Chapter 8

Validity and Correlates of Mental Tests

Validity is the most central concept in the whole testing enterprise. It is the main goal toward which reliability and stability are aimed. However elegantly a test may be constructed in terms of all the other essentials of psychometrics, without validity it comes to naught. *A test's validity is the extent to which scientifically valuable or practically useful inferences can be drawn from the scores.*

From this most general definition of validity, it is obvious that validity is a complex concept requiring further analysis and explication. There are four main types of validity, and the demonstration of a test's validity may be based on any one or any combination of these, easily remembered as the four *C*'s: *content* validity, *criterion* validity, *concurrent* validity, and *construct* validity.

Content Validity

This type of validity is most relevant to achievement tests, job-knowledge tests, and work-sample tests. A test has content validity to the extent that the items in the test are judged to constitute a representative sample of some clearly specified universe of knowledge or skills. This judgment is usually based on a consensus of experts in the field of knowledge or skill that the test items are intended to sample. For example, if it is a test of general musical knowledge (not musical talent), the items, in the judgment of musicians, would have to represent a sufficiently broad and varied selection of factual information about music: notation, musical vocabulary, orchestral instruments, music history and theory, composers, and so on. Any musician examining the test should be able to agree that it is a test of musical knowledge. The test then would be said to show good content or face validity. Tests of knowledge about specific jobs are evaluated in terms of content validity. So are work-sample tests, which are performance tests consisting of a representative sample of the kinds of skills that analysis of a particular job reveals a person must actually possess to perform adequately on the job, for such occupations as typist, computer programmer, welder, electrician, and machinist, among others.

Specific aptitude tests (clerical, mechanical, musical, etc.), often aim for content validity, but it is not a crucial feature in such tests, which must depend mainly on other types of validation. The kinds of items that measure an aptitude need not closely resemble the final kind of performance for which the aptitude is a prerequisite. It may be possible to

measure a person's musical aptitude before he has learned to play an instrument or before he has had any kind of training in music. The musical aptitude test, tapping such elemental capacities as discrimination and short-term memory of pitch, loudness, duration, timbre, and rhythmic patterns, would obviously have some degree of content validity, but its validation would have to rest chiefly on criterion validity, that is, the correlation of the aptitude test scores with assessments of later success in musical training. The aptitude test itself may predict but does not sample the criterion performance, knowledge, or skills acquired through training.

Criterion Validity

This is the ability of test scores to predict performance in some endeavor that is external to the test itself, called the *criterion*. A test's validity coefficient is simply the correlation between the test scores and measurements of the criterion performance. For example, a college aptitude test would be said to have good criterion validity (also called *predictive validity*) to the extent that the test scores are correlated with grades in college (the criterion).

The criterion performance may be measured by other tests (e.g., scholastic achievement tests and job-knowledge tests), by grades in courses, by supervisor's ratings of performance on the job, or by direct indices of work proficiency and productivity, such as the number of articles assembled per hour, number of sales per month, number of pages typed per hour, and the like.

Criterion validity is probably the most important, defensible, and convincing type of validation in the practical use of psychological tests. In many cases it is regarded as crucial. Increasingly in recent years the use of tests in educational and employment selection can be justified only in terms of the test's criterion validity. It is a reasonable and scientifically and economically defensible requirement benefiting both the student and the school, the job applicant and the employer.

Criterion validity does not rest on the content or statistical analysis of the test per se, or on expert opinion or testimony, but depends entirely on empirical demonstration. This consists of establishing the correlation between the test scores and some clearly specified and quantified criterion. The criterion itself, of course, must be open to critical scrutiny.

Analysis of the criterion, in fact, is often the first step in the procedure of establishing a test's criterion validity. The criterion performance itself is systematically examined and analyzed to determine the kinds of abilities, knowledge, and skills that it involves, as a means to formulating hypotheses as to the kinds of test items that would most likely predict the criterion. A test is constructed accordingly and is then tried out. In a test to predict performance in a clerical job, for example, it would seem reasonable to include items that measure perceptual speed and accuracy, knowledge of alphabetizing, tabulation, and the like.

Once a test is thus selected or designed for predicting a particular criterion, one of four procedures is used for determining the test's validity coefficient (i.e., correlation with the criterion):

Method 1. The most completely satisfactory method is to test all the job applicants but not use the test scores in any hiring decisions. The scores are then later correlated with

measurements or ratings of success in training or job performance. It is important that the scores be kept secret so as not to risk contaminating the criterion measures used in determining the test's validity for the particular criterion.

This method has many statistical advantages over all the others, because it enables one to estimate accurately the degree of improvement in selection decisions that can result from using the test as compared with whatever other basis for selection was being used. If the test shows a significant validity coefficient (i.e., correlation between test scores and the criterion measure), one can determine the most suitable cutoff score for selecting applicants in such a way as to maximize success in the criterion performance, which may be successfully completing a course of training, proficiency on the job, job satisfaction, or probability of qualifying for promotion to a higher level job that requires the experience of the hiring-in job.

The only disadvantages of this method are that (1) it takes more time than other methods, because of the necessary interval between initial testing and the later assessment of performance, and (2) no direct benefit in employee selection can be gained from the test results in the first validation group of applicants, as their scores are not used in hiring, and, if there is some prior evidence of the test's validity for similar criteria, this information can have no effect on the initial hiring decisions.

Method 2. The second method gets around these disadvantages of the first method. Employees already on the job are tested and their scores are correlated with assessments of their job performance. This correlation is best described as a *restricted* validity coefficient. It is valuable information, but it usually underestimates the true validity of the test because of the restriction of range of test scores and of the criterion measure. Present employees to some extent have already been selected for success on the job. The least successful have quit or been fired. If there is a correlation between test scores and job performance, we may presume that the on-the-job selection process by and large would have eliminated persons with lower test scores. When the range of predictor scores and criterion measures is thus restricted, the correlation between them is necessarily shrunken and may greatly underestimate the validity that the test would have if it were used in the initial selection of job applicants. The effect of restriction of range is seen most strikingly in highly selective colleges. If all applicants with a high school diploma were indiscriminantly admitted to the college, the college aptitude test scores would correlate very highly with grades and persistence to graduation, assuming that the college maintains its academic standards. But, when admissions are limited only to students who earned excellent grades in high school and obtained high scores on the college aptitude test, the scores will show relatively little correlation with success in college.

In general, a test that has been used as the basis for selection cannot then be adequately validated on the selected group. This is especially true when there is a nonlinear relationship between test scores and the criterion, as when the test score acts as a threshold variable, discriminating well between those who fail and those who succeed, but not discriminating well between varying degrees of success. This is often the case when success on the job depends on a number of traits, each of which is necessary but not sufficient and only one of which is measured by the predictor test. For example, a person with poor pitch discrimination will not succeed as a violinist, whatever other assets he may possess, and so unsuccessful violin students could be confidently predicted by a test

of pitch discrimination alone. But prediction of degree of success at the violin would be only very weak for pupils with good pitch discrimination, because talent for the violin involves many other aptitudes and personal qualities as well.

Method 3. The third method is usually used only in choosing or designing a suitable test to predict a particular criterion, but it may be the sole method of test validation when selection of applicants is clearly essential but validity studies of the first two types are not feasible. Selecting the first astronauts is a good example of the necessity of careful selection of personnel in the absence of any validation of the selection tests in terms of later actual performance.

The best method in such a case is systematic analysis of the job or criterion performance into its various component knowledge and skills and selection (or construction) of validated tests of these components of the criterion performance. For example, in the judgment of psychologists trained in "task analysis," the job may require perceptual speed, motor coordination, and the capacity quickly to grasp and interpret numerical information presented simultaneously on an instrument control panel. The selection battery then will include tests of these abilities and any others that the job analysis suggests are important in successful performance, including perhaps certain physical attributes and personality traits.

What cannot be determined in this type of analysis are the ideal weights that should be assigned to each of the component measures to achieve maximal validity in predicting final performance. (Determining the ideal weights to give every subtest in the total predictor score is completely possible by method 1 and, to a limited extent, by method 2.)

Also, it should be emphasized that selecting predictor tests on the basis of job analysis is really a psychologically sophisticated art involving experience and good judgment concerning the higher-order cognitive abilities required by the person to integrate the more obvious subskills revealed by job analysis. The capacity to integrate a number of subskills effectively may be a more important source of individual differences in the criterion performance than are individual differences in the separate subskills per se. This kind of general integrative capacity is usually best measured by tests that are highly loaded in what we have termed g, such as most tests of general intelligence.

Method 4. Another method, which is better viewed merely as an aid to finding potentially useful selection tests rather than as a means of validating them, is to look for tests that show differences, on the average, between successful employees in different kinds of jobs (or students in different kinds of colleges). If the average test scores differ significantly for persons who are successful in different jobs, it is evident that the test measures factors relevant to job selection, success, and possibly satisfaction and persistence in the job. If test scores do not show significant differences, on the average, between persons in quite different jobs, it is less likely that they will have substantial validity for predicting success in any specific job.

Multiple Prediction

Very often in practice a validity coefficient is based on a *multiple correlation* (symbolized as R) rather than on a simple correlation (r) between a single test and the criterion. A multiple correlation is the correlation between (1) a best-weighted composite score from a number of different tests (called the predictor variables) and (2) the criterion. If the two or more different predictor variables in the composite are well chosen, the

multiple correlation, R, of the composite score with the criterion may be appreciably larger than the simple correlation, r, of any one of the predictor tests with the criterion.

The predictor tests that work best in combination are those that are not highly correlated with one another but are each separately correlated significantly with the criterion. Even if each test separately has only a quite moderate or even low correlation with the criterion, the tests in combination may correlate very substantially with the criterion, provided that the tests are not highly correlated among themselves. For a multiple correlation to be worthwhile, one needs a number of predictor tests that do not overlap too much in the abilities that they measure. Thus each test measures some aspect of ability relevant to the criterion that is not measured by any of the other tests in the combination of predictors.

The scores on each of the several predictor tests are combined into a composite score in such a way as to maximize the multiple correlation between the composite score and the criterion. The statistical device for achieving this is called a multiple regression equation. The method, which is mathematically quite complex, is explicated in most statistical textbooks. The main aim of the method is to determine precisely the optimum values (called *regression coefficients*) by which to weight each of the predictor scores so as to make the composite score (i.e., the sum of the separate weighted predictor scores) have the highest correlation with the criterion.

The multiple correlation is rarely increased significantly by adding in more tests beyond the first few, because there is diminishing likelihood that any new test added to the composite will measure any appreciable part of the criterion-relevant abilities that are not already included in the first few tests. The statistical technique of multiple regression, in fact, permits the investigator optimally to select from a very large number of tests the few tests that are capable of maximizing the prediction of the criterion. It has been amply demonstrated that even the most expert human judgment and intuition cannot compete with the precision of the multiple regression equation as a means for choosing and weighting the final combination of tests (or other variables) that can best predict any given criterion (Meehl, 1954).

A multiple R validity coefficient is interpreted in exactly the same way as a simple r validity coefficient.

Concurrent Validity

This label has been used in the testing literature to refer to two quite distinct types of validity. It is always confusing in science when different concepts are given the same label. The remedy is simply to redefine or delimit old definitions and consistently stick to the new definition.

"Concurrent validity" traditionally has referred to (1) the correlation between a test and a criterion when both measurements are obtained at nearly the same point in time (as when a scholastic aptitude test and scholastic achievement test are administered on the same day or within a few days) and (2) the correlation between a new, unvalidated test and another test of already established validity.

The first case is really a form of criterion validity, that is, a correlation between a test and a criterion. A test's criterion validity can be studied as a function of the temporal interval between giving persons the test and measuring their criterion performance. The

temporal interval between test and criterion should be an essential part of reporting the criterion validity of any test. The interval over which a test will predict a criterion, and with how much precision, is a wholly empirical matter. One could label criterion validity *predictive* criterion validity when there is a reasonably long interval between the test and the criterion, but this would be merely an arbitrary rather than a conceptual distinction, as the interval between test and criterion is a continuous variable.

Therefore, the term concurrent validity should be used only to refer to the second case, that is, the correlation of a previously unvalidated test with an already validated test.

There are dangers in this type of validation. The risk is perhaps least when the unvalidated test is merely a parallel form or shortened version of the validated test, as when only a few of the eleven subscales of the Wechsler Intelligence Scale are used to determine the Full Scale IQ.

Concurrent validation may be resorted to, with greater risk, by finding a shorter, more efficient test, or one that is easier to administer, that can be shown to correlate highly with a much longer or more cumbersome test of established validity for the criterion of interest. Thus group-administered tests are sometimes validated against tests that require individual administration.

Concurrent validity rests on the soundness of the inference that, since the first test correlates highly with the second test and the second test correlates with the criterion, the first test is also correlated with the criterion. It is essentially this question: If we know to what extent A is correlated with B, and we know to what extent B is correlated with C, how precisely can we infer to what extent A is correlated with C? The degree of risk in this inference can be best understood in terms of the *range* within which the actual criterion validity coefficient would fall when a new test is validated in terms of its correlation with a validated test. Call the scores on the unvalidated test U, scores on the validated test V, and measures on the criterion C. Then r_{VC}, the correlation between V and C, is the *criterion validity* of test V; and r_{UV}, the correlation between U and V, is the *concurrent validity* of test U. The crucial question, then, is what precisely can we infer concerning r_{UC}, that is, the probable criterion validity of test U?

If we know r_{VC} and r_{UV}, the upper and lower limits of the possible range of values of r_{UC} are given by the following formulas:

$$\text{Upper limit of } r_{UC} = r_{VC}r_{UV} + \sqrt{r_{VC}^2 r_{UV}^2 - r_{VC}^2 - r_{UV}^2 + 1}$$

$$\text{Lower limit of } r_{UC} = r_{VC}r_{UV} - \sqrt{r_{VC}^2 r_{UV}^2 - r_{VC}^2 - r_{UV}^2 + 1}$$

It may come as a sad surprise to many to see how very wide is the range of possible values of r_{UC} for any given combination of values of r_{VC} and r_{UV}. The ranges of r_{UC} are shown in Table 8.1, from which it is clear that concurrent validity inspires confidence only when the two tests are very highly correlated and the one test has a quite high criterion validity. Because it is rare to find criterion validities much higher than about .50, one can easily see the risk in depending on coefficients of concurrent validity. The risk is greatly lessened, however, when the two tests are parallel forms or one is a shortened form of the other, because both tests will then have approximately the same factor composition, which means that all the abilities measured by the first test that are correlated with the criterion

Table 8.1. Upper and lower limits of the possible range of criterion validity coefficients (r_{UC}) for test U, when the criterion validity of test V is r_{VC} and the concurrent validity of test U is r_{UV}.

r_{UV} (or r_{VC})	r_{VC} (or r_{UV})								
	.10	.20	.30	.40	.50	.60	.70	.80	.90
.95	.41	.49	.58	.67	.75	.82	.89	.95	.99
	−.21	−.11	−.01	.09	.20	.32	.44	.57	.72
.90	.52	.61	.69	.76	.83	.89	.94	.98	1.00
	−.34	−.25	−.15	−.04	.07	.19	.32	.46	.62
.85	.61	.69	.76	.82	.88	.93	.97	1.00	.99
	−.44	−.35	−.25	−.14	−.03	.09	.22	.36	.53
.80	.68	.75	.81	.87	.92	.96	.99	1.00	.98
	−.52	−.43	−.33	−.23	−.12	.00	.13	.28	.46
.75	.73	.80	.85	.91	.95	.98	1.00	1.00	.96
	−.58	−.50	−.41	−.31	−.20	−.08	.05	.20	.39
.70	.78	.84	.89	.93	.97	.99	1.00	1.00	.94
	−.64	−.56	−.47	−.37	−.27	−.15	−.02	.28	.32
.65	.82	.87	.92	.96	.98	1.00	1.00	.97	.92
	−.69	−.61	−.53	−.44	−.33	−.22	−.09	.06	.25
.60	.85	.90	.94	.97	.99	1.00	.99	.96	.89
	−.73	−.66	−.58	−.49	−.39	−.28	−.15	.00	.19

also exist in the second test. The two tests should thus have fairly comparable correlations with the criterion, which is a necessary inference to justify concurrent validity.

Construct Validity

Although criterion validity is the most important kind of validity in the practical use of tests, *construct validity* is the most important from a scientific standpoint. The idea of construct validity is more difficult to explain. It concerns our attempt scientifically to understand, in psychological terms, what the test measures. For criterion validity we need not have this understanding. If the test predicts the criterion, that is all we need to know for it to be potentially useful in educational and vocational counseling and personnel selection. We can use the test's criterion validity to advantage without ever needing to understand what it involves in psychological terms.

Construct validity becomes a consideration as soon as we have some theory (or "construct") as to the psychological nature of the trait that we wish to measure. A theoretical formulation in psychology, as in any other science, is a formal set of propositions about the nature of something, from which we can logically deduce certain consequences or hypotheses, given certain conditions. A hypothesis, usually in the form of a prediction of what will happen under certain specified conditions, can be put to an empirical test to determine its truth or falsity. The theory from which the hypotheses are

derived is progressively modified by the results of these empirical tests of the hypotheses. The greater the range and variety of the hypotheses that are borne out by methodologically sound investigations, the more credence we have in the theory. This of course assumes that the tested hypotheses are really proper deductions from the theory itself.

A theory about a psychological trait calls for a great deal of this kind of hypothesis-testing research. First, it is necessary to determine whether such a trait can even be claimed to exist, because it is possible to posit a trait (i.e., a more or less consistent and enduring constellation of behavioral tendencies) and not be able to adduce any objective evidence of its existence. The next step is to discover the psychological nature of the trait.

A test devised to measure the trait is said to show construct validity if the test predicts the behavior in specific situations that would be deduced from our theory of the trait. For example, if our theory of intelligence involves the idea of an ability to deal effectively with complexity in any form, we might then hypothesize that an intelligence test should be a better predictor of performance on complex jobs than on simpler jobs. We could then ask a group of judges to rank a number of jobs in terms of their complexity. Finally, we would correlate our intelligence tests with employees' performance ratings in these various jobs. If the correlation increases as a function of the job's rank order in judged complexity, we would say that the intelligence test scores behave as our theory of intelligence should predict. Such a finding would be evidence for the test's construct validity as a measure of intelligence.

The task of construct validation is never really completed. A test's construct validity is further enhanced by every such theoretical prediction that is borne out in fact.

Factor analysis is another means of demonstrating a test's construct validity. If the factors that emerge in a factor analysis of a large battery of measurements are unambiguous and well established, and a new test has a high loading on one of the factors, the test is said to show *factorial validity*. This is a form of construct validity, because factors may be viewed as theoretical constructs used to explain the sources of individual differences in a variety of psychological measurements.

Construct validity is most important for tests that claim to measure some broad psychological trait and for which the demonstration of validity in terms of the test's correlation with any *single* criterion would be either inadequate or impossible.

The term "face validity" is frequently heard in discussions of tests, but it is actually a misnomer, since it has only a subjective and incidental relationship to the other forms of validity, that is, the four *C*'s. It refers to the degree to which the test items give the *appearance,* in the eyes of the person taking the test or of the person interpreting the test scores, of being a reasonable and appropriate indicator of what the test is supposed to measure.

Such appearance may or may not be related to the actual validity of the test, which is a matter for empirical determination. However, face validity could influence a person's attitude toward the test and affect his effort and test performance. Especially in the domains of ability and achievement, it is advantageous for the test items to have the appearance of being reasonable and appropriate questions for what the test purports to assess. Tests or test items that fail to meet this condition are said to have poor face validity. Such items make easy targets for popular ridicule. No test, of course, can stand on mere face validity alone; but it is possible that an objectively valid test may be scorned if it is deficient in face validity. Test makers now are paying more attention to this public

relations aspect of people's attitudes toward tests, which depend to a considerable extent on the test's face validity.

The Interpretation of a Validity Coefficient

There are three main ways of interpreting a validity coefficient. Each gives a view from a different perspective. The perspective of choice depends on one's purpose in using the validity coefficient. The three types of interpretations of validity are (1) improvement over chance prediction of a point estimate, (2) prediction of odds for success, and (3) prediction of the efficiency of performance.

Improvement over Chance Prediction. This interpretation of validity is based on the standard error of estimate, which we have encountered before (p. 281), as $SE_{est} = \sigma_c\sqrt{(1 - r_{xc}^2)[(N - 1)/(N - 2)]}$, where σ_c is the standard deviation of all the criterion measures, r_{xc} is the validity coefficient, that is, the correlation between test scores (x) and the criterion (c), and N is the sample size. (The expression $(N - 1)/(N - 2)$ may be omitted from the formula when $N > 200$.) SE_{est} is the standard deviation of the actual criterion measures around the *predicted* values of the criterion (i.e., the regression line). One can see from the formula that, if the validity is zero, SE_{est} is equal to σ_c and therefore that using the test does not make for better than chance prediction. The percentage of improvement in accuracy of prediction by using the test scores, over mere chance prediction, therefore, is equal to $100[1 - (SE_{est}/\sigma_c)]$. We can rewrite this formula solely in terms of the validity coefficient: $100\ (1 - \sqrt{1 - r_{xc}^2})$. This value is known as the *index of forecasting efficiency*.

