Lehman D. R. and P. W. Cheng. (1987) Teaching reasoning, *Science*, 238, 625–31.

Perfetti, C. A. (1983) Individual differences in verbal processes. In *Individual differences in cognition*, Vol. 1, R. F. Dillon and R. R. Schmeck (eds.). New York: Academic Press.

Posner, M. I. (1978) *Chronometric explorations of mind.* Hillsdale: Erlbaum.

Posner, M. I., Boies, S., Eichelman, W. and R. Taylor. (1969) Retention of visual and name codes of single letters. *Journal of Experimental Psychology*, 81, 10–15.

Reynolds, C. R. and R. T. Brown. (eds.) (1984) *Perspectives on bias in mental testing.* New York: Plenum.

Schafer, E. W. P. (1985) Neural adaptability: A biological determinant of *g* factor intelligence, *Behavioral and Brain Sciences*, 8, 240–41.

Schmid, J. and J. M. Leiman. (1957) The development of hierarchical factor solutions, *Psychometrika*, 22, 53–61.

Schmidt, F. (1988) The problem of group differences in ability test scores in employment selection, *Journal of Vocational Behavior*, 33, 272–92.

Shiffrin, R. M. and S. T. Dumais. (1981) The development of automatism. In *Cognitive skills and their acquisition*, ed. J. R. Anderson. Hillsdale: Erlbaum.

Shiffrin, R. M. and W. Schneider. (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory, *Psychological Review*, 84, 127–90.

Shockley, W. (1957) On the statistics of individual variations of productivity in research laboratories, *Proceedings of the IRE*, 45, 279–90.

Slavin, R. E. (1987a) Mastery learning reconsidered, *Review of Educational Research*, 57, 175–213.

Slavin, R. E. (1987b) Taking the mystery out of mastery: A response to Guskey, Anderson, and Burns, *Review of Educational Research*, 57, 231–35.

Snow, R. E. and D. F. Lohman. (1988) Implications of cognitive psychology for educational measurement. In *Education measurement.* 3d ed., ed. R. L. Linn.

New York: Macmillan.

Snow, R. E. and E. Yalow. (1982) Education and intelligence. In *Handbook of human intelligence*, ed. R. J. Sternberg. Cambridge: Cambridge University Press.

Spearman, C. E. (1904) "General intelligence" objectively determined and measured, *American Journal of Psychology*, 15, 201–93.

Spearman, C. E. (1927) *The abilities of man: Their nature and measurement.* London: Macmillan.

Spitz, H. H. (1986) *The raising of intelligence: A selected history of attempts to raise retarded intelligence*. Hillsdale: Erlbaum.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B. and T. R. Cervo. (1977) *A planned variation model. Vol. IV-A: Effects of follow through models*. Washington, DC: U.S. Office of Information.

Sternberg, S. (1966) High speed scanning in human memory, *Science*, 153, 652–54.

Sternberg, R. J. and K. Bhana. (1986) Synthesis of research on the effectiveness of intellectual skills programs: Snake-oil remedies or miracle cures?, *Educational Leadership*, 44, 60–67.

Sternberg, R. J. and D. K. Detterman. (eds.) (1986) *What is intelligence?* Norwood: Ablex.

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

Sticht, T. G., Hooke, L. R. and J. S. Caylor. (1981) *Literacy, oracy, and vocational aptitude as predictors of attrition and promotion in the armed services.* HumRRO.FR-ETSD-81–11. Alexandria: Human Resources Research Organization.

Sticht, T. G., Armstrong, W. B., Hickey, D. T. and J. S. Caylor. (1987) *Cast-off youth: Policy and training methods from the military experience*. New York: Praeger.

Thorndike, R. L. (1984) *Intelligence as information processing: The mind and the computer*. Bloomington: Center on Evaluation, Development and Research.

Thorndike, R. L. (1987) Stability of factor loadings, *Personality and Individual Differences*, 8, 585–86.

Thurstone, L. L. (1947) *Multiple factor analysis*. Chicago: University of Chicago Press.

Vernon, P. A. (1981) Level I and Level II: A review, *Educational Psychologist*, 16, 45–64.

Vernon, P. A. (ed.) (1987) *Speed of information-processing and intelligence.* Norwood: Ablex.

Waldrop, M. M. (1987) The workings of working memory, *Science*, 237, 1564–67.

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

Wigdor, A. K. and W. R. Garner. (eds.) (1982) *Ability testing: Uses, consequences, and controversies. Part 1: Report of the committee; Part 2: Documentation section*. Washington, DC: National Academy Press.

Wilson, R. S. (1983) The Louisville Twin Study: Developmental synchronies in behavior, *Child Development*, 54, 298–316.