It is important to note that the *index of forecasting efficiency* (IFE) tells us the percentage of improvement over chance in the accuracy of prediction when we are predicting *a specific point on the scale of criterion measurement*. To predict the *precise* value of an individual's criterion performance requires very high validity if the prediction is to be a marked improvement over chance. (The best *chance* prediction is the same value for every person and is the mean of the distribution of all criterion measurements.) To predict 50 percent better than chance (i.e., an IFE of 50 percent), we would need a validity coefficient of .866. In practice, however, most validity coefficients do not exceed .50 or .60, and usually are even lower. Yet a validity coefficient of .60 corresponds to an IFE of only 20 percent.

In terms of the index of forecasting efficiency, then, a realistic validity coefficient does not look very impressive, and validities below about .50 would hardly seem worth considering. But recall that the IFE concerns the accuracy of predicting a *specific point* on the criterion scale. This is an extremely stringent demand on any prediction test, and one that, in practice, is almost never required. For example, we are much more often concerned with predicting who will succeed or fail on a particular criterion, but we are not so concerned with predicting the specific rank order in performance on the criterion among those who fail or among those who succeed. If we can predict that Bill and Sue will both easily succeed on the job, we may not care, as far as our selection procedure is concerned, whether Bill or Sue performs better.

Therefore, when we are not primarily concerned with the accuracy of prediction at every single point throughout the entire range of measurement of the criterion, the index of forecasting efficiency is a much too stringent interpretation of the validity coefficient; it

grossly underestimates the benefit gained from using the test in selection. For this reason the index of forecasting efficiency is very seldom considered in the practical use of tests in educational and personnel selection. It does not tell the test user what he or she usually most wants to know. For this we turn to the following more practical interpretations of validity.

Prediction of Odds for Success. In the practical use of tests for selection we are usually concerned with the test's accuracy of predicting success or failure on the criterion. The dividing line between success and failure is defined in terms of some level of performance on the criterion, such as ability to obtain passing marks in a course of training or to perform on a job in a way deemed satisfactory by the employer. The test user tries to determine the best *cutoff score* on the predictor test (or combination of tests) to maximize the selection of applicants who will prove successful, given the constraints of (1) the number of persons who can be selected and (2) the total number of applicants. The ratio of condition 1 to condition 2 is termed the *selection ratio*. It is expressed as a proportion and is a crucial factor in the interpretation of a test's validity coefficient. The one other crucial parameter in the interpretation of validity is the *failure rate* on the criterion performance among selectees who were selected entirely without reference to their scores on the test in question. If there are no failures on the criterion, there is of course no need for the selection test in the first place. The test could not improve on whatever selective factors were already in effect. It should be remembered that applicants for any particular college or occupation are usually an already highly self-selected group with respect to the relevant requirements. The crucial question is how much the use of a test will improve selection (in terms of decreasing the failure rate) over and above whatever other selective factors are already operating.

The optimal cutoff score on a selection test with *perfect* validity would divide the pool of applicants into two groups: (1) those who score above the cutoff and *succeed* on the criterion (called *positives*) and (2) those who score below the cutoff and *fail* on the criterion (called *hits*). But when the test's validity is less than perfect, there will be created two other groups of applicants: (3) those who score above the cutoff but fail on the criterion (called *misses*) and (4) those who score below the cutoff but succeed on the criterion (called *false positives*). A test is regarded as valid to the extent that it minimizes the proportions of misses and false positives (and, conversely, maximizes the proportion of selectees who succeed on the criterion) for any given selection ratio (i.e., the proportion of the applicant pool that can be selected).

Taylor and Russell (1939) have devised a set of tables that take all these parameters into account to show the proportion of successes that should result from the use of a selection test with a given validity. One of the Taylor–Russell tables is presented in Table 8.2. The figures in the body of the table are the proportion of successes that would be expected for a given test validity and a given selection ratio, when the proportion of successes without selection on the basis of test scores is .60. The selection ratio determines the location of the cutoff score: the smaller the proportion of applicants that one needs, the higher can be the cutoff score on the selection test, and the greater is the payoff (in terms of the proportion of successful selectees) for any given validity coefficient.

A greater appreciation of the practical meaning of a given validity coefficient may be had by converting the proportions of success in the Taylor–Russell tables into the predicted odds for success when the test is or is not used. If the success rate without test

Table 8.2. Proportion of successes expected through the use of a test of given validity, when the proportion of successes is .60 without use of the test. (From Taylor & Russell, 1939, p. 576)

	Selection Ratio										
Validity	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
.05	.64	.63	.63	.62	.62	.62	.61	.61	.61	.60	.60
.10	.68	.67	.65	.64	.64	.63	.63	.62	.61	.61	.60
.15	.71	.70	.68	.67	.66	.65	.64	.63	.62	.61	.61
.20	.75	.73	.71	.69	.67	.66	.65	.64	.63	.62	.61
.25	.78	.76	.73	.71	.69	.68	.66	.65	.63	.62	.61
.30	.82	.79	.76	.73	.71	.69	.68	.66	.64	.62	.61
.35	.85	.82	.78	.75	.73	.71	.69	.67	.65	.63	.62
.40	.88	.85	.81	.78	.75	.73	.70	.68	.66	.63	.62
.45	.90	.87	.83	.80	.77	.74	.72	.69	.66	.64	.62
.50	.93	.90	.86	.82	.79	.76	.73	.70	.67	.64	.62
.55	.95	.92	.88	.84	.81	.78	.75	.71	.68	.64	.62
.60	.96	.94	.90	.87	.83	.80	.76	.73	.69	.65	.63
.65	.98	.96	.92	.89	.85	.82	.78	.74	.70	.65	.63
.70	.99	.97	.94	.91	.87	.84	.80	.75	.71	.66	.63
.75	.99	.99	.96	.93	.90	.86	.81	.77	.71	.66	.63
.80	1.00	.99	.98	.95	.92	.88	.83	.78	.72	.66	.63
.85	1.00	1.00	.99	.97	.95	.91	.86	.80	.73	.66	.63
.90	1.00	1.00	1.00	.99	.97	.94	.88	.82	.74	.67	.63
.95	1.00	1.00	1.00	1.00	.99	.97	.92	.84	.75	.67	.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.86	.75	.67	.63

selection is .60, then the predicted odds in favor of any person's succeeding is in the ratio of .60 to .40, or 1.5 to 1, regardless of the selection ratio. But say that we use a test with a validity of .40 and the selection ratio is .30. Then the predicted odds in favor of any selectee succeeding will be in the ratio of .78 to .22, or 3.5 to 1. In other words, the use of a test with a validity of .40 in this situation would increase the odds in favor of success by 3.5 to 1.5, or 2.33 times greater than the odds in favor of success if we had not used the test. The higher the validity and the lower the selection ratio, the more one can increase the odds in favor of success by using the test in selection. With a validity of .60 (which is about the top validity for college entrance exams) and a selection ratio of .10, the odds in favor of succeeding are almost 16 to 1, which is more than ten times better odds than if no selection test had been used. It can be seen that *test validity gains greater potency as the selection ratio becomes more stringent.*

Another set of tables has been devised to show the probability of success for persons selected from different deciles in the distribution of scores on the selection test (Wesman, 1953). (Deciles divide the total frequency distribution equally into tenths, going from the lowest 10 percent of the scores, decile 1, to the highest 10 percent, decile 10.) Table 8.3 shows that percentage of successes for persons selected from different deciles when the failure rates are 20 percent, 30 percent, or 50 percent and the test validity (r) ranges from .30 to .60.

From Table 8.3 we can determine the predicted odds in favor of success for persons scoring in any given decile on the selection test. When the overall failure rate is 20 percent

Table 8.3. Percentage of successful persons in each decile on test score. (From Wesman, 1953)

Standing on the Test		When the Total Percentage of Failures is 20%, and				When the Total Percentage of Failures is 30%, and				When the Total Percentage of Failures is 50%, and			
Percentile	Decile	$r = .30$	$r = .40$	$r = .50$	$r = .60$	$r = .30$	$r = .40$	$r = .50$	$r = .60$	$r = .30$	$r = .40$	$r = .50$	$r = .60$
90–99	10	92%	95%	97%	99%	86%	91%	94%	97%	71%	78%	84%	90%
80–89	9	89	91	94	97	81	85	89	92	63	68	73	78
70–79	8	86	89	91	94	78	81	84	88	59	62	65	69
60–69	7	84	86	88	91	75	77	80	83	55	57	59	61
50–59	6	82	84	85	87	72	74	75	77	52	52	53	54
40–49	5	80	81	82	83	70	70	70	71	48	48	47	46
30–39	4	78	77	77	78	67	66	65	64	45	43	41	39
20–29	3	75	73	72	71	63	61	59	56	42	38	35	31
10–19	2	71	68	64	61	59	55	50	45	37	33	28	22
1–9	1	63	56	49	40	50	43	35	27	29	23	16	10

and the test validity is .60, for example, the odds in favor of success for persons in the 10th decile are 99 to 1, as compared with the odds of 0.67 to 1 for persons in the 1st decile. The 10th decile's chances of success are 148 times greater than the 1st decile's chances. If no test were used, the odds for all persons would be 4 to 1 that they would succeed. Notice how markedly the odds are increased by selecting persons in the higher deciles. Even when the test validity is as low as .30, the odds favoring success for selectees in the tenth decile are 11.5 to 1, which is almost three times better odds than if no selection test were used. If one views test validity as would a gambler trying to find the best odds to maximize the payoff on his or her bets, even a test with a quite low validity would yield odds favoring the gambler that are impressively better than the base rate of success if no selection test were used. Gamblers would all be billionnaires if they could predict half as well as do tests with validities even much lower than .30. It is also worth noting that the typical validity coefficients of psychological tests compare quite favorably with the reliability of medical diagnoses, which is near .40 (Cronbach, 1960, p. 349).

Validity as a Proportional Increase in Criterion Performance. Another interpretation of validity is in terms of the average level of productivity, proficiency, or other index of performance that would result from using a selection test of a given validity, as compared with not using the test. Brogden (1946) has proved the following relationship:

$$r_{xc} = \frac{S - U}{P - U},$$

where

r_{xc} = the test's validity (i.e., the correlation r between the test scores x and the criterion measures c),
S = the mean level of performance of the persons who were selected on the basis of test scores,
U = the mean level of performance of persons who were selected at random (i.e., without the aid of the test), and
P = the mean level of performance of *perfectly* selected persons, as would be the case if $r_{xc} = 1$.

For example, say that we were selecting salesmen and that the performance criterion is the salesman's average number of sales per month over a period of one year. And say that we need to hire 50 salesmen from a pool of 200 job applicants. We could then do the following experiment. First, give the test to all the applicants. Then, draw 50 applicants at random; they are group U, that is, unselected by the test. Then, from the remaining 150 applicants we would select the 50 with the highest scores on the test; they are group S. Then we employ all 200 applicants and determine the average monthly sales of group U and of group S, which are, say, 40 and 30 sales per month, respectively. Finally, we determine the average monthly sales of the 50 salesmen who actually turned out to have the highest sales records; they are group P, that is, the 50 best applicants who hypothetically would have been chosen if we had used a *perfect* selection test. Say that their average sales are 60 per month. The validity of our selection test, then, can be expressed as the proportional improvement in employee performance over the base level resulting

from using the test, as compared with a hypothetical test of perfect predictive validity. Thus, in the example,

$$\text{Validity} = \frac{S - U}{P - U} = \frac{40 - 30}{60 - 30} = \frac{1}{3},$$

which is to say that employee selection by means of this particular test increases the average level of performance $33\frac{1}{3}$ percent over what it would be without the use of the test. This is numerically the very same validity coefficient that is usually defined as the correlation between test scores and criterion measures. The $33\frac{1}{3}$ percent improvement in employee performance or productivity, which in this example could be achieved by hiring the 50 applicants who scored highest on a test having a validity of only .33, would not be regarded as a trivial gain by most employers.

Brown and Ghiselli (1953) have expressed criterion performance as a standard score scale with a mean of zero and standard deviation of one when there has been no selection of applicants. The effects of selection (for any given selection ratio) by means of a test (with any given validity coefficient) can then be expressed in terms of the mean increase in criterion performance (in standard score units) over what it would be if the test were not used. These values computed by Brown and Ghiselli are presented in Table 8.4. For example, if one hired the highest-scoring 50 percent of applicants on the selection test (i.e., a selection ratio of .50), and, if the test's validity coefficient is .25, then the average level of job performance of these selected employees would be 0.2 standard deviations higher than the average performance level of an unselected group. But notice that, even if we had *perfect* validity in selection and a selection ratio of .50, the average performance of the selected employees would be 0.8 standard deviations higher than that of an unselected group. Thus, selection by means of our test (with a validity of only .25) results in an average employee performance gain over unselected employees that is 25 percent as good as the maximal selection procedure could possibly yield. Again, as with Brogden's formula, the *test's validity* ($\times$ 100) *may be interpreted as the average percentage gain in criterion performance resulting from use of the test in selection.*

Factors Influencing Validity

One may wonder why there is a "prediction ceiling" such that validity coefficients seldom exceed .50 or .60 and in most cases are considerably lower. Several main factors have been found to influence the validity coefficient.

Criterion Reliability. Very often the criterion is not measured with sufficient precision or consistency to permit any other variable to correlate with it highly. The highest possible validity coefficient cannot exceed the square root of the reliability of the criterion measurements. The criterion, when consisting of grades or ratings, often has considerably lower reliability than the predictor test itself. Considering the reliabilities of both the test and the criterion, the highest possible validity coefficient is the square root of the product of the two reliabilities, that is, $\sqrt{r_{tt} \times r_{cc}}$.

Restriction of Range. The possible size of the correlation between any two variables is affected by the range or spread of scores on each variable. For example, we would expect to find a much lower correlation between height and weight among the

Table 8.4. Mean standard criterion score of selected persons in relation to test validity and selection ratio. (From Brown & Ghiselli, 1953, p. 342)

Selection Ratio	Validity Coefficient																				
	.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
.05	.00	.10	.21	.31	.42	.52	.62	.73	.83	.94	1.04	1.14	1.25	1.35	1.46	.156	.166	.177	1.87	1.98	2.08
.10	.00	.09	.18	.26	.35	.44	.53	.62	.70	.78	.88	.97	1.05	1.14	1.23	1.32	1.41	1.49	1.58	1.67	1.76
.15	.00	.08	.15	.23	.31	.39	.46	.54	.62	.70	.77	.85	.93	1.01	1.08	1.16	1.24	1.32	1.39	1.47	1.55
.20	.00	.07	.14	.21	.28	.35	.42	.49	.56	.63	.70	.77	.84	.91	.98	1.05	1.12	1.19	1.26	1.33	1.40
.25	.00	.06	.13	.19	.25	.32	.38	.44	.51	.57	.63	.70	.76	.82	.89	.95	1.01	1.08	1.14	1.20	1.27
.30	.00	.06	.12	.17	.23	.29	.35	.40	.46	.52	.58	.64	.69	.75	.81	.87	.92	.98	1.04	1.10	1.16
.35	.00	.05	.11	.16	.21	.26	.32	.37	.42	.48	.53	.58	.63	.69	.74	.79	.84	.90	.95	1.00	1.06
.40	.00	.05	.10	.15	.19	.24	.29	.34	.39	.44	.48	.53	.58	.63	.68	.73	.77	.82	.87	.92	.97
.45	.00	.04	.09	.13	.18	.22	.26	.31	.35	.40	.44	.48	.53	.57	.62	.66	.70	.75	.79	.84	.88
.50	.00	.04	.08	.12	.16	.20	.24	.28	.32	.36	.40	.44	.48	.52	.56	.60	.64	.68	.72	.76	.80
.55	.00	.04	.07	.11	.14	.18	.22	.25	.29	.32	.36	.40	.43	.47	.50	.54	.58	.61	.65	.68	.72
.60	.00	.03	.06	.10	.13	.16	.19	23	.26	.29	.32	.35	.39	.42	.45	.48	.52	.55	.58	.61	.64
.65	.00	.03	.06	.09	.11	.14	.17	.20	.23	.26	.28	.31	.34	.37	.40	.43	.46	.48	.51	.54	.57
.70	.00	.02	.05	.07	.10	.12	.15	.17	.20	.22	.25	.27	.30	.32	.35	.37	.40	.42	.45	.47	.50
.75	.00	.02	.04	.06	08	.11	.13	.15	.17	.19	.21	.23	.25	.27	.30	.32	.33	.36	.38	.40	.42
.80	.00	.02	.04	.05	.07	.09	.11	.12	.14	.16	.18	.19	.21	.22	.25	.26	.28	.30	.32	.33	.35
.85	.00	.01	.03	.04	.05	.07	.08	.10	.11	.12	.14	.15	.16	.18	.19	.20	.22	.23	.25	.26	.27
.90	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20
.95	.00	.01	.01	.02	.02	.03	.03	.04	.04	.05	.05	.06	.07	.07	.08	.08	.09	.09	.10	.10	.11

players on a professional basketball team than in the general population, in which there is much greater variability in height and weight. If applicants are already highly self-selected on job-relevant characteristics, then a predictor test cannot show as high a validity coefficient as it would in an unselected group. Validity coefficients may even be reduced to zero if the test scores themselves are the basis for stringent selection and then are correlated with the criterion grades or performance ratings. If the test has high predictive validity and succeeds in eliminating applicants who would not excel on the criterion, then we should expect to find a low correlation between test scores and criterion performance in a highly selected group. The correlation between test scores and criterion in such a case is not a proper estimate of the test's validity. Yet many reported validity coefficients are determined on test-selected groups and consequently underestimate the test's potential validity.

Criterion Contamination. In determining validity, it is essential that the criterion ratings be made "blind," that is, without the rater's having any knowledge whatsoever of the ratee's score on the selection test. Such contamination usually inflates the validity coefficient.

Variable Criteria. The criterion itself may not actually be the same for all selectees, which may drastically lower the validity coefficient. Examples of this are grades in school or college. Grading standards differ from one teacher to another, from one course to another, and from one department to another. A grade of A in one course may be equivalent to a C in another, in terms of the level of aptitude and effort required. Also, weaker students tend to seek out the easy courses and the teachers reputed to have the most lenient grading standards. These factors all work against high correlations between scholastic aptitude test scores and grades.

A similar effect occurs in ratings of job performance when the rated employees have somewhat different duties that may make for different rating standards.

Training versus Final Level of Performance. It is a consistent finding that validity coefficients are higher for predicting success in training than for predicting the final level of performance reached after training and experience on the job. In a thorough survey of the reported validities of many tests in numerous and varied training and employment settings, Ghiselli (1966, p. 125) concluded: "Taking all jobs as a whole . . . it can be said that by and large the maximal power of tests to predict success in training is of the order of .50, and to predict success on the job itself is of the order of .35. . . ." The range of the average validity coefficients was .27 to .59 for training criteria and .16 to .46 for job proficiency criteria. Abilities that are important in the early stages of learning a new set of skills, and that may be measured well by aptitude tests, are sometimes of much less importance in the performance of the skills once they are learned and have become well practiced. The training phase may involve skills, such as reading ability, that are not even a part of the final job. In some cases, quite different tests may be needed to predict speed of progress in training and final level of job performance, because these two aspects can involve different abilities and personality characteristics.

Usually, new learning makes greater demands on *g* than does the final practiced performance, and most aptitude tests have a fairly substantial *g* loading. As a simple example, think of when you first learned to drive an automobile. It took all your concentration and considerable mental effort to recall and coordinate in sequence all the correct actions that were explained by the instructor. The ease and quickness with which the learner "catches on" could be predicted to some extent by a *g*-loaded paper-and-pencil

test. But with prolonged practice the task of driving becomes less and less mentally demanding. One's final level of driving performance will be more related to personality factors and very specific aptitudes (such as developing a "feel" for the performance capabilities of the vehicle itself) than to the *g* of mental ability per se. And so it is with training on many jobs. If the training itself is not too extensive or costly to the employer, and if trainability is not highly correlated with the final level of proficiency, one may question the use of selection tests that predict only success in training.

In predicting success in training, the validity of the predictor test will depend also on the form of instruction. Success in training that emphasizes abstract and conceptual understanding of what is being taught will be more predictable from *g*-loaded tests than will training that emphasizes practice on specific component skills and imitative learning in an apprenticeship fashion. In one training program, for example, changes in the methods of instruction reduced the validity of a mathematical aptitude test in predicting success in training from better than .50 to approximately zero, although under the modified instruction, those who were low in mathematical aptitude took 50 percent longer to complete the training. Most of these low-math-aptitude trainees would have failed the course altogether under the old method of instruction (Ford & Meyer, 1966).

Correlates of IQ and *g*-Loaded Tests

It would be practically impossible to review all the published evidence on the validity of all types of mental tests, even in a volume twice the length of this book. Validity coefficients are highly specific to the particular test, the particular criterion with which the test is correlated, and the particular population involved. Generally the best sources of information on the validity of any published test are the publisher's test manual and the *Mental Measurements Yearbook* edited by Oscar K. Buros. The *MMY* contains detailed, critical reviews, often by two or more expert reviewers, of all published psychological tests. Reviewers usually pay particular attention to the nature, extent, and statistical quality of the evidence for the test's validity. Readers seeking descriptive and evaluative information on any specific test are urged to consult the *Mental Measurements Yearbook,* of which seven large volumes have appeared, along with a new volume devoted exclusively to critical reviews of all published intelligence tests.

Rather than try to summarize the evidence for the validity of numerous specialized tests, I will confine the following review to typical examples of the wide variety of well-established correlates of intelligence tests or highly *g*-loaded tests. All intelligence or IQ tests are highly *g* loaded, but not all highly *g*-loaded tests are labeled as intelligence tests. In recent years test publishers have often substituted new labels for intelligence tests, such as tests of "cognitive ability," "general aptitude," and "learning potential." Most all such tests are found to be highly *g* loaded when factor analyzed along with other tests and are virtually indistinguishable from traditional intelligence tests in item content, statistical properties, and correlations with external criteria.

IQ has more behavioral correlates than any other psychological measurement. The external correlates of IQ are an empirical fact that must be recognized regardless of one's theoretical position regarding the existence or nature of intelligence, the causes of individual differences in IQ, or the causes of the correlations between IQ and other behavioral criteria. We may gain further insights into the nature and importance of IQ by surveying

Table 8.5. Correlations between various standard intelligence tests reported in the literature. (Data from Buros, 1972, Vol. I; Matarazzo, 1972, pp. 245–246; Sattler, 1974, pp. 125, 155, 236–246, Appendix B)

Tests	Correlations[1]
Wechsler–Bellevue I ×	
Stanford–Binet (1937)	.62, .86, .89
Raven Progressive Matrices	.55
Army Alpha	.74
Army General Classification Test	.83
Kent EGY	.65, .69
Shipley–Hartford	.72, .76
Thorndike CAVD	.69
Otis	.73
Wechsler Adult Intelligence Scale ×	
Stanford–Binet	.40–.83 (.77)
Raven Progressive Matrices	.53, .72, .83
SRA Nonverbal	.81
Army General Classification Test	.74
Army Beta (Revised)	.37, .82, .83
Ammons Picture Vocabulary	.76–.84 (.83)
Peabody Picture Vocabulary	.86
Kent EGY	.70, .77
Shipley–Hartford	.73–.86 (.77)
Otis	.78
Thurstone Test of Mental Alertness	.62
Wechsler Intelligence Scale for Children ×	
Stanford–Binet (47 studies)	.43–.94 (.80)
Columbia Mental Maturity Scale	.50–.76 (.64)
Draw-a-Man	.04–.59 (.36)
Raven Progressive Matrices	.27–.91 (.15)
Quick Test	.35–.84 (.41)
Peabody Picture Vocabulary Test	.30–.84 (.63)
Pictorial Test of Intelligence	.65, .71, .75
Slosson Intelligence Test	.50–.84 (.67)
Hiskey–Nebraska Test of Learning Aptitude	.82
Stanford–Binet ×	
Peabody Picture Vocabulary (37 studies)	.22–.92 (.66)
Pictorial Test of Intelligence	.38–.78 (.69)
Columbia Mental Maturity Scale	.39–.87 (.71)
Slosson Intelligence Test	.60–.94 (.90)
Cooperative Preschool Inventory	.39–.65
Hiskey–Nebraska Test of Learning Aptitude	.78–.86
Kahn Intelligence Test	.62, .75, .83
California Test of Mental Maturity	.66–.74
Peabody Picture Vocabulary Test ×	
Pictorial Test of Intelligence	.77
Columbia Mental Maturity Scale	.53
A Variety of (24) Other Ability Tests (not including WISC and S–B)	.06–.90 (.53)
Pictorial Test of Intelligence ×	
Columbia Mental Maturity Scale	.53
Leiter International Performance ×	
S–B and WISC (8 studies)	.56–.92 (.83)

Table 8.5 (*continued*)

Tests	Correlations[1]
Lorge–Thorndike × Analysis of Learning Potential	.83
Academic Alertness × Wonderlic Army Beta AGCT	.69–.88

[1]Where more than three correlations are reported, only the range and median (in parentheses) are indicated. In some cases in the literature, the range of correlations is reported but not the median.

the variety and characteristics of the many variables that show a correlation with IQ. Measurements of human individual differences that show few or negligible correlations with other aspects of life are usually of little or no interest. There are marked and highly reliable individual differences, for example, in fingerprints and form of the outer ear, but, because these features show correlations with hardly anything else, they are of no interest to anyone, except as a reliable means of identification. The great and persistent interest in IQ, on the other hand, is a direct result of the readily perceived and undeniable fact that IQ is correlated with so many other variables that are deemed important in life by almost everyone.

Concurrent Validity of IQ Tests. How well do scores on different IQ tests agree with one another? Do different IQ tests measure one and the same intelligence? There are hundreds of studies of correlations between various IQ tests. Table 8.5 shows a compilation of reported correlations between some of the most well-known standardized tests of intelligence.

It can be seen that the correlations range widely, with an overall mean of +.67. Many studies have been summarized in terms of the total range of correlations (i.e., the lowest and highest r's that are found in any of the studies) and the median value of the entire set of correlations (indicated in parentheses in Table 8.5). The mean of the median values is +.77. The mean of all the lower values of the range of correlations is +.50, and the mean of all the higher values of the range is +.82. Thus the correlations among various IQ tests can be said to be most typically in the range from about +.67 to +.77. The lower limit of the range of correlations between certain tests is often the result of studies based on small samples or on atypical groups, such as retardates, psychiatric patients, college students, or other groups with a restricted range of scores. Correlations are generally higher in studies based on representative samples of the general population. Also, some of the tests showing the lowest correlations with other tests (e.g., "Draw-a-Man" and the "Quick Test") may be questioned as measures of intelligence even on the basis of other psychometric criteria than their poor correlations with a quite good test of intelligence such as the WISC.

Correlations between IQ tests in the range from .67 to .77 are just about what one should expect if the g loadings of most IQ tests range from .80 to .90 and the tests have little variance other than g in common. The reader may recall from Chapter 6 that the correlation between any two tests can be expressed as the sum of the products of the tests' loadings on each of the common factors. By far the largest common factor in IQ tests is g.

Tests with *g* loadings in the .80 to .90 range, therefore, would show intercorrelations ranging from .64 to .81. Other common factors, such as verbal ability, would tend to raise the correlations only slightly. The fact that the median correlation between the Wechsler Intelligence Scale for Children and the Stanford–Binet in forty-seven studies is .80 suggests that these two tests have *g* loadings of close to .90 (i.e., $\sqrt{.80}$), which is only slightly less than the reliabilities of these tests (i.e., about .95).

It should be remembered that the correlation between tests indicates mainly the degree to which persons maintain the same relative standing on the various tests. A high correlation does not guarantee that the IQ scores themselves will be alike on every test. It is often noticed that even though individuals remain in very much the same rank order on two different IQ tests, meaning there is a high correlation between the tests, the actual IQ scores may be quite discrepant on the two tests. The discrepancies in the two IQs may show up consistently throughout the whole range, or they may differ in direction and magnitude in the lower, middle, and upper ranges of the IQ scale. Hence the various IQ scales themselves, although they may be highly correlated, are not exactly equivalent in an absolute sense. In this respect mental testing is currently in the situation similar to the measurement of distance and weight before the adoption of uniform international standards of measurement. Unfortunately, at present we have no standard IQ test corresponding to the platinum meter bar that is kept in the International Bureau of Standards in Paris.

The most common causes of the IQ scale discrepancies among various intelligence tests are the following:

1. The tests were standardized on somewhat different populations, with different absolute means or different standard deviations, or both.
2. The IQ scales were arbitrarily assigned different standard deviations. For example, the standard deviation of IQ on the Wechsler scales is 15 and on the Stanford–Binet it is 16.
3. The IQ is a standardized score on one test and on another is derived from the MA/CA ratio (which results in a variable standard deviation at different ages).
4. The IQ scores of one or both tests are not on an equal-interval scale throughout the whole range.
5. The factorial composition of the two tests is not quite the same, at all levels of difficulty. Scores in the high, medium, or low range may be more *g* loaded on the one test than on the other, even though both tests overall are equally *g* loaded.

Scholastic Achievement

In the seventy years since the publication of the Binet–Simon intelligence scale there have been thousands of studies of the correlation between intelligence test scores and scholastic performance. So generally consistent and statistically incontestable are these massive results that even the harshest critics of mental testing wholly concede the substantial relationship between IQ and scholastic achievement. They even exaggerate the relationship and claim that IQ differences reflect *nothing but* differences in educational advantages, in school, and in the home. In general, no other single fact that we can determine about a child after the age of 5 better predicts his or her future educational

progress and attainments than the IQ. Children with higher IQs generally acquire more scholastic knowledge more quickly and easily, get better marks, like school better, and stay in school longer.

A detailed review of all the evidence cannot be attempted here. Readers who wish to pursue the evidence further are referred to reviews by Cattell and Butcher (1968, Ch. 3), Lavin (1965, Ch. 4), Matarazzo (1972, Ch. 12), and Tyler (1965, Ch. 5). The evidence, however, is fairly easy to summarize, because there is now so much of it that a number of firm generalizations quite clearly emerge.

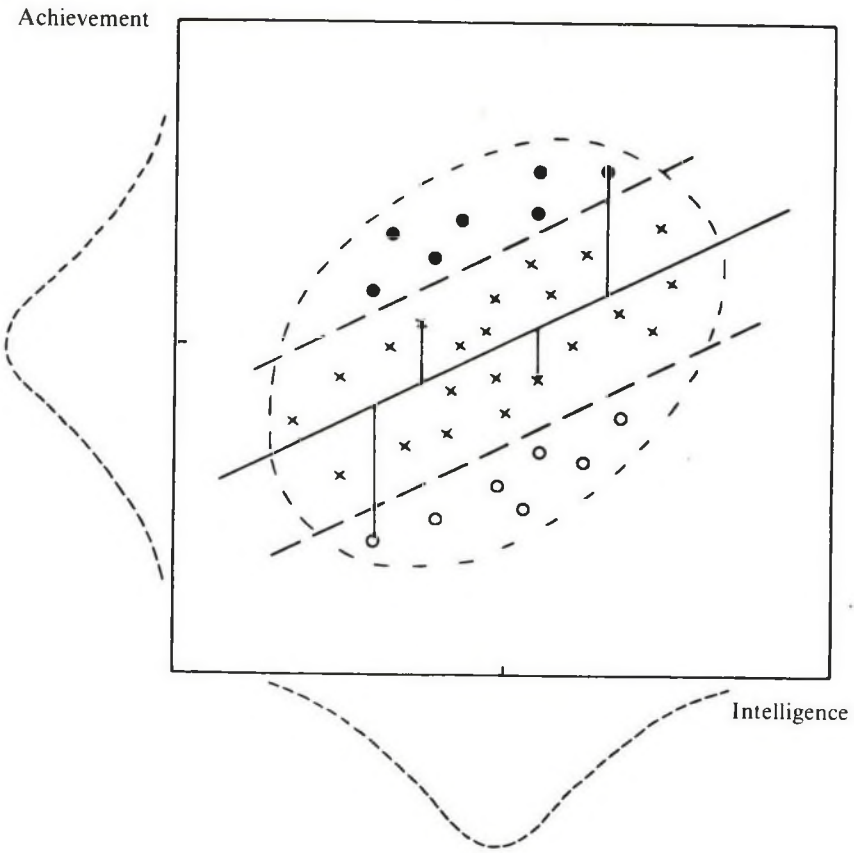
Intelligence and achievement are correlated but are not synonymous. Because they are not the same thing, it should not be surprising that the correlation between them is considerably less than perfect, even when the correlation is corrected for measurement error. Even if measurements of intelligence and of scholastic achievement were found to be perfectly correlated, however, it would not necessarily mean that intelligence, as a theoretical construct, is the same thing as achievement. Correlation does not prove synonymity. Diameters and circumferences of circles are perfectly correlated, but no one would claim that they are the same thing. Many actual tests of intelligence and scholastic achievement, however, are not completely distinct; they often have many elements in common. Tests of abilities, aptitudes, and achievement can be thought of as lying on a *continuum* (in that order) that consists of the amount of specific learning and information involved in the test items. Tests of "scholastic aptitude" contain more items involving the kinds of knowledge and skills specifically taught in school than do most tests of general intelligence. Aptitude test items sample rather broadly from the domain of past achievements that are quite closely related to the future achievements that the aptitude test is specifically intended to predict. Past progress in a specific field is the best predictor of future progress in the same field. Hence the shift from the use of IQ tests to the use of scholastic aptitude tests at higher and more specialized levels of education. Increasing the proportion of test items that tap specific past academic achievements increases the test's predictive validity, the more as students advance in school and increase their scholastic skills and knowledge, as these in turn are good predictors of students' future acquisitions in the academic sphere.

It is an important fact, however, that IQ *pre*dicts about as well as it *post*dicts scholastic achievement. That is, the IQ can predict individual differences in a particular area of cognitive or scholastic achievement before the individuals have any achievement at all in that area, and even when there are no items in the IQ test that show any resemblance to the predicted achievement in specific skills or informational content. In fact, in the elementary school grades, at least, present IQ predicts future scholastic achievement slightly but significantly better than present scholastic achievement predicts future IQ (Crano, Kenny, & Campbell, 1972). But the cause-and-effect relationship between performance on intelligence tests, and even more so on scholastic aptitude tests, on the one hand, and scholastic achievement, on the other, is best thought of as working to some extent in both directions. Scholastic achievement involves more different identifiable causal factors and correlates than IQ, which is simply the single most important factor. Achievement may be thought of as the product of intelligence × motivation, emotional stability, persistence, work habits, interests and values, and certain personality traits.

Educational Level and Predictive Validity of IQ. The most frequently mentioned value of the typical correlation between IQ and scholastic achievement is .50. This is a

good estimate of the overall average of all estimates of the predictive validity of IQ for scholastic achievement. To gain some sense of the degree of relationship represented by a correlation coefficient of .50, it is instructive to examine a correlation scatter diagram, as shown in Figure 8.1. The "bivariate normal" scatter plot is here divided into four sections by the regression of achievement scores on IQ (middle line) and the deviations of one standard deviation above and below the regression line, as indicated by the upper and lower broken lines. Traditionally, persons whose achievement score is more than one standard deviation above the regression line (i.e., their predicted achievement on the basis of IQ) have been referred to as "overachievers" (dots in Figure 8.1), and persons whose achievement is more than one standard deviation below their predicted scores have been called "underachievers." Actually these designations are quite arbitrary and really mean little more than the fact that IQ and achievement are far from perfectly correlated even if one corrects for measurement error. Because intelligence is not the only determinant of achievement, it is inevitable that there should be less than a perfect correlation, and hence the existence of "underachievers" and "overachievers." R. L. Thorndike (1963a) has expanded on this point.

Figure 8.1. A bivariate normal scatter diagram showing a correlation of .50, typical of correlations between scholastic achievement and IQ. "Underachievers" (circles) and "overachievers" (dots) are those persons whose achievement scores deviate more than one standard deviation from the regression line.



Every bit as important as the overall correlation of .50, from a psychological standpoint, is the great variation above and below this average value from one study to another. Much of this variation in the size of the correlation between IQ and scholastic achievement is quite systematic and is worth noting.

The first fact that stands out when reviewing all the evidence is that the higher that one moves up the educational ladder—from elementary school to high school, to college, and finally to graduate school—the lower in general is the correlation between IQ (or scores on scholastic aptitude tests) and indices of achievement. The typical range of most of the validities of a single IQ score for predicting academic achievement are the following for the various levels of schooling:

Elementary school	.60–.70
High school	.50–.60
College	.40–.50
Graduate school	.30–.40

The lower correlations at higher educational levels have a number of causes, which are discussed later in this chapter. But it should be emphasized that none of the causes implies in the least that intelligence becomes any less important at more advanced levels of education.

IQ Not a Threshold Variable for Scholastic Achievement. Note that the regression line of achievement on intelligence in Figure 8.1 is linear throughout the entire range of IQ scale. This is typical of the findings of the many studies that have investigated the form of the regression of achievement on IQ. The findings are unequivocal. There is no point on the IQ scale below which or above which IQ is not positively related to achievement. This means that IQ does not act as a threshold variable with respect to scholastic achievement, as has been suggested by some of the critics of IQ tests, for example, McClelland (1958, p. 13), who wrote: "Let us admit that morons cannot do good school work. But what evidence is there that intelligence is not a threshold type of variable; that once a person has a certain minimal level of intelligence, his performance beyond that point is uncorrelated with ability?" There is plenty of evidence that this is not the case. The evidence is overwhelming that scholastic achievement increases linearly as a function of IQ throughout the entire range of the IQ scale so long as scholastic achievement itself is measured on a continuous scale unrestricted by the artifacts of ceiling or floor effects due to the achievement tests not including simple enough or advanced enough items. Even items such as "can button shirt," "can tie own shoe laces," "can eat with a fork," and "can say own name" are achievements that are positively correlated with IQ at the lower end of the intelligence scale. At the other end of the scale, for IQs of 140 and above, there are still achievement differences related to IQ, as can be seen by contrasting the typical school-age intellectual achievements of Terman's (1925) gifted group with IQs above 140 (the top 1 percent) with Hollingworth's (1942) even more highly gifted group with IQs above 180. Some of the differences in the intellectual achievements even among children in the IQ range from 140 to 200 are quite astounding. Over fifty years ago, Hollingworth and Cobb (1928) strikingly demonstrated marked differences in a host of scholastic achievements between a group of superior children clustering around 146 IQ and a very superior group clustering around IQ 165. The achievement differences between these groups are about as great as between groups of children of IQ 100 and IQ 120. It is also

noteworthy that the superior (IQ 146) and very superior (IQ 165) groups do not differ in the least in ratings of the quality of their home backgrounds.

Grades versus Objective Test Scores. Part of the variation in validity coefficients from one study to another is due to *heterogeneity of the criteria,* as it is called by psychometricians. This means that different studies have used different criteria for measuring scholastic achievement. The most conspicuous source of differences in validity coefficients involves using teachers' *grades* versus *scores* on objective achievement tests as the criterion of achievement to be predicted by IQ or scholastic aptitude scores. Grades assigned by teachers typically have correlations with IQ some .10 to .20 lower than the correlation between IQ and achievement test scores.

Not all this difference is due merely to the lower reliability of grades than of test scores. There are also systematic biases in teacher-assigned grades. For example, teachers give higher marks to girls than to boys who are their equals on IQ and achievement as measured by objective tests. In elementary school, the more outgoing, socially extraverted children receive higher marks than more introverted children with the same IQ and achievement scores. Teachers' grades tend to confound achievement with deportment. Many teachers use good grades to reward effort as well as achievement. Grades are also influenced by the general level of aptitude of the particular class; a grade of A in a class of low average ability may be equivalent to the grade of C in a high-ability class, in terms of actual achievement. All these conditions work to lower the correlation between IQ and school grades as compared with achievement test scores.

Sex. In elementary school girls score slightly higher on achievement tests than do boys of the same IQ, although the sex difference is not nearly as pronounced as in the case of school grades. Girls excel particularly in subject matter involving language. With increasing grade level boys outperform girls in arithmetic and subjects involving numerical reasoning. The sex difference in overall scholastic achievement as measured by tests decreases at the higher grade levels and is practically negligible in high school, although girls still receive considerably higher grades from their teachers than do boys. Because of these sex differences, the correlation between IQ and achievement (whether assessed by grades or test scores) is generally about .10 higher when computed separately for boys and girls than when computed for the combined sexes.

Differences in School Subjects. IQ does not correlate equally with all school subjects. There are systematic differences in the average correlations between IQ and various subjects. Performance in the more highly academic and abstract subjects, such as English, mathematics, and science, is more highly predictable from IQ than is performance in subjects that depend more on special abilities, such as music and art, or acquisition of specific skills requiring narrower perceptual-motor abilities and improvement of skills through prolonged practice, such as typing and shorthand and the manual arts. Even within a specialized field such as music, IQ is probably differentially predictive for certain aspects of the field. Learning to play a musical instrument is probably less predictable from IQ than is learning harmony and counterpoint. Achievement in foreign languages shows low or intermediate correlations with IQ as compared with other academic subjects.

These differences in correlations are not the result of various subjects' having more or less content in common with IQ tests, but arise from the fact that some subjects are more *g* loaded than others; that is, they involve more of the ability for abstraction and the "eduction of relations and correlates," to use Spearman's characterization of *g*. A

content analysis of achievement tests in, say, algebra and shorthand shows no more resemblance of one or the other test to such highly *g*-loaded components of general intelligence tests as vocabulary, block designs, figure analogies, and embedded designs. Yet these tests predict performance in algebra much more highly than in shorthand. The same is true of achievement in English composition as compared with spelling or arithmetic concepts and reasoning problems as compared with arithmetic computation or so-called mechanical arithmetic.

IQ predicts achievement better in subjects that are hierarchically ordered in complexity and in the sequence of cognitive skills and knowledge that are prerequisite to more advanced achievement, as in mathematics, the physical sciences, and engineering, than in less hierarchical subjects such as history and the social sciences. The biological sciences are intermediate in this respect.

In general, the correlation between IQ and achievement is *lower* for subject matter in which there is a high correlation between achievement or the amount learned and the amount of time spent in study. But, when study time is held constant, the IQ–achievement correlation is increased. The IQ–achievement correlation is highest for subject matter that involves difficulty due to the increasing complexity of the material. It has also been found that in any subject area the correlation between IQ and achievement can be increased by eliminating the easier items in the achievement test. An easy item that, say, 80 percent of the students pass and 20 percent fail has the same statistical capacity to correlate with IQ as a more difficult item that is passed by only 20 percent and failed by 80 percent. Yet the easier item in fact correlates less with IQ. High-IQ pupils tend to miss those achievement test items that involve material that they did not study or that was presented by the teacher on the day that the student was absent. In addition to these kinds of failures, low-IQ pupils tend more to miss those achievement items that are abstract or conceptually complex, for example, *thought* problems in arithmetic, the proper *application* of the relevant formulas in physics and chemistry, the logical *inferences* and interpretation of meaning involved in the reading comprehension of literature, history, and the social sciences. In short, IQ predicts scholastic achievement because, and to the extent that, achievement is dependent on those kinds of cognitive processes that characterize *g*. *The correlation does not come about because IQ tests only measure knowledge that has been taught in school.*

Other Factors That Lower the Predictive Validity of IQ. While there can be no doubt that IQ measures aptitude for academic education, it is also important to note that IQ accounts for only about half, or less, of the variance in measured achievement at any given point in the course of schooling. Why is the correlation between IQ and achievement not higher?

For one thing, the achievement tests themselves differ in *content validity* from one class to another and from one school to another. That is, the item composition of any given standardized achievement test is a less than perfect sample of the subject matter that has actually been presented in a particular class in the months preceding the administration of the achievement test. Scores on an achievement test that is tailored specifically to what was actually taught in a given class, assuming that the test has all the other desirable psychometric features of the best standardized tests, would yield a higher correlation with IQ than do the usual standardized tests. A corollary of this is that, when teachers are familiar with the content and nature of the standardized achievement tests to be used in their class at the end of the school year and "teach to the test," it most likely has the effect

of increasing the correlation between IQ and achievement test scores. *IQ correlates best with achievement scores when there has been uniformity of exposure of pupils to all of the subject matter sampled by the achievement test.*

Another factor that lowers the IQ–achievement correlation is the greater restriction of range on many achievement tests as compared with IQ tests, on which there is virtually no restriction of range in the general school population. Achievement tests are usually designed to assess achievements for the subject matter content of a particular grade level in school. Such tests fall far short of sampling the full range of scholastic knowledge that exists in the total pupil population at any one grade level. The achievement tests thus have "ceiling" and "floor" effects that restrict the possible size of the obtained correlation with IQ. Laymen scarcely appreciate the actual spread of scholastic achievement that exists within any one grade level in the typical large city school system. The total range of scholastic knowledge and skills increases with every grade, and by high school the students falling below the 10th percentile may average six or seven grade levels below the students who score above the 90th percentile in achievement. The top high school students possess greater scholastic knowledge than the majority of college graduates, whereas the poorest high school students are academically on a par with the average second or third grader. But the high school junior who knows calculus and Boolean algebra cannot show this on the standard achievement tests typically used in high school. Nor are the academic limitations of the poorest students revealed by such tests, because rarely are high school students given tests that are appropriate for second and third-graders—the level of test that would have to be used fully to reveal the actual range of academic achievement among high school students.

Motivational factors undoubtedly play a part in achievement, and the imperfect correlation between motivation and IQ can therefore lower the predictive validity of IQ. But the contribution of motivational differences is probably overrated, for the following reason: there is a positive correlation between academic aptitude (i.e., IQ) and academic motivation, and the correlation increases throughout the course of schooling. The positive correlations among ability, achievement, and motivation work to enhance the correlation between IQ and achievement. Nothing reinforces the behavioral manifestations of motivation as much as success itself. Abler students are rewarded by greater success, which in turn reinforces the kinds of behavior—attention, interest, persistence, and good study habits—that lead to further academic success. The repeated failures of less able students generally have just the opposite effect. A pupil's self-perceived failure, even when it is not explicitly pointed out by the teacher, is a kind of punishing or at least unrewarding kind of experience from which the student is anxious to escape. Thus the academically less successful students tend to withdraw from academic pursuits and seek out other areas for enhancing their self-esteem. The greatest efforts are made by the most successful. The teacher and pupil alike are usually not very highly motivated to move the pupil's performance from the 1st percentile to the 10th percentile. At the other extreme, one often sees phenomenal efforts on the part of some students, such as scholarship winners, who are already at the 99th percentile and wish to climb to the 99.9th percentile, or from the 99.9th to the 99.99th percentile. The most striking examples of this kind of motivation and effort to exceed excellence are seen in great musical virtuosos, world chess champions, and Olympic athletes.

There are temporal fluctuations in people's performance both in achievement and in

IQ tests due to irregularities and the ups and downs that everyone experiences, some much more than others. Persons differ from day to day and even over somewhat longer periods in their lives in how effectively they use the abilities they possess and how they deploy their investments of attention, interest, and energy. At the extreme, for example, we know that psychiatric patients or persons suffering from severe emotional disorders often show great fluctuations in performance on IQ tests amounting to twenty or thirty points or more, along with a temporary impairment of the ability to accomplish anything that demands sustained mental effort. At a lesser extreme, everyone is familiar with "off days" or being in the doldrums for even a week or more.

Because of such fluctuations in the conditions affecting performance, more accurate assessments of ability and achievement and a consequently higher predictive validity of IQ can be secured with repeated measurements spread over a period of time, as was first demonstrated by Noel Keys (1928). By averaging various IQ and achievement test results over several years, the correlations between and among the tests approach the magnitude of the general factor common to all of the IQ and achievement tests—correlations between .80 and .90. Hence, with temporal fluctuations averaged out over the years, the correlation between IQ and scholastic achievement approximates the *g* saturation of IQ tests. Over the entire course of schooling from kindergarten to college, IQ and objective measures of academic achievement probably have as much as 90 percent or more of their variance in common.

This is illustrated by test data from 274 pupils in an integrated suburban school district on whom were obtained IQs, achievement scores, and teachers' marks in every grade from third through high school (Vane, 1966). The average correlation between IQ and scholastic achievement in any single grade was .67, with a range of correlations from .56 to .71 across grades. From the data given in Vane's Table 1, we can extract a general factor (i.e., first principal component) from the matrix of intercorrelations among achievement measures at every grade level; this large general achievement factor accounts for no less than 82 percent of the total variance in achievement at all grade levels. The IQ, measured at any single grade level, correlates on the average .79 with the general achievement factor, which is close to the correlation between the IQ obtained at any single administration and the *g* factor loading of any single group IQ test.

Prediction of Achievement by Multifactor Tests of Ability. Numerous studies have shown that by far the largest share of the validity of mental ability tests for predicting scholastic achievement is attributable to the *g* factor. Mental ability factors other than *g*, as measured by multifactor tests such as Thurstone's Primary Mental Abilities, the Differential Aptitude Tests, and the General Aptitude Test Battery of the U.S. Employment Service, add surprisingly little to the prediction of overall scholastic achievement or even of achievement in specific academic subjects. Rarely is the multiple correlation between a number of differential aptitude tests and achievement measures more than .10 greater than the simple correlation between IQ and achievement.[1] Verbal ability contributes more to the prediction of scholastic achievement independently of *g* than does any other ability factor.

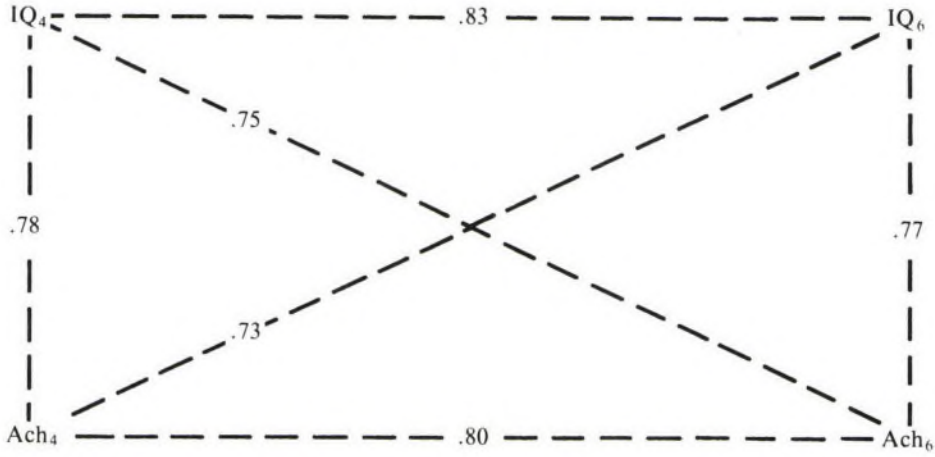
Although multifactor ability tests predict only slightly better than a single IQ score, the prediction of achievement is considerably enhanced by the use of composite achievement scores, mainly because a number of different achievement tests provide a broader sampling of students' achievements than any single achievement test. Composite scores

on multiachievement batteries have been found to correlate close to .80 with single IQ scores at the elementary school level where there is no restriction of range and all children are exposed to the same curriculum.

Achievement in Elementary School. Results quite typical of those found in most studies of the predictive validity of IQ are seen in a large-scale study by Crano, Kenny, and Campbell (1972). It has the added advantage of showing both the concurrent and predictive validities of IQ. Achievement was measured by a composite score on the Iowa Tests of Basic Skills, which measure achievement and skills in reading, language (spelling, punctuation, usage, etc.), arithmetic, reading of maps, graphs, and tables, and knowledge and use of reference materials. IQ was measured by the Lorge-Thorndike Intelligence Test. The tests were taken by a representative sample of 5,495 children in the Milwaukee Public Schools in Grade 4 and parallel forms of the tests were obtained again in Grade 6. Figure 8.2 shows all of the correlations among the four sets of measurements. Notice that the predictive validity of IQ over an interval of two years (IQ_4–Ach_6) is nearly as high as the concurrent validity (IQ_4–Ach_4 and IQ_6–Ach_6). As is typically found, past achievement predicts future achievement slightly better than IQ.

One might wonder to what extent the common factor of reading ability per se involved in group tests of IQ and achievement plays a part in such intercorrelations. It is not as great as one might imagine. Although the verbal items of group IQ tests usually involve reading, the reading level is deliberately made simpler than the conceptual demands of the items, so that individual differences in the IQ scores are more the result of general cognitive ability than of reading ability per se. The reading requirements of an IQ test for sixth-graders, for example, will typically involve a level of reading ability within the capability of the majority of fourth-graders. The Lorge–Thorndike IQ test has both Verbal and Nonverbal parts; the Verbal requires reading, the Nonverbal does not. In a large study (Jensen, 1974b) of children in Grades 4 to 6, a correlation of .70 was found between the Verbal and Nonverbal IQs. The correlation between Verbal IQ and the reading comprehension subtest of the Stanford Achievement Test was .52. The correlation between Nonverbal IQ (which involves no reading) and reading comprehension scores

Figure 8.2. Correlations among IQ and achievement test scores on 5,495 children in Grades 4 and 6. (From Crano, Kenny & Campbell, 1972)



was .47. The correlation between Verbal IQ and reading comprehension after Nonverbal IQ is partialled out is only .29. The Verbal IQ test obviously measures considerably more than just reading proficiency.

IQ and Learning to Read. Pupils' major task in the primary grades (i.e., Grades 1 to 3) is learning to read. There are two main aspects of reading skill: *decoding* and *comprehension*. Decoding is the *translation* of the printed symbols into spoken language, and *comprehension*, of course, is *understanding* what is read. The learning of decoding (also called *oral reading*) is somewhat less predictable from IQ than is reading comprehension, which, once decoding skill has been achieved, quite closely parallels mental age. When elementary school children (all of the same age) are matched on decoding skill, their rank on a test of reading comprehension is practically the same as on IQ. In fact, reading comprehension per se is almost indistinguishable from oral comprehension once decoding is acquired. Most students with poor *reading* comprehension perform no better on tests of purely *oral* comprehension. But the reverse does not hold: there are some children (and adults) whose oral comprehension is average or superior, yet who have inordinate difficulty in the acquisition of decoding. When such disability is severe and unamenable to the ordinary methods of reading instruction, it is referred to as *developmental dyslexia*. Dyslexia seems to be a specific cognitive disability that does not involve *g* to any appreciable extent. Some dyslexics obtain high scores on both the verbal and nonverbal parts of individual IQ tests that require no reading, and they can be successful in college courses, especially in mathematics, physical sciences, and engineering, provided that someone reads their textbooks to them. There is no deficiency in comprehension per se. The vast majority of poor readers, however, are poor readers not because they lack decoding skill, but because they are deficient in comprehension, which, as measured by standard tests of reading comprehension is largely a matter of *g* (E. L. Thorndike, 1917; R. L. Thorndike, 1973–74.)

Here are some typical results. The Wechsler Preschool and Primary Scale of Intelligence (WPPSI), which does not involve reading, was given to children in kindergarten prior to any instruction in reading and was correlated with tests of reading achievement in first grade after one year's instruction in reading (Krebs, 1969). Achievement was measured by the Gilmore Oral Reading Test (a test of decoding) and the reading subtests of the Stanford Achievement Test (SAT), which involves word meaning and paragraph comprehension as well as decoding. The one-year predictive validities of the WPPSI IQ scales are as follows:

WPPSI	*Gilmore Oral Reading*	*SAT Reading Comprehension*
Verbal Scale IQ	.57	.61
Performance Scale IQ	.58	.63
Full Scale IQ	.62	.68

When the sample was divided into lower- and upper-socioeconomic-status groups, it was found that the predictive validity of IQ was higher in the lower-SES group than in the higher-SES group (e.g., SAT reading scores correlated .66 versus .40 with Full Scale IQ).

Group tests of *reading readiness* look a good deal like group IQ tests in item content. They are intended to predict reading achievement in the primary grades and can

be taken by children prior to having received any instruction in reading. Lohnes and Gray (1972) factor analyzed seven reading readiness tests and an IQ test given to 3,956 pupils in 299 classrooms in the first weeks of the first grade, before they could read. The IQ test correlated .84 with the general factor (i.e., first principal component) common to the reading readiness tests, a higher correlation than that of any of the readiness tests themselves, which showed correlations with the general factor ranging from .44 to .81, with a median of .60. Two years later, when the same pupils were in the second grade, they were given ten reading and language achievement tests and one arithmetic computation test. These were factor analyzed, yielding correlations with the general factor of the achievement battery ranging from .64 (arithmetic computation) to .87 (reading vocabulary), with a median correlation of .80. The general factor of the reading readiness battery (including IQ) correlated .81 with the general factor of the achievement battery. Lohnes and Gray conclude:

> There is no question that reading skills of pupils were observed by the criterion measurement instruments [i.e., the achievement tests given in second grade]. What these analyses reveal is that the most important single source of criterion variance, or to put it differently, the best single explanatory principle for observed variance in reading skill, was variance in general intelligence. (p. 475)

IQ, Learning Ability, and Retention. The relation between intelligence and learning ability has long been a puzzle to psychologists. It is still not well understood, but a number of consistent findings permit a few tentative generalizations.

Part of the problem has been that "learning ability" has been much less precisely defined, delimited, and measured than intelligence. The psychometric features of most measures of "learning ability" are not directly comparable with tests of intelligence, and it is doubtful that much further progress in understanding the relation between learning and intelligence will be possible until psychologists treat the measurement of individual differences in learning with at least the same degree of psychometric sophistication that has been applied to intelligence and other abilities.

One still occasionally sees intelligence defined as learning ability, but for many years now, since the pioneer studies of Woodrow (1938, 1939, 1940, 1946), most psychologists have dropped the term "learning ability" from their definitions of intelligence. To many school teachers and laymen this deletion seems to fly in the face of common sense. Is not the "bright," or high-IQ, pupil a "fast learner" and the "dull," or low-IQ, pupil a "slow learner?" Simple observation would surely seem to confirm this notion.

The ability to learn is obviously a mental ability, but it is not necessarily the same mental ability as intelligence. Scientifically the question is no longer one of whether learning ability and intelligence are or are not the same thing, but is one of determining the conditions that govern the magnitude of the correlation between measures of learning and measures of intelligence.

The Woodrow studies showed two main findings. (1) Measures of performance on a large variety of rather simple learning tasks showed only meager intercorrelations among the learning tasks, and between learning tasks and IQ. Factor analysis did not reveal a general factor of learning ability. (2) Rate of improvement with practice, or gains in proficiency as measured by the difference between initial and final performance levels, showed little or no correlation among various learning tasks or with IQ. Even short-term

pretest–posttest gains, reflecting improvement with practice, in certain school subjects showed little or no correlation with IQ. Speed of learning of simple skills and associative rote learning, and rate of improvement with practice, seem to be something rather different from the *g* of intelligence tests. Performance on simple learning tasks and the effects of practice as reflected in gain scores (or final performance scores statistically controlled for initial level of performance) are not highly *g* loaded.

Many other studies since have essentially confirmed Woodrow's findings. (Good reviews are presented by Zeaman and House, 1967, and by Estes, 1970.) The rate of acquisition of conditioned responses, the learning of motor skills (e.g., pursuit rotor learning), simple discrimination learning, and simple associative or rote learning of verbal material (e.g., paired associates and serial learning) are not much correlated with IQ. And there is apparently no large general factor of ability, as is found with various intelligence tests, that is common to all these relatively simple forms of learning.

The same can be said of the *retention* of simple learning. When the degree of initial learning is held constant, persons of differing IQ do not differ in the retention of what was learned over a given interval of time after the last learning trial or practice session.

But these findings and conclusions, based largely on simple forms of learning traditionally used in the psychological laboratory, are only half the story. Some learning and memory tasks do in fact show substantial correlations with IQ. This is not an all-or-none distinction between types of learning, but a continuum, which in general can be viewed as going from the simple to the complex. What this means needs to be spelled out more specifically. Individual differences in learning proficiency show increasingly higher correlations with IQ directly in relation to the following characteristics of the learning task.

1. Learning is more highly correlated with IQ when it is *intentional* and the task calls forth conscious mental effort and is paced in such a way as to permit the subject to "think." It is possible to learn passively without "thinking," by mere repetition of simple material; such learning is only slightly correlated with IQ. In fact, *negative* correlations between learning speed and IQ have been found in some simple tasks that could only be learned by simple repetition or rote learning but were disguised to appear more complex so as to evoke "thinking" (Osler & Trautman, 1961). Persons with higher IQs engaged in more complex mental processes (reasoning, hypothesis testing, etc.), which in this specially contrived task only interfered with rote learning. Persons of lower IQ were not hindered by this interference of more complex mental processes and readily learned the material by simple rote association.

2. Learning is more highly correlated with IQ when the material to be learned is *hierarchical,* in the sense that the learning of later elements depends on mastery of earlier elements. A task of many elements, in which the order of learning the elements has no effect on learning rate or level of final performance, is less correlated with IQ than is a task in which there is some more or less optimal order in which the elements are learned and the acquisition of earlier elements in the sequence facilitates the acquisition of later elements.

3. Learning is more highly correlated with IQ when the material to be learned is *meaningful,* in the sense that it is in some way related to other knowledge or experience already possessed by the learner. Rote learning of the serial order of a list of meaningless

three-letter nonsense syllables or colored forms, for example, shows little correlation with IQ. In contrast, learning the essential content of a meaningful prose passage is more highly correlated with IQ.

4. Learning is more highly correlated with IQ when the nature of the learning task permits *transfer* from somewhat different but related past learning. Outside the intentionally artificial learning tasks of the experimental psychology laboratory, little that we are called on to learn beyond infancy is *entirely* new and unrelated to anything we had previously learned. Making more and better use of elements of past learning in learning something "new"—in short, the transfer of learning—is positively correlated with IQ.

5. Learning is more highly correlated with IQ when it is *insightful,* that is, when the learning task involves "catching on" or "getting the idea." Learning to name the capital cities of the fifty states, for example, does not permit this aspect of learning to come into play and would therefore be less correlated with IQ than, say, learning to prove the Pythagorean theorem.

6. Learning is more highly correlated with IQ when the material to be learned is of *moderate difficulty* and *complexity*. If a learning task is too complex, everyone, regardless of his IQ, flounders and falls back on simpler processes such as trial and error and rote association. Complexity, in contrast to sheer difficulty due to the amount of material to be learned, refers to the number of elements that must be integrated simultaneously for the learning to progress.

7. Learning is more highly correlated with IQ when the *amount of time* for learning is fixed for all students. This condition becomes increasingly important to the extent that the other conditions listed are enactive.

8. Learning is more highly correlated with IQ when the learning material is more *age related.* Some things can be learned almost as easily by a 9-year-old child as by an 18-year-old. Such learning shows relatively little correlation with IQ. Other forms of learning, on the other hand, are facilitated by maturation and show a substantial correlation with age. The concept of *learning readiness* is based on this fact. IQ and tests of "readiness," which predict rate of progress in certain kinds of learning, particularly reading and mathematics, are highly correlated with IQ.

9. Learning is more highly correlated with IQ at an *early stage* of learning something "new" than is performance or gains later in the course of practice. That is, IQ is related more to rate of acquisition of new skills or knowledge rather than to rate of improvement or degree of proficiency at later stages of learning, assuming that new material and concepts have not been introduced at the intermediate stages. Practice makes a task less cognitively demanding and decreases its correlation with IQ. With practice the learner's performance becomes more or less automatic and hence less demanding of conscious effort and attention. For example, learning to read music is an intellectually demanding task for the beginner. But for an experienced musician it is an almost automatic process that makes little conscious demand on the higher mental processes. Individual differences in proficiency at this stage are scarely related to IQ. Much the same thing is true of other skills such as typing, stenography, and Morse code sending and receiving.

It can be seen that all the conditions listed that influence the correlation between learning and IQ are highly characteristic of much of school learning. Hence the impression of teachers that IQ is an index of learning aptitude is quite justifiable. Under the listed

conditions of learning, the low-IQ child is indeed a "slow-learner" as compared with children of high IQ.

Very similar conditions pertain to the relation between memory or retention and IQ. When persons are equated in degree of original learning of simple material, their retention measured at a later time is only slightly if at all correlated with IQ. The retention of more complex learning, however, involves meaningfulness and the way in which the learner has transformed or encoded the material. This is related to the degree of the learner's understanding, the extent to which the learned material is linked into the learner's preexisting associative and conceptual network, and the learner's capacity for conceptual reconstruction of the whole material from a few recollected principles. The more that these aspects of memory can play a part in the material to be learned and later recalled, the more that retention measures are correlated with IQ.

These generalizations concerning the relationship between learning and IQ may have important implications for the conduct of instruction. For example, it has been suggested that schooling might be made more worthwhile for many youngsters in the lower half of the IQ distribution by designing instruction in such a way as to put less of a premium on IQ in scholastic learning (e.g., Bereiter, 1976; Cronbach, 1975). Samuels and Dahl (1973) have stated this hope as follows: "If we wish to reduce the correlation between IQ and achievement, the job facing the educator entails simplifying the task, ensuring that prerequisite skills are mastered, developing motivational procedures to keep the student on the task, and allocating a sufficient amount of time to the student so that he can master the task."

IQ and College Grades. The predictive validity of IQ for success in college has to be dealt with separately, as it involves problems peculiar to this level. For one thing, omnibus achievement tests are seldom given to college students. Students' academic achievements are assessed only in those courses that they have taken in college, and this criterion measure is usually just the final grade received in the course. College grades constitute a five-point scale (A, B, C, D, and F), usually with a highly skewed distribution (i.e., many more A's and B's than D's and F's). Hence this is a quite crude scale and far from statistically optimal as a criterion measurement against which to determine the validity of any predictor variables. Most studies of predictive validity use grades averaged over all of the courses the student has taken in college, the grade-point average or GPA, obtained by assigning numerical values to the letter grades.

According to Lavin's (1965, p. 51) review of the literature, the validity of college entrance exams, such as the Scholastic Aptitude Test (SAT), for predicting college grade-point averages ranges from about .30 to .70, with an average correlation of about .50. Other reviews of this voluminous literature cited by Lavin give highly similar values as typical. When multiple predictors based on tests of specific aptitudes relevant to different courses of study are used to predict GPA, the multiple correlation coefficients reported may be as high as .60 to .70. But it now appears that this increase in validity is not so much the result of using multiple predictors (i.e., differential aptitude tests) as it is a result of using more homogeneous criteria, namely, predicting GPA *within* groups majoring in the same subjects. If grades are not strictly comparable across different fields of study, this fact can only weaken the correlation between any predictor variable and grades when students from all fields are pooled together.

College aptitude tests such as the SAT are not, strictly speaking, general intelligence

tests, although they would no doubt show a quite high correlation with IQ. The aptitude tests are a kind of hybrid, combining items typical of those found in IQ tests and items typical of those found in high school achievement tests. Including the assessment of high school achievement significantly enhances the predictive validity of the college aptitude test. This should not be surprising, because academic knowledge and skills gained in high school are a prerequisite for many college courses. High school grades or the student's rank in his or her graduating class generally predict college GPA at least as well as scores on college aptitude tests.

There are few studies of the correlation between standard intelligence tests (in contrast to scholastic aptitude tests) and college grades. The Full Scale IQ of the Wechsler Adult Intelligence Scale correlated .44 with college freshman GPA in a college where the mean IQ of freshmen is 115. (The correlation of WAIS IQ with rank in high school graduating class, with a mean IQ of 107, was .62; see Matarazzo, 1972, p. 284.)

The most extensive evidence that I have been able to find of the correlation between general intelligence and college grades is based on the general intelligence test of the General Aptitude Test Battery (GATB) developed by the U.S. Employment Service (Manpower Administration, U.S. Department of Labor, 1970). The general intelligence test of the GATB is a good measure of *g* and correlates .89 with the WAIS Full Scale IQ. Table 8.6 shows the frequency distribution of correlations between GATB intelligence test scores and college grades (usually GPA) in 48 different samples (totaling 5,561 students) from diverse colleges and for various majors within the colleges. The median correlation is .40.

A number of conditions contribute to the considerable variation in correlations between IQ and college grades.

Sex differences in the predictability of college grades are a quite consistent finding, with higher predictive validities for females than for males. This is also true at the high school level. The causes of this sex difference are obscure. (See Chapter 13, pp. 628–630, for a discussion of this.)

Table 8.6. Correlations between scores on the general intelligence test of the GATB and college grades in 48 college samples. (From *Manual of the GATB*, Sec. III, pp. 205–219, Manpower Administration, U.S. Department of Labor, Washington, D.C., 1970)

Correlation	Frequency	Percentage
.60–.64	1	2%
.55–.59	1	2
.50–.54	8	17
.45–.49	4	8
.40–.44	10	21
.35–.39	5	10
.30–.34	8	17
.25–.29	3	6
.20–.24	5	10
.15–.19	1	2
.10–.14	2	4

Selection of students in terms of high school grades and scholastic aptitude scores lowers predictive validity by restriction of range on both predictor and criterion variables.

Field of study is also related to the predictive validity of IQ. Grades in mathematics and the sciences generally show the highest predictive validities, followed by the social sciences and humanities, and finally by the arts. These differences are attributable mainly to three factors: (1) the hierarchical nature of the subject matter in math and science and the fact that IQ at all ages is quite highly related to capacity for mastering material that is hierarchically ordered in terms of increasing complexity in which the simpler elements are prerequisite for the more complex; (2) the greater objectivity and reliability of the criteria for assessing achievement in math and science than in the humanities and the arts; and (3) the important role of special talents in the arts.

Self-dependence of college students is greater than of pupils in elementary and high school and works to lower the validity of IQ and scholastic aptitude tests in predicting college grades. There is much less parent and teacher supervision of the student's study habits in college. Class attendance is seldom mandatory, and the time spent in classes is only a small fraction of the total study time needed to obtain passing grades, time that the student must allocate wisely on his own. Individual differences in self-discipline, study habits, and the like lessen the correlation between ability and achievement.

Variable grading standards is the most important cause of attenuation in predicting college grade-point average from intelligence and aptitude tests. The GPA is a composite scale of *nonequivalent* components, that is, grades in different courses. Hence the GPAs of students who have taken different courses are not equivalent in their regressions on ability. This has been conclusively shown in research by Roy Goldman and his associates (Goldman, Schmidt, Hewitt, & Fisher, 1974; Goldman & Slaughter, 1976; Goldman & Hewitt, 1976). Their studies justify the conclusion that GPA is a very poor criterion against which to judge the validity of tests for predicting *actual* achievement in college. GPA per se is a poor index of actual achievement.

Goldman et al. have shown that grading standards differ from one field to another and from one course to another within fields. What is more important is that the stringency of grading standards is positively related to the average level of student ability within a field or within a course in a given field. High-ability students can perform well in any field, but low-ability students can pass only in certain fields and courses with lax grading standards. This is evident in the preponderant direction of changes in college major, which is from "harder" to "easier" fields in terms of the level of ability required to obtain passing grades. Low-ability students tend to gravitate toward fields and courses with lax grading standards. This condition obviously plays havoc with the predictive validity of aptitude tests for predicting composite GPA. Taking into account differences in reliability of grades and restriction of range in different courses, the validities of the SAT are considerably higher for grades within separate fields than for overall GPA and are still higher for grades in separate classes.

In other words, the aptitude tests predict actual achievement in college better than they have been given credit for when judged in terms of their correlation with GPA. Goldman and Slaughter (1976) conclude:

> As long as there are radical differences in grading standards, and students are able to choose most of their classes, then *no predictor* will have more than moderate

> validity for predicting GPA. There are several remedies for this situation, but none of them seem politically palatable. One solution would be to create a conversion system for equating grades in one class with grades in another. Although this presents some technical difficulties, they are not insurmountable. Nevertheless, we imagine that there will be a great deal of resistance to such a suggestion, although it has been implicitly adopted for many purposes: medical schools, for example, tend to weight grades from different classes with greater or lesser values depending on the perceived difficulties of the classes. Another possible solution would be to use the GPA in a particular field as the success criterion. This too would present problems. In sum, we believe that the validity problem in GPA prediction is a result of the shortcomings of the GPA criterion rather than the tests that are used as predictors. Recognition of this phenomenon would eliminate much pointless argument about the merits of standardized tests for college student selection. (p. 14)

Goldman and Hewitt (1976) also discovered the interesting fact that the SAT–Verbal score is more predictive of grades (even in the sciences) than the SAT–Math score. On the other hand, SAT–Math is more predictive of the student's major field and career choice. College courses can be arranged along a quantitative–nonquantitative continuum, and the mean scores of students in the courses so ordered show a steeper gradient in SAT–Math scores than in SAT–Verbal, although the two abilities are correlated. The continuum is as follows: physical sciences (including mathematics), biological sciences, social sciences, humanities, fine arts. Which of these areas a student is most apt to major in is determined largely by his mathematical aptitude; but the grades that he receives within his chosen field are determined more by verbal ability. Nonscience students are below science students in mathematical ability; but, contrary to popular belief, nonscience students are not higher than science students in verbal ability.

The ratio of males to females in different college majors seems to be largely "explainable" in terms of the sex difference in mathematical aptitude. The correlations between sex and major field (which range from .15 to .28 in four colleges) are greatly reduced (to values from .07 to .17) when males and females are statistically equated on the verbal and the math SAT scores.

Prediction of Grades in Graduate School. The same conditions that work against predictive validity in undergraduate college operate even more strongly in graduate and professional schools. In most graduate and professional schools, such as law and medicine, the ratio of applicants to selectees varies from about 10 to 1 to 100 to 1. This high degree of selection implies a severe restriction of range on the predictor variables. Last year at the University of California in Berkeley, for example, none of the finally selected students admitted to the graduate program in the Mathematics Department scored below the 98th percentile on the Graduate Record Examination, a high-level scholastic aptitude test standardized on college seniors and graduates. Moreover, at the graduate level, course grades usually constitute only a three-point scale (grades of A, B, C), with the vast majority of grades consisting of A's and B's. Such conditions militate severely against high correlations between the predictor and criterion variables. Hence grades in graduate school typically correlate with aptitude scores in the range from .30 to .40. The median correlation between first-year average grades in twenty-eight law schools and scores on the Law School Aptitude Test (essentially a high-level verbal intelligence test)

was .30, ranging from .01 to .40 (Pitcher & Schrader, 1969). (Correlation of undergraduate grades with law school grades was only .27.)

A much more potent factor than stringent selection (and the consequent restriction of range) operates to lower the predictive validity of aptitude tests in graduate school. The method of selecting students, by most graduate schools, statistically guarantees a low correlation between the predictor and the criterion. Students are selected on the basis of undergraduate grades and aptitude scores such that among the accepted students there is a substantial *negative* correlation between undergraduate grades and aptitude scores. Because both undergraduate grades and aptitude scores are each positively correlated with grades in graduate school, when the two variables are negatively correlated with each other in the selected sample, it makes it mathematically impossible for either predictor variable (i.e., undergraduate grades and aptitude scores) to be highly correlated with the criterion (i.e., grades in graduate school). This observation has been clearly explicated by Dawes (1975). The most prestigious and highly selective graduate schools admit almost exclusively students who have *high* GPA and *high* aptitude scores. There is poor predictive validity in this select group mainly due to restriction of range. Students who do not make it on both points gain admittance to other graduate schools if they have *high* GPA and *low* aptitude (i.e., high or low relative to the median of all other applicants) or *low* GPA and *high* aptitude. (Few graduate schools would normally select low-GPA–low-aptitude students.) This results in a negative correlation between GPA and aptitude, making it impossible for either predictor variable to correlate highly with the criterion, as explained. There is simply no way in which the single correlations between predictors and criterion can be high. Because of these peculiar conditions in the selection of graduate students, the correlations between aptitude scores and grades are not a fair assessment of the actual validity of aptitude tests for predicting the achievement of graduate students.

IQ and Amount of Formal Education. Number of years of schooling is a common criterion of educational achievement, mainly because it is so easy to ascertain. For the past decade or more, however, it has been a relatively poor index of actual educational achievement, so great are the average differences between various colleges as well as the range of individual differences in achievement within any given college population. One large study using a very broad general scholastic achievement test (the General Culture Test) found that about 10 percent of high school seniors exceeded the median achievement level of college seniors (Tyler, 1965, pp. 104–105). If diplomas were awarded to the upper fifth of the entire college student body on the basis of tested knowledge rather than on hours and credits, the composition of the "graduating" class would consist of 28 percent seniors, 21 percent juniors, 19 percent sophomores, and 15 percent freshmen; and 10 percent of high school seniors could be awarded the diploma before ever having entered college.

The differences between colleges are enormous. When the Selective Service College Qualification Test was given to more than 74,000 men in colleges throughout the United States in 1952, in some colleges as many as 65 percent failed the test as compared with 2 percent in some other colleges (Tyler, 1965, p. 106). There are even colleges that graduate some students who fail the Armed Forces Qualification Test, the failing score on which is equivalent to an IQ of less than 80. There are also large differences in the percentage passing the Selective Service College Qualification Test in different college majors, from 69 percent passing in physical science and mathematics to 30 percent

passing in education, with an average of 54 percent over nine majors (Tyler, 1965, p. 106).

Considering the wide margin of discrepancy between objectively tested attainments and years of education, diplomas, and credentials, it seems an obvious conclusion that most employers and our social institutions in general have put far too much stock in sheer amount of schooling and formal credentials and not enough in objectively assessed actual achievement.

Despite these conditions that attenuate amount of education as an index of real educational achievement, there is still a quite substantial correlation between IQ and amount of schooling. The IQ is clearly not a *result* of the amount of schooling, as childhood IQ predicts the final level of education attained by adulthood. For example, in a group of 437 adults there was found a correlation of .58 between their IQs measured in the sixth grade (at age 12) and the amount of education they had attained by the age of 45 (Bajema, 1968).

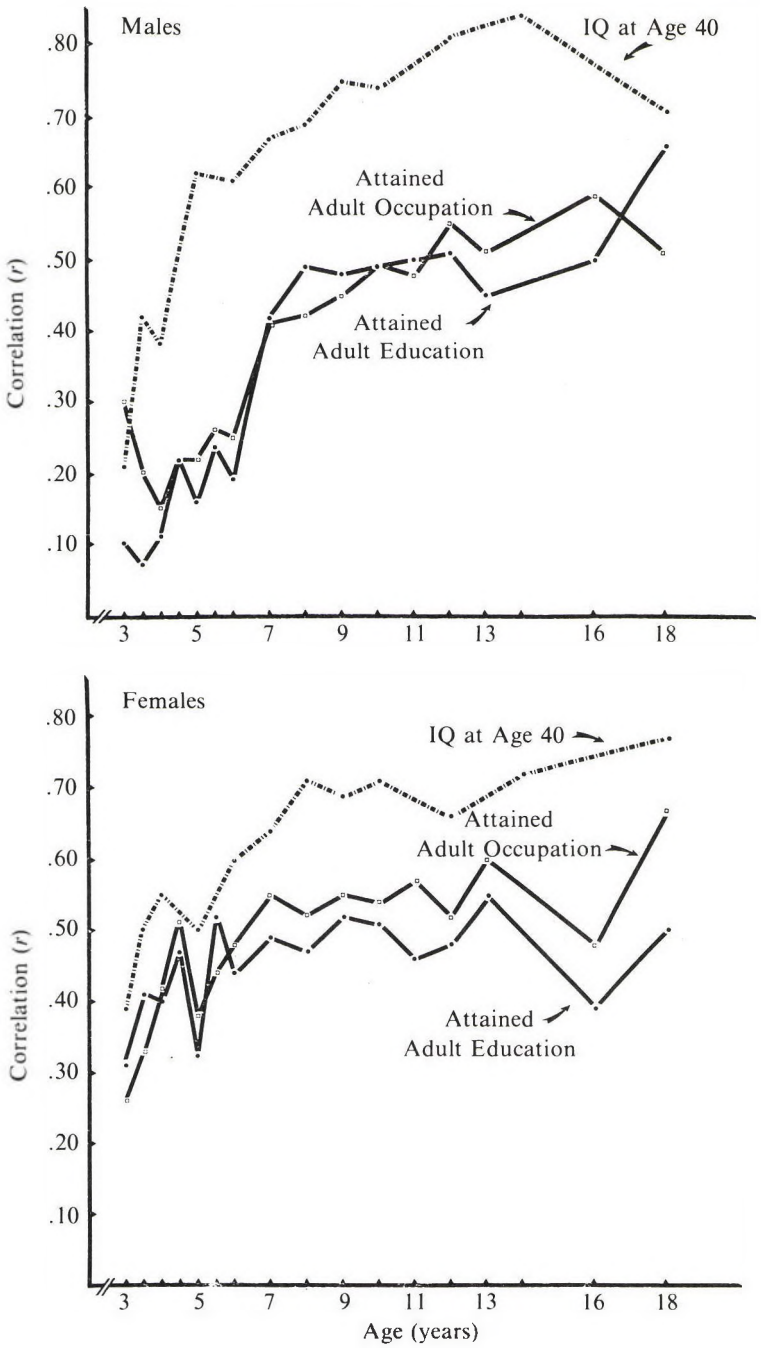
IQ is positively correlated with school persistence even if we do not consider schooling beyond high school graduation. Dillon (1949) followed the high school careers of 2,600 pupils who took an IQ test in the seventh grade. The percentage of pupils within each of five IQ intervals who persisted to high school graduation was as follows:

IQs less than 85	4%
IQs 85–94	54%
IQs 95–104	63%
IQs 105–114	76%
IQs 115 and above	86%

Longitudinal data from the Fels Research Institute (McCall, 1977) shows the correlations between IQ measured at various intervals between 3 and 18 years of age and adult educational and occupational attainment in samples of 94 males and 96 females all of at least 26 years of age. The group was considerably above average, with a mean IQ of 117, standard deviation of 15.9. Among the women 3 percent did not graduate from high school, 31 percent graduated but did not go beyond high school, and 34 percent graduated from college. The comparable figures for men were 1 percent, 22 percent, and 56 percent, respectively. Figure 8.3 shows the correlations of IQ × adult educational and occupational attainments for males and females. Notice that by 7 years of age, the IQ predicts adult educational and occupational levels with a validity coefficient of .40 to .50. IQ at age 40 (data from a study by Honzik, 1972), as shown in Figure 8.3, is considerably more predictable from childhood IQ than are educational and occupational attainments. Also it is interesting that females' adult attainments are more predictable from IQ at an early age than is the case for males. The cause of this sex difference is open to speculation.

IQ Not a Stand-in for Socioeconomic Status. The claim has been made that IQ as a predictor of amount of education attained by adulthood is merely a "stand-in" for socioeconomic status. SES is indexed mainly by the father's occupational status and the educational level of both parents. If a child's SES determines his educational achievement or number of years spent in school, we should not expect to find a significant correlation between IQ and years of schooling among brothers reared together in the same family. Yet among brothers there is a correlation of about .30 to .35 between IQ and years of schooling as adults when IQ is measured in elementary school. (This correlation can be inferred

Figure 8.3. Correlation between IQ at ages from 3 to 16 years and IQ at age 40, adult occupational status, and final educational level attained by adulthood, shown separately for males and females. (From McCall, 1977)



from data presented by Jencks, 1972, p. 144.) Within-family differences in educational attainments for same-sex siblings cannot be attributed to differences in SES, "cultural differences," or "family background."

A study in Britain (Kemp, 1955) determined the correlations among IQ, tested scholastic achievement, and SES, with all of the intercorrelated variables consisting of the mean values obtained on these characteristics in fifty schools. The intercorrelations were as follows:

IQ and scholastic achievement = .73
IQ and SES = .52
SES and scholastic achievement = .56

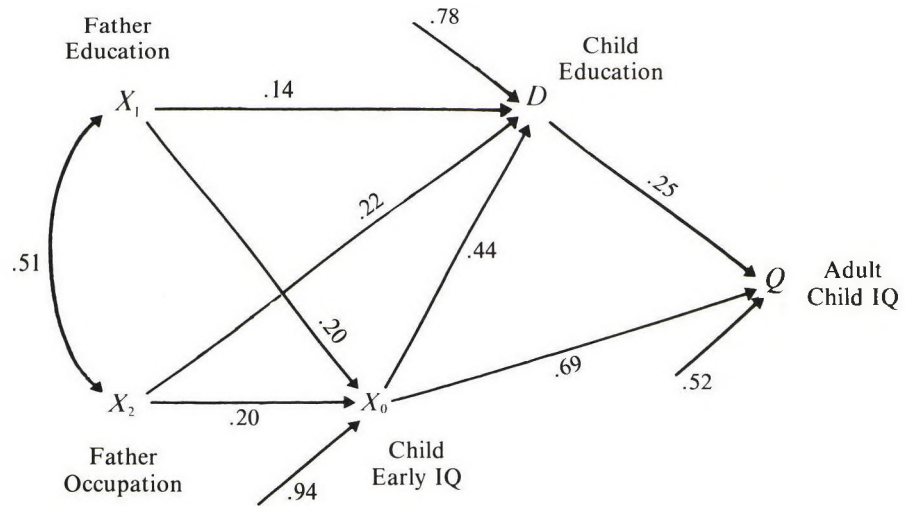
When IQ is partialled out (i.e., held constant statistically) of the correlation between SES and scholastic achievement, the partial correlation drops to .30. However, when SES is partialled out of the correlation between IQ and achievement, the partial correlation drops only to .62. This means that IQ independently of SES determines achievement much more than does SES independently of IQ.

Because father's education and occupation are the main variables in almost every composite index of SES or "family background," it is instructive to look at the degree of causal connection between these variables and a child's early IQ (at age 11), the child's level of education (i.e., highest grade completed) attained by adulthood, and the child's IQ as an adult. The intercorrelations among all these variables were subjected to a "path coefficients analysis" by the biometrician C. C. Li (1975, pp. 324–325).

Path analysis is a method for inferring causal relationships from the intercorrelations among the variables when there is prior knowledge of a temporal sequence among the variables. For example, a person's IQ can hardly be conceived of as a causal factor in determining his or her father's educational or occupational level. The reverse, however, is a reasonable hypothesis. The path diagram as worked out by Li (from data presented by Jencks, 1972, p. 339) is shown in Figure 8.4.

In path diagrams the observed correlations are conventionally indicated by curved lines (e.g., the observed correlation of .51 between father's education and father's occupation). The temporal sequence goes from left to right, and the direct paths, indicating the unique causal influence of one variable on another independently of other variables, are represented by straight lines with single-headed arrows to indicate the direction of causality. (Arrows that appear to lead from nowhere (i.e., from unlabeled variables) represent the square roots of the residual variance that is attributable to variables that are unknown or unmeasured in the given model.) We see in Figure 8.4 that the direct influences of father's education and occupation contribute only $.14^2 + .20^2 = 6$ percent of the variance in the child's final educational attainment (i.e., years of schooling) as an adult, whereas the direct effect of the child's IQ at age 11 in determining final educational level is $.44^2$, or 19 percent of the variance. In brief, childhood IQ determines about three times more of the variance in adult educational level than father's educational and occupational levels combined. Notice also that the father's education and occupation combined determine only $.20^2 + .20^2 = 8$ percent of the variance in childhood IQ. Li concludes: "The implication seems to be that it is the children with higher IQ who go to school rather than that schooling improves children's IQ. The indirect effect from early IQ to adult IQ via education is $(0.44)(0.25) = 0.11$" (p. 327).

Figure 8.4. Path diagram showing analysis of the network of some of the causal influences on IQ from early childhood (early child IQ) to adult (adult child IQ). (From Li, 1975, p. 325)



IQ and Evaluations by Parents, Teachers, and Peers

As was pointed out in Chapter 6, teachers' ratings of pupils' intelligence correlate between .60 and .80 with IQ scores. Teachers' ratings of ability generally show a sex bias slightly favoring girls, probably because deportment and scholastic achievement enter into teachers' judgments of intellectual ability. Yet teachers' and pupils' ratings of mental ability show a remarkably high agreement as evinced by the correlations between teacher ratings of pupils' intelligence and pupils' ratings of one another on intelligence—a raw correlation of .85 (or .95 when corrected for unreliability of the ratings) for girls and of .76 or .90 corrected) for boys (Thorndike, Lay, & Dean, 1909). The correlations of the ratings with school marks was .60 for teachers' ratings and .40 for pupils' ratings, which are values similar to the correlations of IQ scores with school marks.

IQ and other measurements of general intelligence also show significant correlations with interpersonal ratings and behaviors even when the raters have not been instructed to rate on intellectual ability per se. A study of 7,417 children of ages 6 to 11, sampled so as to be representative of the school population of the United States, showed significant relationships between children's intellectual ability and various behavioral indices of social acceptance by their classmates and peers (Roberts & Baird, 1972). In competitive game activities, for example, the first few children chosen for a team were much more often (54 percent versus 8 percent) selected from among the higher-IQ children (upper quartile) in the class than from among the lower-IQ children (lower quartile). The frequency with which children were chosen as leaders by their peers was also related to ability level. A number of other studies have shown essentially the same thing: the more intelligent school children are usually better accepted socially by their classmates than are the less intellectually favored children and the retarded (Baldwin, 1958; Barbe, 1954; Epperson, 1963; Gallagher, 1958).

Parental attitudes along a continuum of acceptance–rejection of their children are significantly correlated with the children's IQs, especially in the case of mothers and daughters, showing an average correlation of .42 (or .55 corrected for unreliability) between daughters' IQs and accepting attitudes by their mothers (Hurley, 1965). The correlations were significantly lower for boys. The direction of the causality of the correlation between parental attitudes and child's IQ is unknown, but the lack of sex differences between all four possible sex combinations of parent–child IQ correlations (which average close to .50) would suggest that the child's IQ causes the parental attitudes rather than the reverse.

Adults' ratings of one another also reflect IQ to some extent even when the basis of rating is not intelligence per se but a quality only indirectly correlated with intelligence. Izard (1959) found low but significant correlations between group intelligence test scores (as well as certain personality traits) and peer nominations for leadership ability among military personnel. Among student nurses, peer predictions of success in nurses' training were significantly related to measures of verbal and numerical ability (Poland, 1961).

In a study by Schmidt (unpublished manuscript), trainees for foremen in a large manufacturing concern were asked to rate each of their peers' probable degree of future success on the job as foremen. Among white ratees, the peer ratings (made by whites and blacks) correlated .385 with the unweighted composite score on a battery of eight diverse mental tests. (Such a composite score should be a good measure of g). Future success ratings of blacks by black raters correlated .421 with the composite mental test score, but the ratings of blacks by whites was not significantly correlated with test score ($r = .088$). Trainees also rated one another for "drive and assertiveness." These ratings, too, correlated significantly ($r = .409$) with composite test score in the white sample, but again ratings of blacks by whites were not significantly correlated with test scores, although blacks' ratings of whites showed a significant correlation (.367) with test score. It is not known why the ratings of blacks by white raters are so much less correlated with the ability measures than the ratings of blacks by blacks or the ratings of whites by whites *or* blacks. One may wonder if a similar pattern of correlations would have resulted if the raters were asked to estimate the general intelligence level of their co-workers.

"Assortative mating" is the term used by geneticists to refer to the degree of positive correlation between mates on any given observable or measurable characteristic. It is a fact of considerable interest that among married couples the degree of assortative mating for IQ is higher than for any other trait, physical or mental. Studies of assortative mating for IQ show correlations between spouses ranging between +.40 to +.60, with a mean of +.50 (Jencks, 1972, p. 272). (This is about the same as the correlation between brothers and sisters.) A correlation of .50 is equivalent to an average absolute difference between spouses (or siblings) of 12 IQ points. Assuming that the IQ tests have a reliability of .95, the correlation between spouses after correction for attenuation becomes +.53. (The marital correlation for physical stature is about +.30.) It is hard to imagine how such a correlation could come about if IQ were not a socially important variable and if a host of personal cues and other social determinants in mate selection were not correlated with IQ, since prospective marriage partners ordinarily do not give one another IQ tests or have access to one another's IQ scores in old school files. The correlation of about .50 between spouses' IQs implies a correlation of at least .70 between measured IQ and one's ability to estimate the intelligence of oneself and of others who are known quite well in a variety of

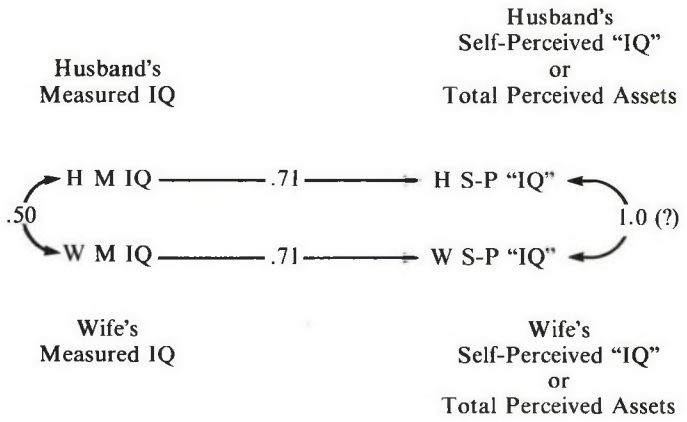
contexts. Apparently a person is perceived as less attractive as a possible marriage partner when his or her intelligence as perceived by the partner (a perception that may be based on a host of observable correlates of intelligence) departs more than a certain amount from the partner's self-perception of his or her own intelligence level. Too great a discrepancy between these perceptions of the partner's intelligence and the person's own intelligence decreases the probability of the persons' marrying one another. If we assume that the validity of the perceptions of intelligence is .71, as measured against the criterion of measured IQ, and that mates' self-perceptions of their intelligence are perfectly correlated, then we should expect their IQs to be correlated $.71 \times 1.0 \times .71 = .50$, as shown in Figure 8.5.

But mates' self-perceptions of their own intelligence are most likely not perfectly correlated, because there is a mutual "trade-off" between perceived intelligence and other assets that determine overall attractiveness. Hence it could be argued that there is an even higher correlation than .71 between perceived intelligence and measured IQ to account for the correlation of .50 between mates' measured IQs. Another possibility is that one does not marry another person unless the other person's overall assets, as perceived by oneself, are *at least* equivalent to one's own overall self-perceived assets. This requirement on the part of both mates would ensure a near perfect correlation between their overall self-perceived assets. The correlation, then, between these self-perceived total assets and measured IQ would have to be at least .71 to account for the observed correlation of .50 between spouses' measured IQs.

Occupational Level, Performance, and Income

Not the judgment of the "average" person, but the *averaged judgments* of many persons can show an extraordinary consistency across quite diverse groups of persons, and from one generation to the next, as well as remarkably high correlations with certain independent objective criteria. Such is the case with people's average subjective judgments of occupational "level" and their high correlation with the average tested IQs of persons in various occupations. This striking finding has been demonstrated to about the same high degree in numerous studies and has been contradicted by none. (For more

Figure 8.5. Path diagram of marital correlation.



extensive reviews of this evidence the reader is referred to Matarazzo, 1972, Chs. 7 and 12; and Tyler, 1965, Ch. 13.)

People's average ranking of occupations is much the same regardless of the basis on which they were told to rank them. The well-known Barr scale of occupations was constructed by asking 30 "psychological judges" to rate 120 specific occupations, each definitely and concretely described, on a scale going from 0 to 100 according to the level of general intelligence required for ordinary success in the occupation. These judgments were made in 1920. Forty-four years later, in 1964, the National Opinion Research Center (NORC), in a large public opinion poll, asked many people to rate a large number of specific occupations in terms of their subjective opinion of the *prestige* of each occupation relative to all of the others. The correlation between the 1920 Barr ratings based on the average subjectively estimated *intelligence requirements* of the various occupations and the 1964 NORC ratings based on the average subjective opined *prestige* of the occupations is .91. The 1960 *U.S. Census of Population: Classified Index of Occupations and Industries* assigns each of several hundred occupations a composite index score based on the average income and educational level prevailing in the occupation. This index correlates .81 with the Barr subjective intelligence ratings and .90 with the NORC prestige ratings.

Rankings of the prestige of 25 occupations made by 450 high school and college students in 1946 showed the remarkable correlation of .97 with the rankings of the same occupations made by students in 1925 (Tyler, 1965, p. 342). Then, in 1949, the average ranking of these occupations by 500 teachers college students correlated .98 with the 1946 rankings by a different group of high school and college students. Very similar prestige rankings are also found in Britain and show a high degree of consistency across such groups as adolescents and adults, men and women, old and young, and upper and lower social classes. Obviously people are in considerable agreement in their subjective perceptions of numerous occupations, perceptions based on some kind of amalagam of the prestige image and supposed intellectual requirements of occupations, and these are highly related to such objective indices as the typical educational level and average income of the occupation. The subjective desirability of various occupations is also a part of the picture, as indicated by the relative frequencies of various occupational choices made by high school students. These frequencies show scant correspondence to the actual frequencies in various occupations; high-status occupations are greatly overselected and low-status occupations are seldom selected.

How well do such ratings of occupations correlate with the actual IQs of the persons in the rated occupations? The answer depends on whether we correlate the occupational prestige ratings with the *average* IQs in the various occupations or with the IQs of individual persons. The correlations between *average* prestige ratings and *average* IQs in occupations are very high—.90 to .95—when the averages are based on a large number of raters and a wide range of rated occupations. This means that the average of many people's subjective perceptions conforms closely to an objective criterion, namely, tested IQ. Occupations with the highest status ratings are the learned professions—physician, scientist, lawyer, accountant, engineer, and other occupations that involve high educational requirements and highly developed skills, usually of an intellectual nature. The lowest-rated occupations are unskilled manual labor that almost any able-bodied person

could do with very little or no prior training or experience and that involves minimal responsibility for decisions or supervision.

The correlation between rated occupational status and *individual* IQs ranges from about .50 to .70 in various studies. The results of such studies are much the same in Britain, the Netherlands, and the Soviet Union as in the United States, where the results are about the same for whites and blacks. The size of the correlation, which varies among different samples, seems to depend mostly on the *age* of the persons whose IQs are correlated with occupational status. IQ and occupational status are correlated .50 to .60 for young men ages 18 to 26 and about .70 for men over 40. A few years can make a big difference in these correlations. The younger men, of course, have not all yet attained their top career potential, and some of the highest-prestige occupations are not even represented in younger age groups. Judges, professors, business executives, college presidents, and the like are missing occupational categories in the studies based on young men, such as those drafted into the armed forces (e.g., the classic study of Harrell & Harrell, 1945).

Evidence contradicts the notion that IQ differences between occupations are the result rather than a cause of the occupational difference. Professional occupations do not score higher than unskilled laborers on IQ tests because the professionals have had more education or have learned more of the test's content in the pursuit of their occupations. A classic study (Ball, 1938) showed that childhood IQs of 219 men correlated substantially with adult occupational status as measured on the Barr scale some 14 to 19 years later—a correlation of .47 for a younger sample of men and of .71 for a sample of older men just five years further into their careers. Thorndike and Hagen (1959) analyzed data on the tested abilities of 10,000 World War II airforce cadets, *all of them high school graduates with IQs above 105,* who were tested at age 21, and their postwar occupations (classified into 124 occupational categories) at 33 years of age. Recall that this was an above-average group in IQ (and education) to begin with, constituting the upper 35 percent of the general population. Yet their scores on a test of general intelligence at age 21 show a marked relationship to their occupational classifications 12 years later. For example, men in the following high-status occupations (listed alphabetically) averaged .53 standard deviations *above* the mean of the whole group of 10,000 in "general intelligence" score: accountants, architects, college professors, engineers, lawyers, physicians, scientists, treasurers and comptrollers, and writers. The following occupations of lower-status occupations (not the lowest, as these were not represented in this above-average sample) averaged .54 standard deviations *below* the overall mean: bus and truck drivers, guards, miners, production assemblers, tractor and crane operators, railroad trainmen, and welders. (Day laborers and unskilled manual occupations are not represented in this group.) In an informal study, mean prestige ratings made by a group of college students of 42 of the 124 occupations in the Thorndike and Hagen list correlated .74 with the occupations' average intelligence scores (R. J. Herrnstein, personal communication, 1971).

To gain an accurate impression of the full range of mean intelligence differences between occupational levels, we must look at a representative sample of the working population that has not been previously selected on intelligence or education. The U.S. Department of Labor has obtained such information (see Manpower Administration, 1970). A representative sample of 39,600 of the employed U.S. labor force in the age

range from 18 to 54 years was given the U.S. Employment Services General Aptitude Test Battery. The sample contains 444 of the specific occupations listed in the U.S. Department of Labor's *Dictionary of Occupational Titles* (1965). (Certain occupations were not included in the sample: all farmers and farm workers and foremen, proprietors, managers and officials, and service workers.) The overall mean GATB General Intelligence score is 100, with a standard deviation of 20. The means of the 444 specific occupations range from 55 (tomato peeler) to 143 (mathematician). Thus the total range of occupational means embraces 4.45 standard deviations on the scale of the distribution of intelligence test scores in the general working population of men and women in the United States, with the exception of the excluded occupations noted. Figure 8.6 shows the frequency distribution of the means of the GATB intelligence test scores of the 444 occupations. Superimposed on it is the normal distribution. Even though the *Dictionary of Occupational Titles* was not made up with reference either to intelligence or to the normal distribution nor was the sampling of these 444 occupations or the 39,600 men and women in them, it is interesting to see that the distribution of occupational means is roughly symmetrical but departs significantly from the normal curve. The distribution of occupational means is quite *leptokurtic;* that is, there is a piling up of scores in the middle of the distribution with too few scores at the extremes for a truly normal curve. The lack of perfect symmetry of the distribution, with too small frequencies in the range from 60 to 80, may well be due in part to the exclusion from the sample of all farm workers and the unemployed. We know from other studies that intelligence test scores from the lower half of the normal distribution are overrepresented among farm workers and the unemployed. The smaller frequencies in the

Figure 8.6. Frequency distribution (shaded histogram) of *mean* intelligence test scores of 444 occupational categories. A normal curve (smooth line) shows the theoretical distribution of *individual* scores in the population.

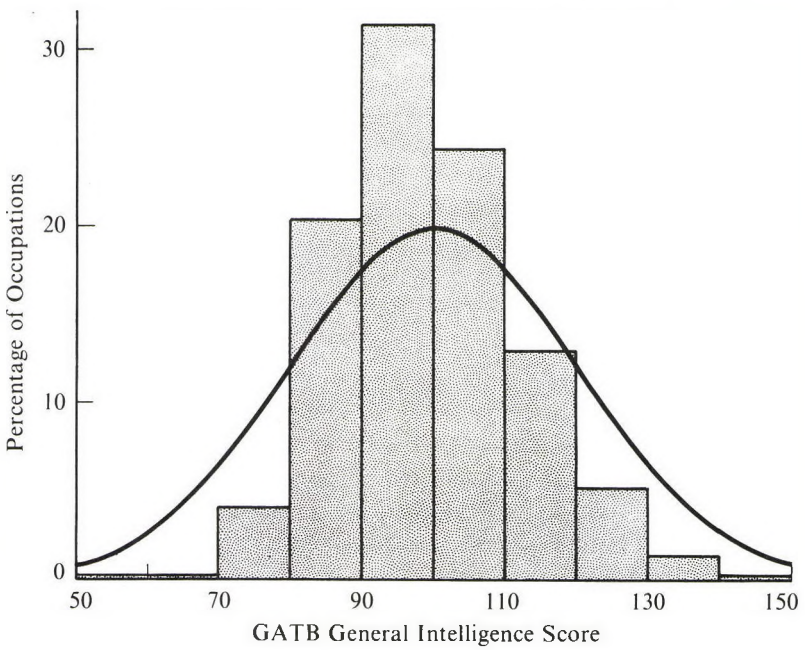
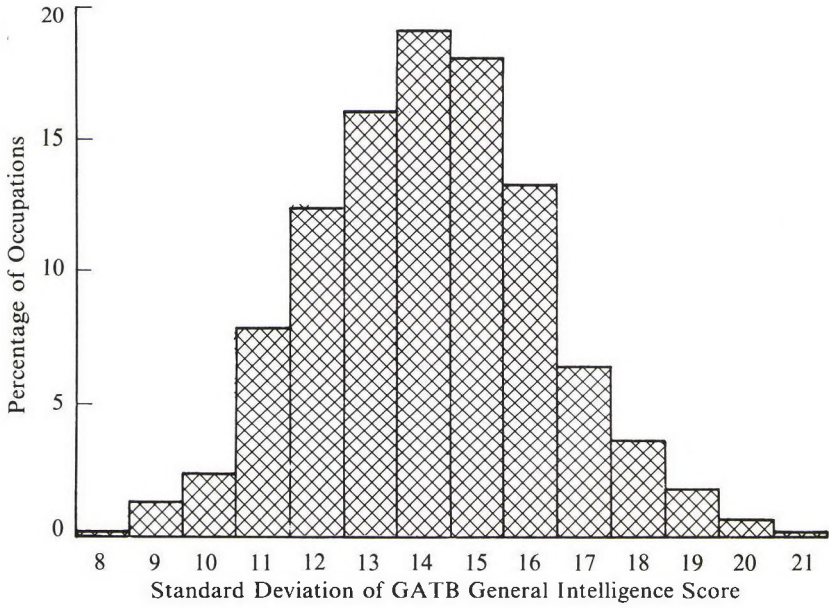


Figure 8.7. Frequency distribution of the standard deviations of intelligence test scores of 444 occupational categories.



lower tail of the distribution are also due to the fact that the distribution of scores within the lowest occupational categories is quite positively skewed, which pulls up the mean of the occupational group. (Median scores would be expected to show a more normal distribution than mean scores.) Hence it appears that the frequency distribution of the average intellectual requirements of existing occupations corresponds somewhat to the distribution of general intelligence in the working adult population with the marked exception that the percentage of jobs of average ability (i.e., within plus or minus one standard deviation of the mean) is greater than the percentage of persons of average ability. This is because so many of these middle-level jobs contain persons scoring over quite a wide range, so that the *average* score within each job category is pulled toward the middle of the distribution.

Thus at least as important a fact as the differences *between* the means of occupations is the wide distribution of individual scores *within* occupations. Figure 8.7 shows the frequency distribution of the standard deviations of GATB intelligence scores in 444 occupations. They range from a standard deviation of 8 to a standard deviation of 21, with a mean standard deviation of 14.6. These standard deviations may be compared with the standard deviation of 20 for the distribution of individual scores in the total population. The full *range* of scores in different occupations is of course much greater. Analysis of variance shows that of the total population variance in test scores, 47 percent of it is variation *between* the means of occupations and 53 percent is individual variation *within* occupations. (From these figures, it follows statistically that the correlation between individual intelligence scores and occupational classifications is $\sqrt{.47} = .69$). The distribution of individual scores *within* occupations is roughly normal for most occupations but tends to become skewed to the right in the occupations lowest on the intelligence scale; in

these groups there is a greater piling up of low scores with a long narrow tail of the distribution extending up to quite high scores.

It is a consistent finding in all the studies of occupations and IQ that the standard deviation of scores within occupations steadily *decreases* as one moves from the lowest to the highest occupational levels on the intelligence scale. In other words, a diminishing percentage of the population is intellectually capable of satisfactory performance in occupations the higher the occupations stand on the scale of occupational status. Almost anyone can succeed as a tomato peeler, for example, and so persons of almost every intelligence level except the severely retarded may be found in such a job. But relatively few can succeed as a mathematician; no persons in the lower half of the intelligence distribution are to be found in this occupation in which nearly all who succeed are in the upper quarter of the population distribution of IQ. Thus the lower score of the total range of scores in each occupation is much more closely related to occupational status than is the upper score of the range.

For example, in a study using the Army General Classification Test (with an overall mean of 100 and a standard deviation of 20), the range of scores for engineers is 100 to 151 (with a mean of 127) whereas the range for farmers is 24 to 147 (with a mean of 93) (Harrell & Harrell, 1945). From the "lowest" to the "highest" occupations, the top of the score range varies only 12 points (from 145 to 157), whereas the bottom score of the range varies 86 points (from 16 to 102). This threshold quality of general intelligence with respect to occupational status was first stated clearly by Harrell and Harrell (1945, p. 239):

> Evidently a certain minimum of intelligence is required for any one of many occupations and a man must have that much intelligence in order to function in that occupation, but a man may have high intelligence and be found in a lowly occupation because he lacks other qualifications than intelligence.

A certain threshold level of intelligence is a necessary but not sufficient condition for success in most occupations. Therefore a low IQ is much more predictive of occupational level than is a high IQ. A person with a high IQ may be anything from an unskilled laborer to a Nobel Prize-winning scientist. But low-IQ persons are not found at all in the sciences or in any of the learned professions. The range of WAIS IQs of 80 medical students in one year's class of the University of Oregon Medical School goes from 111 to 149 (with a median of 125.5). But the lowest IQ of 111 still exceeds the IQs of 77 percent of the general population (Matarazzo, 1972, p. 177). WAIS IQs of 148 members of the Cambridge University faculty ranged from 110 to 141 (with a mean of 126.5) (Matarazzo, 1972, p. 180). A group of 243 policemen and firemen had a range of WAIS IQs from 96 to 130, with a median of 113. In this group there was one exceptional outlier with an IQ of 86 (Matarazzo, 1972, p. 175). Almost no professional, technical, or highly skilled job has a median IQ below 100. Typical jobs with median IQs falling slightly below this point are crane operator, cook, weaver, truck driver, laborer, barber, lumberjack, farmhand, and miner (Harrell & Harrell, 1945).

Terman's famous follow-up study of some 1,500 school children with IQs of 140 and above (with an average IQ of 152) showed that this group by middle age attained a far higher level of occupational status than would have been expected for a random sample of persons of comparable childhood backgrounds or even of college graduates (Terman & Oden, 1959). Among the men in Terman's study, the ten most frequent occupations were

lawyers, engineers, college professors, major business managers, financial executives, scientists, physicians, educational administrators, top business executives, and accountants—in that order. Over 85 percent of the men in the Terman group were employed in these high-level occupations. Only about 3 percent were farmers or semiskilled laborers, and virtually none were unskilled laborers.

The Terman sample also had earnings well above the income level of the general population and higher than persons of the same education and occupation. In fact, in this very superior group, amount of formal education seemed to make relatively little difference in income. For example, even those in the Terman sample who had not gone beyond high school had earnings comparable with those who had graduated from college with a bachelor's degree; and of the six men with the highest incomes in the entire sample, only one was a college graduate.

In the general population, however, there is a closer link between education and income than was found for the Terman gifted group. It has been argued by a number of sociologists and economists that the correlation between IQ and income (as well as between IQ and occupational status) is largely *indirect;* it is *mediated* via the correlation of IQ with amount of education and of education with occupation and income (e.g., Bajema, 1968; Bowles & Gintis, 1973; Eckland, 1965; Jencks, 1972). This conclusion is based on the observation that the *partial correlation* between amount of education (i.e., highest grade completed) and occupational status (statistically holding IQ constant) is much higher than the partial correlation between IQ and occupational status (statistically holding education constant). Table 8.7 shows the simple correlations and partial correlations found in two typical studies. What these partial correlations mean is that occupational status is related to IQ for all persons having the same educational level.

But the interpretation of such partial correlations is very tricky. They are easily misleading. The high partial correlation of education and occupation, for example, would seem to imply that almost anyone, given the necessary amount of education, could attain the corresponding occupational status more or less regardless of his or her IQ, as the partial correlation of IQ and occupation is quite low. But this would be a false inference, because not everyone can attain the educational thresholds required by the higher occupa-

Table 8.7. Simple and partial correlations between IQ, amount of education, and occupational status in two samples. (Above diagonal, data from Bajema, 1968; below diagonal, data from Waller, 1971)

	Simple Correlation		Partial Correlation	
	Education	*Occupation*	*Education*	*Occupation*
IQ[1]	.52 / .58	.50 / .46	.27 / .42	.21 / .15
Education		.72 / .63		.63 / .50

[1]IQ in the Bajema study is Terman Group Test given in the sixth grade. IQ in the Waller study is Otis and Kuhlman group tests given in school at a mean age of 13.38 years.

tions. Holding IQ constant statistically, as a partial correlation, only means that, among those whose IQs are above the threshold required for any given occupation, educational attainment then becomes the chief determinant of occupational level. The low partial correlation between IQ and occupation does not contradict the importance of the threshold property of IQ in relation to occupational status. If the true relationship between IQ and occupation were as low as the partial correlations would seem to suggest, we should find every level of IQ in every type of occupation. But of course this is far from true, even in occupations to which entry involves little or no formal education. Moreover, not all high-IQ persons choose to enter the professions or other high-status occupations, but those who do so work to attain the required educational levels; and hence educational level is more highly correlated with occupational level than is IQ per se.

The causal relationships between IQ, education, and occupational status are too complex to be explained satisfactorily in terms of partial correlations, as the forms of the correlation scatter diagram between these variables do not all involve to the same degree the property of being "necessary and sufficient" types of correlations.[2] (A statistician would note that the bivariate distributions are not *homoscedastic,** which makes partial correlations hazardous.) Because of the practically inextricable causal connections among IQ, education, and occupations, probably the least contentious kind of correlation that one can report is the multiple correlation R between the *combined* effects of IQ and education, on the one hand, and occupational level, on the other. The R based on the data in Table 8.7 is .64 for the Bajema study and .73 for the Waller study.[3]

Some economic and social theorists (e.g., Bowles & Gintis, 1973) would like to have us believe that occupational level is not causally related to IQ but is almost wholly a result of privilege associated with the individual's social-class background, especially the educational and occupational level of the parents, as well as sheer "luck" (which only means as yet unexplained sources of variance). This claim is only partly true. It is partly true because of the correlation between education and occupation and the fact that education has some relationship to social class independently of IQ. Youths from low-socioeconomic backgrounds are less likely to finish high school or graduate from college than are youths of higher SES but of the same IQs. This is more true at mediocre levels of IQ; SES makes a considerably greater educational difference for the mediocre than for those at the upper end of the IQ scale.

The fact that privilege associated with family background is not the whole story in occupational level is shown by the great variability in occupation and income among members of the same family. Inequality in occupational status between brothers is about 82 percent of status inequality in the general population; the correlation between brothers' occupational statuses is only about .30 (Jencks, 1972, pp. 198, 343). Moreover, a substantial part of the status inequality within families is related to IQ differences within families. This is clearly seen in a study by Waller (1971), which found that the discrepancy between father's and son's adult occupational status correlated +.368 with the difference in their IQs. The IQs were obtained from high school records for both fathers and sons. Sons with IQs higher than their fathers' IQs tended to attain higher occupational levels than their fathers, and sons with lower IQs than their fathers' generally fell below their fathers' occupational levels. Waller concludes that intelligence produces variation in persons' occupational attainments that is unrelated to the status of their family origin.

Income, like occupational level, is causally related to IQ in a complex fashion. The

simple correlation between IQ and earnings is about .30 for white males and only about .10 for black males, and the simple correlations between education and earnings are about the same (Brown & Reynolds, 1975). Jencks (1972, p. 240) estimates the correlation between intelligence (AFQT scores) and income, after correction for unreliability of measurement, to be .349. A correlation this size, though seemingly small, still has considerable consequences in terms of dollars and cents. As Jencks (1972, p. 220) notes, men who scored above the 80th percentile on the Armed Forces Qualification Test (AFQT) after the Korean War had personal incomes 34 percent above the national average, whereas men who scored below the 20th percentile had incomes about 34 percent below the average; this amounts to the first group's earning about twice as much as the second group. Jencks's analyses of the best available income and IQ data led to the conclusion that "about half the observed relationship between test scores and income persists after we control family background and [educational] credentials" (p. 221).

Citing other evidence on the relation of earnings to IQ and education, Leona Tyler (1974, p. 47) draws the following conclusion:

> These figures show that the more college education a high school graduate obtained, the higher his income turned out to be; but they also show that with any amount of education beyond high school, persons who as children scored in the top ten percent [i.e., IQ 119 and above] on an intelligence test had a distinct advantage over the rest. What the figures suggest is that measured intelligence is related not solely to school success and survival but also to the kinds of real life success reflected in income differences. If this is true, some advantage of high scores may remain even if schools and colleges cease to carry out the screening that has been a major source of the relationship between educational aptitude and occupational success.

Tested Ability and Performance within Occupations. The IQ and other ability test scores are considerably better at predicting persons' occupational statuses than at predicting how well they will perform in the particular occupational niche they enter. Some one-fourth to one-half of the total IQ variance of the employed population is already absorbed in the allocation of persons to different occupations, so that there is less IQ variation left over that can enter into the correlation between IQ and criteria of success *within* occupations.

Restriction of range, however, is not the major factor responsible for the often low correlations between test scores and job performance. For one thing, in the vast majority of jobs, once the necessary skills have been acquired, successful performance does not depend primarily on the ability we have identified as *g*. Other traits of personality, developed specialized skills, experience, and ability to get along with people become paramount in job success as it is usually judged. It has been said that in the majority of jobs, as far as employers are concerned, the most important ability is not intellectual ability but *dependability*.

Another factor that lowers predictive validity is the lack of standardization of the criterion. The criterion of successful job performance is usually a judgment of the worker's immediate supervisor. Supervisor ratings have notoriously poor reliability compared with objective measures of performance. The assessment of job success is often based on more objective indices such as actual production records, sales records, and tested job-related knowledge and skills. In many test validation studies based on various criteria of

job performance, inconsistent criteria are used for different employees on the same job. For example, Mr. X, a stockman, receives a low rating because, although seemingly conscientious, he is judged to be slow and inaccurate in his work; Mr. Y in the same job, although he is fast and accurate, receives the same rating as Mr. X because Mr. Y is often late to work, overextends coffee breaks and lunch hour, and is often seen socializing rather than working on the job. Both men are given low job ratings by their supervisor for such different reasons that it would seem miraculous if their IQs were at all correlated with their ratings.

Yet despite these kinds of limiting conditions, there are still sufficiently substantial predictive validities of ability tests for many jobs to be of considerable value in personnel selection.

A review of the entire literature on the validity of tests for predicting job performance found that intelligence tests correlate on the average in the range of .20 to .25 with ratings of actual proficiency on the job (Ghiselli, 1955). An equally important finding of Ghiselli's review is that the average validity of intelligence scores for predicting proficiency differs systematically for various types of jobs. For example, here are the ranges for the majority of validity coefficients for the following groups of occupations:

.00 to .19	Sales, service occupations, machinery workers, packers and wrappers, repairmen
.20 to .34	Supervisors, clerks, assemblers
.35 to .47	Electrical workers, managerial and professional

These results suggest the hypothesis that the predictive validity of tests is related in part to the *g* demands of the job.

Another important conclusion from the Ghiselli monograph is that test validities for training criteria (e.g., course grades, instructor ratings, time required to meet training criteria) are considerably higher than for actual job proficiency criteria after training. Training criteria correlate close to .50 with IQ and other ability tests. Often the abilities that best predict success in training for a particular job are not the same abilities that best predict success on the job after training. The most *g*-loaded tests have their highest validity for predicting success in *training*. After training is completed, special abilities—numerical, spatial, perceptual, motor—gain in importance, relative to *g*, for predicting actual job performance.

Ghiselli's 1955 review includes a great variety of tests with different psychometric properties, a fact that could itself contribute to the great variability he observed in the validity coefficients for many jobs. The extensive validation studies carried out by the U.S. Employment Service using a single standardized test battery—the General Aptitude Test Battery, or GATB—overcomes this objection. Yet the results are still very similar to those by Ghiselli.

The GATB was devised according to factor analytic principles and yields scores on the following factors:

G - General Intelligence
V - Verbal Aptitude
N - Numerical Aptitude
S - Spatial Aptitude

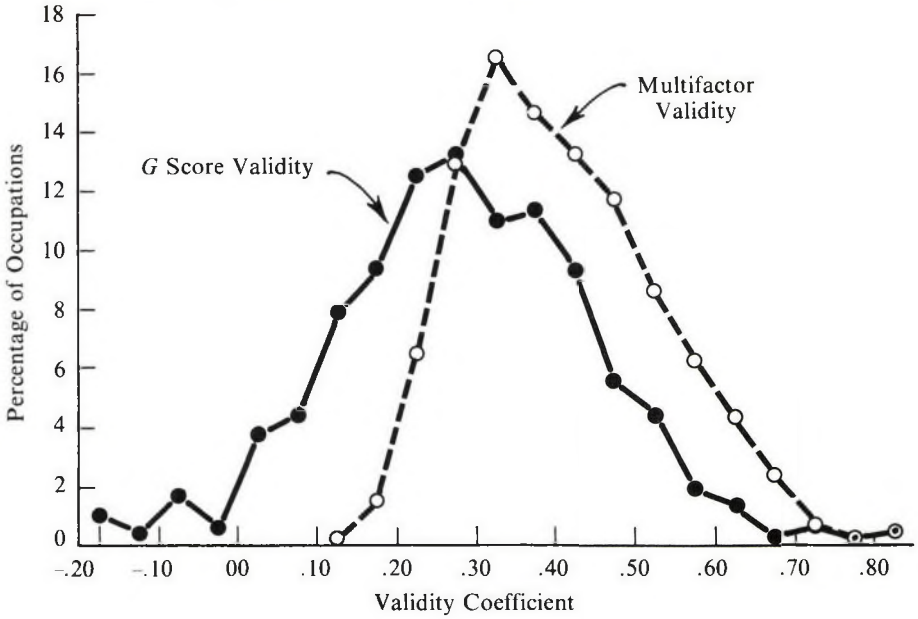
P - Form Perception
Q - Clerical Perception
A - Aiming } *K* - Motor Coordination
T - Motor Speed }
F - Finger Dexterity
M - Manual Dexterity

A total of 537 predictive and concurrent validity coefficients, based almost entirely on supervisory ratings, were obtained on large samples of workers in 446 different occupational categories listed in the *Dictionary of Occupational Titles*. Validity coefficients were determined on the basis of the optimally weighted composite of the various GATB scores for predicting performance ratings in each occupation, as well as for each of the ten separate factors. Figure 8.8 shows the frequency distributions of validity coefficients for the multifactor battery and for *G*, the general intelligence score, which is factorially about the same as scores on most standard IQ tests. It can be seen that multiple predictors yield somewhat higher validities than *G* alone. The median validity for multiple predictors is +.36; for *G* it is +.27.

Notice also that the distribution of validities for *G* has a wider spread than the distribution of multifactorial validities. In the latter case, of course, a different best-weighted combination of tests is used to maximize the validity for each occupation. Against this multifactor standard, *G* alone stands up remarkably well.

As noted in the Ghiselli review, there are occupational differences in validity coefficients, and these are much more highly related to *G* validities than to multifactor validities. In general, occupations with the greatest cognitive demands, the "knowledge

Figure 8.8. Frequency distribution of 537 validity coefficients of the General Aptitude Test Battery for 446 different occupations. *G score* is general intelligence; multifactor validity is based on a weighted composite of GATB subtests.



jobs'' and those requiring higher education or technical training show the highest *G* validities. Jobs showing negligible (or even negative) *G* validities are on the whole the least skilled and least complex jobs that involve rather simple, repetitious, or routine work, such as onion corer (−.15), metal chair assembler (−.19), and letter-opener operator (−.08).

It is also interesting to note that although the *G* validities are higher for training criteria than for job proficiency criteria, the multifactor validities show no difference in this respect. It is also interesting that *predictive* validites average slightly higher (by .04) than *concurrent* validities, probably because the latter are based on tests given to persons who have already survived in the given job for some time and they are therefore a more highly selected sample in terms of job performance and suitability for the job.

The fact that general intelligence correlates significantly with performance ratings in so very many ordinary jobs that have little or no formal educational requirements calls for closer psychological scrutiny into the specific aspects of these jobs that account for the correlation between job proficiency ratings and general intelligence.

The Human Resources Research Organization, better known as ''HumRRO,'' has done intensive research on just this question (Vineberg & Taylor, 1972). The test of general intelligence they used was the Armed Forces Qualification Test. The subjects of the study were 1,544 inducted and enlisted men in the army distributed about equally in four specific job categories: armor crewman, repairman, supply specialist, and cook. At the time of the study all the men were working daily in their jobs. Job experience ranged from one month to over twenty years. Job proficiency was measured by objective job-sample tests. These tests were made up by having persons in the jobs make up inventories of the specific activities involved in the performance of the job. Because the job of cook will be more familiar than the other jobs to most readers, it provides the best example. The inventory of duties contains twenty-nine items, for example, prepares cook's worksheet and other forms, takes inventory of food products and kitchen equipment, stores and inspects food, prepares beverage, cooks meat, fish, poultry, prepares desserts, cleans or disassembles equipment, cooks soups, and so on. Under each of these categories are a number of even more specific jobs, such as ''makes scrambled eggs,'' ''makes jellyroll,'' ''makes cocoa.'' These many specific jobs constitute the ''items'' of the job-sample test. (The tests contain 359 items for armor crewmen, 176 for repairmen, 156 for supply specialist, and 158 for cooks.) The person's score is simply the percentage of the total items that the person could perform unassisted. (Prompts were allowed on some of the more complex items, which were given part-scores that added to the total score only when prompts for the various subtasks were not needed.)

First, it was shown that there is a significant correlation between AFQT and scores on the job-sample tests. The partial correlations, which remove from the AFQT × job-sample correlation the effects of years of education, number of months on the job, and age, are the most relevant here. The partial correlations for the four jobs are armor crewman, .36; repairman, .32; supply specialist, .38; cook .35.

It is interesting to compare these partial correlations with the corresponding partial correlations between AFQT scores and supervisor ratings of job performance; armor crewman, .26; repairman, .15; supply specialist, .11; cook, .15. The AFQT correlates significantly higher with the objective job-sample tests than with supervisor ratings. (The

simple correlations between job-sample scores and supervisory ratings for the four jobs are .27, .20, .28, and .28.)

A paper-and-pencil job-knowledge test was also given. Its validity is attested to by its partial correlations (controlling for AFQT, education, and age) with *months on the job*, ranging from .48 to .66, with an average of .56 in the four jobs. The partial correlations (controlling education, age, and months on job) between job knowledge and AFQT scores were armor crewman, .54; repairman, .42; supply specialist, .37; cook, .47. Thus AFQT correlates somewhat higher with job-knowledge than with job-sample scores.

Finally, the most important from a theoretical standpoint are the analyses that highlight the aspect of job performance, as assessed by the job-sample tests, that is most responsible for its correlation with general intelligence as measured by the AFQT. Subjects were classified into four mental groups in terms of AFQT score, as follows:

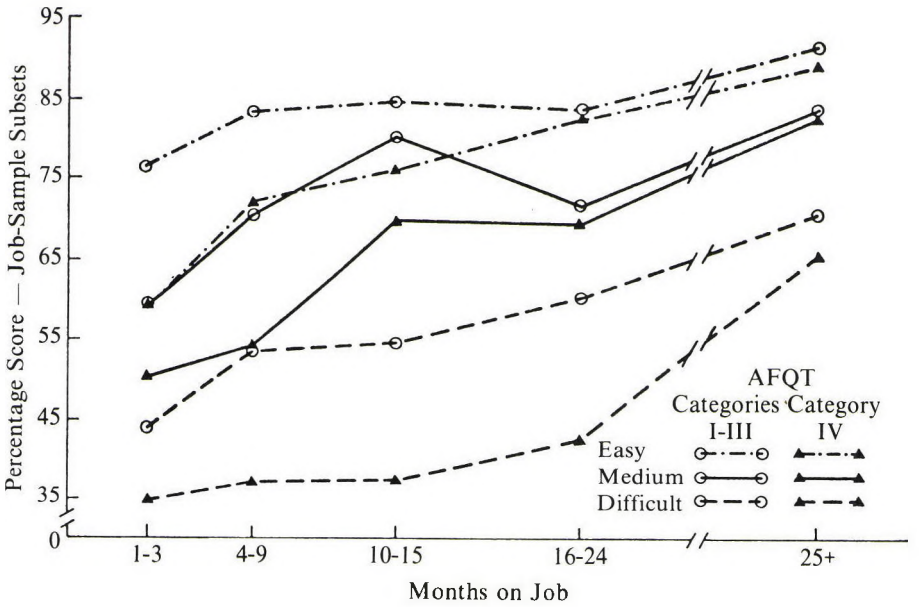
Mental Group	*AFQT Percentile*	*WAIS IQ Equivalent*
I	93–100	122 and above
II	65–92	105–121
III	31–64	93–104
IV	10–30	81–92

The AFQT percentiles are based on all U.S. males, ages 18 to 26, who have come up for the armed forces draft. The Wechsler Adult Intelligence Scale (Full Scale) IQs corresponding to the same percentiles in the general population are listed as a frame of reference. The mean job-sample scores of the AFQT groupings show clear separation out to at least five years of experience on the job, after which there is some, but not complete, convergence of mean scores.

Correlations between AFQT and job-sample scores are mainly a function of the empirically determined difficulty level of the subtests. The easiest job-sample problems (in terms of percentage of subjects passing the item) discriminated the least between the AFQT Mental Groups. The more difficult and complex items discriminated more highly between the groups. For example, comparing the percentage of cooks in category IV with the percentage in categories I to III who can scramble eggs, an easy task, we see a nonsignificant difference of 77.5 percent versus 79.0 percent. For the somewhat more complex task of making a jellyroll, there is a significant difference of 59.9 percent versus 70.3 percent. In fact, the only single subtest in all of the 849 subtests for the four jobs that showed no significant difference between the AFQT categories was making scrambled eggs—the easiest of all the test items for cooks! A graph of the percentage passing the repairman's job-sample subtests as a function of AFQT Mental Category, task difficulty, and months on the job clearly illustrates the greater separation of the AFQT groups on the difficult than on the easy job performance items, a difference that persists to some degree throughout all months on the job, as shown in Figure 8.9. Similar trends were found in the other three job categories.

The trends of the means, which are theoretically important for understanding the nature of general ability, should not detract from the considerable amount of overlap of the AFQT mental groups in their performance on the various jobs. Substantial percentages of

Figure 8.9. Mean scores on job-sample tests for repairman as a function of task difficulty (easy, medium, or difficult), intelligence level (AFQT categories), and number of months on the job. (From Vineberg & Taylor, 1972, p. 56)



every mental group performed above and below the median on the job-sample tests. With increasing months of job experience, an increasing percentage in the AFQT Group IV were able to perform at acceptable levels on these particular army jobs.

Another army study carried out by HumRRO psychologists (Fox & Taylor, 1967) demonstrated most clearly the interaction between general intelligence level and task complexity, even when the tasks are extremely simple and variation in degree of task complexity is minimal. Two artificial "jobs" were devised, both at a simple stimulus–response level of performance, but one task involved only simple reaction time, the other complex reaction time. The authors described these two tasks as follows:

> [S]equential monitoring tasks which fall at the simplest level of complexity. In fact, they are so simple that no learning is required for performance. These tasks have elements in common with many military jobs. . . . Task 1 (T_1) is a Simple Sequential Monitoring Task. The trainee was told that his "control" panel was part of a communications system that became overloaded when a red light came on. His task was simply to "reset" the control panel by pressing the lever when a red light appeared. The control panel apparatus was programmed so that white lights flashed intermittently across the panel accompanied by loud clicking noises. After an interval which varied from 15 to 205 seconds, the white lights went out and one of the four red lights came on. The trainee was required to "reset" the panel a total of twenty times over a forty-minute period. The second task (T_2) is a Choice Sequential Monitoring Task and uses the same apparatus as the previous task except for additional response levers. The trainee was to respond to one of the four red lights,

labeled A, B, C or D, by pressing the corresponding lever. All procedures and programming were identical for both tasks.

The subject's performance was measured in terms of reaction time, that is, the time interval between the appearance of the red light and the subject's pressing the lever.

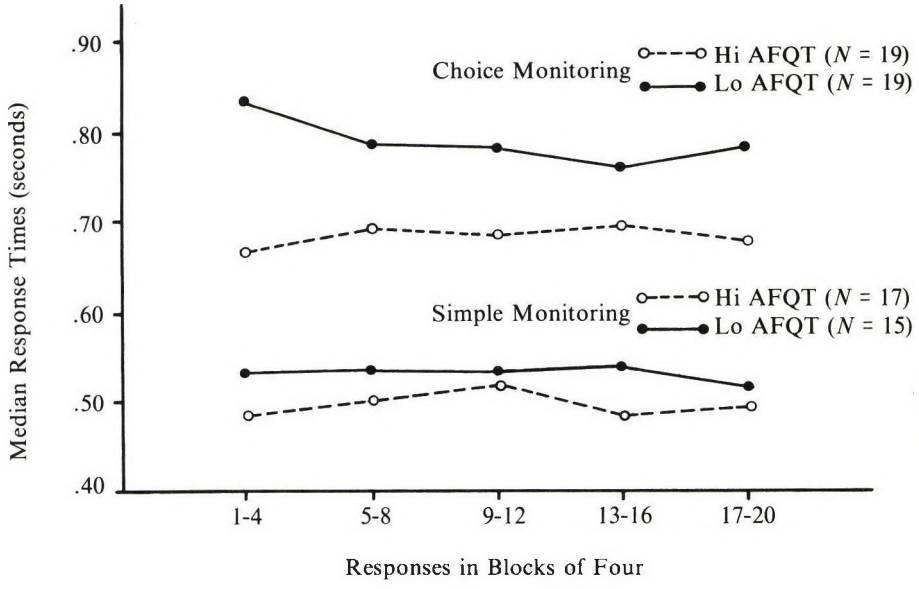
The tasks were given to two groups of army recruits from mental categories I and IV on the AFQT, labeled "Hi" and "Lo" AFQT, respectively. The results are shown in Figure 8.10. Notice that for both AFQT groups choice monitoring resulted in greater response times than simple monitoring and that the separation between the Hi and Lo AFQT groups was greater for the slightly more complex task, even at this relatively simple level. This result shows that subjects who are selected on general intelligence differ even on performance of very simple tasks. The general ability factor thus extends over an enormous range of complexity and types of performance.

IQ and Creativity

In recent years, the popular psychological and educational literature has promulgated the notion of "creativity" as a psychological trait quite distinct from, or even opposed to, general intelligence. The belief is probably born of the hope that, if a person is deficient in intelligence, there is a chance that he may possess an abundance of something at least equally valuable—*creativity*. However, there is no sound scientific basis for this hope.

The term "creativity" is in quotes because it means so many different things to different investigators and has no standard operational definition. The use of the term in the psychological literature often refers to types of behavior that scarcely correspond to

Figure 8.10. Reaction times on simple and choice monitoring tasks by high and low scorers on the Armed Forces Qualification Test. (From Fox & Taylor, 1967)



what the layman means by creativity. The laymen's concept of creativity generally involves the characteristics both of *originality* and *quality* of performance. To devise objective measures of "creativity" that have as little correlation with IQ as possible, psychologists have had to leave out of consideration altogether the concept of quality of performance.

There exists at present no validated test of "creativity" in the sense of being able to predict who will be socially judged as creative in the arts or sciences. The existing tests of "creativity" are more accurately referred to as tests of *divergent thinking*. The usual IQ tests involve mainly *convergent thinking*. That is, the mental manipulations called forth by a given test item lead to a single correct solution. Items in a test of divergent thinking, in contast, are intended to lead to a large number and diversity of possible answers, none of which is either correct or incorrect. The person's responses are scored for *fluency* (i.e., number of responses to a question), *flexibility* (i.e., diversity of responses), and *originality* (i.e., uncommonness of the responses). A typical example of a divergent thinking test item is the "unusual uses" type of item, such as the question "How many uses can you think of for a brick?" Answers such as the following would score low on flexibility and originality: "To build a house," "To build a fireplace," "To build a wall," "To pave a walk." Higher-scoring answers would be "As a doorstop," "A pillow for an ascetic monk," "An abrasive to strike matches on," "As a bedwarmer or footwarmer after heating it in a stove."

Critical reviews of attempts to measure creativity have concluded that various creativity tests show hardly any higher correlations with one another than with standard tests of intelligence (Thorndike, 1963b; Vernon, 1964). The *g* factor is common to both kinds of tests, and there seems to be no independent substantial general factor that can be called creativity. Besides *g*, creativity tests involve long-recognized smaller group factors usually labeled as verbal and ideational fluency. Differences between persons scoring high and persons scoring low on "creativity" tests, when they are matched on IQ, invariably consist of descriptions of *personality* differences rather than of characteristics that would be thought of as any kind of *ability* differences. Thus "creativity," at least as presently measured, apparently is not another type of ability that contends with *g* for importance, as some writers might lead us to believe (Getzels & Jackson, 1962; Wallach & Kogan, 1965). While "creativity" tests may be related to certain personality characteristics, they have not been shown to be related to real-life originality or productivity in science, invention, or the arts, which are what most people regard as the criteria of creativity.

For a time it was believed that the research of Wallach and Kogan (1965) contradicted the conclusion of earlier reviews to the effect that "creativity" and intelligence are not different co-equal abilities or even factorially distinguishable traits. Wallach and Kogan had claimed that the failure of earlier researches to separate creativity and intelligence was a result of the fact that the creativity tests were usually given in the same manner as the usual psychometric tests, with time limits, as measures of some kind of ability, in an atmosphere conducive to competitiveness and self-critical standards. These conditions, it was maintained, were antithetical to the expression of creativity. So Wallach and Kogan gave their tests of creativity (better labeled as *fluency*) without time limits, in a very free, nonjudgmental, play-like, game-like atmosphere. Under these conditions, they found negligible correlations between several verbal intelligence tests and their "creativity" tests. The "creativity" scores, however, account for only a small percentage (2 per-

cent to 9 percent) of the variance in any of the dependent variables measured in this study. High scorers in general tended to be less inhibited or less constricted in producing responses; they responded more energetically and fluently in the game-like setting in which the "creativity" tests were given.

In a penetrating and trenchant methodological critique and reanalysis of the Wallach–Kogan data, Cronbach (1968) concluded:

> My final impression is that the *F* [i.e., fluency or "creativity"] variable has disappointingly limited psychological significance. It can scarcely be considered a measure of ability or creativity; there is no evidence that high *F* children produce responses of superior quality in any situation. It is correlated with other measures of social responsiveness, but not strongly. (p. 509)

In one of the earliest works on the psychology of creativity, Spearman (1930) argued that socially recognized creativity exemplified the very process that most characterizes *g*—"the eduction of relations and correlates." Creativity could be characterized as "the eduction of *new* relations and correlates." This is the essence of invention and innovation—in science, in the arts, in politics. According to Spearman, a high level of *g* is a necessary but not sufficient condition for creativity in the nontrivial sense. Creative persons may possess certain traits of personality and character not found as often in noncreative persons, but none is poorly endowed on *g*.[4]

Most present-day researchers on creativity, such as MacKinnon, Barron, and Torrance, acknowledge the *threshold relationship* of intelligence to creativity. Beyond a certain threshold level of intelligence, which is probably about one standard deviation above the mean IQ of the general population, there is little relationship between IQ and rated creativity; below the threshold there is simply no creativity to speak of in any culturally significant sense. Below the threshold, tests of "creativity" are factorially hard to distinguish from tests of general intelligence.

If there were actually no relationship of any kind between creativity and intelligence, as some popular writers would have us believe, we should expect to find the same proportion of mentally retarded persons (with IQs below 70) among the acknowledged creative geniuses of history as is found in the general population. Biographical research on the childhoods of famous creative persons in history, however, has revealed that in 300 cases on whom sufficient data were available, all of them without exception showed childhood accomplishments that would characterize them as of above-average intelligence, and the majority of them were judged to be in the "gifted" range above IQ 140 (Cox, 1926).

Donald W. MacKinnon has actually obtained the Wechsler (WAIS) Full Scale IQs of 185 noted architects, mathematicians, scientists, and engineers who were selected from a national sample on the basis of ratings by other professionals in these fields as being among the most creative contributors to these socially significant fields (MacKinnon & Hall, 1972). In this highly select group, the judged ranking in creativity correlated only +.11 with WAIS IQ. But the more important fact, which is often neglected in popular accounts, shows the threshold relationship of IQ to creativity: the total IQ variance in this group of creative persons is *less than one-fourth* of the IQ variance in the general population. The entire creative sample ranges between the 70th and 99.9th percentiles of the population norms in IQ (i.e., IQs from 107 to 151), with the group's *mean* at the 98th

percentile (IQ 131). To the extent that these groups are typical of persons whom society regards as creative, it can be said that some 75 to 80 percent of the general population would be excluded from the creative category on the basis of IQ alone. Despite claims that IQ has little or nothing to do with creativity, no one has ever published the crucial evidence that would necessarily follow from this position: a distribution of IQs among persons of generally acknowledged creativity that differs nonsignificantly from the distribution of IQs in the population. It would be absolutely astounding if such evidence could ever be found.

Nonintellectual Correlates of IQ

Intelligence tests have never been specifically devised to measure anything other than general intellectual ability and scholastic aptitude, and so it is especially interesting to note that, despite this fact, IQ has a number of nonintellectual and nonacademic behavioral correlates. The *causal* connection between IQ and nonintellectual behavior, such as personality adjustment, social responsibility, delinquency, and crime, is complicated. The links in the chain of causality are not at all clearly worked out in most studies. Social-class and cultural differences are undoubtedly involved to some degree in such correlations, but they are not the whole story. Frustration due to repeated failures in a society that makes many intellectual demands, particularly during the school years, can lead, for those who are less able to compete intellectually, to aggression, or withdrawal, or other forms of socially maladaptive behavior.

In a highly organized social system of a technological bent that tends to sort out people according to their abilities, and rewards them more or less accordingly, it seems not surprising to find that those traits of personality and temperament that complement and reinforce the development of intellectual skills requiring persistent application, practice, freedom from emotional distraction, and resistance to mental fatigue and boredom in the absence of physical activity, should become genetically assorted and segregated, and thereby be genetically correlated with the socially valued mental abilities that require the most education for their full development and utilization in the world of work.

Thus ability and personality traits tend to work together in determining a person's overall capability in the society. For example, Cattell (1950) has found that certain personality traits are correlated to the extent of about 0.3 to 0.5 with the general mental ability factor. He concludes: "There is a moderate tendency . . . for the person gifted with higher general ability, to acquire a more integrated character, somewhat more emotional stability, and a more conscientious outlook. He tends to become 'morally intelligent' as well as 'abstractly intelligent'" (pp. 98–99). The connection between intelligence and moral behavior has been further investigated in recent years by Harvard psychologist Laurence Kohlberg (1969), who concludes that a person's complexity and maturity of moral judgments depend in large part on his level of general cognitive ability. Because the correlation, though substantial, is not perfect, we can all, of course, point out notable exceptions.

Adjustment and Adaptive Behavior. "Adjustment" is a broad term in psychology and mental hygiene, referring to a complex of behaviors involving such features as emotional stability, freedom from neurotic symptoms, responsibility, getting along with people, social participation, realistic self-confidence, absence of socially disruptive and

self-defeating behavior, healthy attitudes about sex and bodily functions, marital success, possessing more than superficial or fleeting interests and values (whatever they may be), and displaying a capacity for self-discipline and planful and sustained goal-directed effort. What psychoanalysts term "ego strength" is an amalgam of these indices of "adjustment." Adjustment is usually assessed by means of interviews, detailed ratings by parents, teachers, or peers, and self-report adjustment inventories or questionnaires.

A number of studies (see reviews by Anderson, 1960; Kohlberg, LaCrosse, & Ricks, 1970; White et al., 1973, pp. 207–209) have found substantial correlations between IQ and various assessments of adjustment. The validity of IQ for predicting adjustment generally ranges from 0.4 to 0.6, even when several years intervene between the IQ and adjustment assessments. IQs measured in Grades 6 to 9, for example, show correlations of about 0.50 with various assessments of adjustment in early adulthood. A review of longitudinal studies relating IQ to later indices of adjustment reached the following conclusion: "[A] crude quantitative estimate of the predictive power of IQ is the statement that 20% to 30% of the reliable variation in gross ratings or estimates of adjustment in a representative sample of adults can be predicted from elementary school IQ scores" (Kohlberg et al., 1970). It was also noted that high IQ predicts very good adjustment somewhat better than low IQ predicts very poor adjustment.

Curiously, two of the most extreme forms of maladjustment—psychosis and suicide—seem to be unrelated to IQ. Terman's "gifted" group of 1,500 persons with IQs above 140, for example, showed at age 40 better than average adjustment on all criteria considered except the incidence of psychosis and suicide, which were about the same as in the general population (Terman & Oden, 1959).

The extent to which there is a direct *causal* connection between IQ level and adjustment is still obscure. All we can say for certain on the basis of the present evidence is that IQ *predicts* adjustment to some extent as a result of both IQ and adjustment being correlated elements in a complex causal network involving other factors such as social class, cultural values, styles of child rearing, physical health and appearance, and probably some degree of criterion contamination (i.e., indicants of intelligence, per se, influencing ratings of adjustment).

Adaptive behavior is somewhat akin to adjustment, but it also involves to a greater extent the implication of personal and social *competence*. In the words of Matarazzo (1972, pp. 147–148):

> *Adaptive behavior* refers primarily to the effectiveness with which the individual copes with, and adjusts to, the natural and social demands of his environment. It has two principal facets: (a) the degree to which the individual is able to function and maintain himself independently, and (b) the degree to which he meets satisfactorily the culturally imposed demands of personal and social responsibility. It is a composite of many aspects of behavior . . . [subsumed] under the designation intellectual, affective, motivational, social, motor, and other noncognitive elements [that] all contribute to and are a part of total adaptation to the environment.

The concept of adaptive behavior, also referred to as social maturity, has evolved largely in connection with the diagnosis of mental retardation. It is now generally agreed that the diagnosis of mental retardation must be based on a broader set of criteria than just performance on an IQ test. The list of additional criteria are termed indices of adaptive

behavior, and a number of adaptive behavior rating scales including these criteria have been devised to improve the reliability and objective validity of assessments of these forms of adaptive behavior.

The American Association on Mental Deficiency has expended considerable research effort in the development of adaptive behavior rating scales (Nihira, Foster, Shellhaas, & Leland, 1969). These scales consist of more than a hundred specific descriptive behavioral items involving three broad factors: personal independence, social maladaptation, and personal maladaptation. Because the scales are intended primarily for use with the mentally retarded, the items are much more discriminating in the lower half of the IQ distribution than in the upper half. Adaptive behavior scales necessarily have too low a ceiling for the above-average segment of the population. Many items, for example, involve quite simple everyday skills such as handling money, personal care and hygiene, telling time, domestic skills, ability to go shopping alone, and the like. (In normal children items of this kind are correlated .60 to .70 with Stanford-Binet mental age and some 117 such items have been age graded to form the well-known Vineland Social Maturity Scale; see Doll, 1953, 1965.)

Adaptive behavior, as rated on such scales, is substantially, but far from perfectly, correlated with IQ among retardates. Correlations in a number of institutional samples range from .58 to .95 (Leland, Shellhaas, Nihira, & Foster, 1967, p. 368). (See also Chapter 14, pp. 681-685.)

School Deportment. A study of 7,119 school children aged 6 to 11, by Roberts and Baird (1972), shows a relationship between a pupil's intelligence (as rated by their teachers) and the frequency with which the teacher reports the pupil's behavior in school results in disciplinary action on the teacher's part. At all ages, for both boys and girls, there is a negative relationship, corresponding to an overall correlation of about $-.30$, between rated intelligence and frequency of disciplinary action. Although disciplinary action was less frequent for girls than for boys, it had about the same correlation with intelligence as was found for boys. A possibly serious shortcoming of this study, of course, is that, because both the assessments of intelligence and of frequency of disciplining were based on teacher judgments, there could be an undetermined degree of criterion contamination★ or "halo effect"★ involved in the correlation between these two variables. However, other evidence on the negative correlation between measured IQ and delinquent behavior suggests that the correlation between teacher ratings of intelligence and deportment is not mainly due to a halo effect.

Activity Level in Early Childhood. Harvard psychologist Jerome Kagan was the first to report the observation of a *negative* correlation between degree of motoric hyperactivity (hyperkinesis) in young children (ages 3 to 6 years) and their intellectual level at maturity (Kagan, 1971; Kagan, Moss, & Sigel, 1963). It is as if the inability to inhibit gross motor activity in early childhood interferes with development of the capacity for sustained involvement in cognitive tasks. Hyperactivity, impulsivity, and short attention span all work together to hinder the child's acquiring as much information as he should from interaction with the physical and social environment.

Another possible explanation of the negative correlation between degree of early hyperactivity and later cognitive ability is that both variables are related to some third variable that mediates the correlation. This hypothesis is suggested by some excellent recent research by Halverson and Waldrop (1976), who found that early childhood activity

level is "highly related to an index of minor physical anomalies" (p. 107). The correlation between physical anomalies, including motor coordination problems, and activity rating in children at $2\frac{1}{2}$ years of age is .51; in the same children at $7\frac{1}{2}$ years of age the correlation is .44. All the sixty-two unselected nonclinical children in this longitudinal study were white middle class, from intact families, who were attending a nursery school. Activity level during free play was rated through systematic observations by trained observers and was also measured objectively by means of a mechanical activity recorder fastened to the child's clothing during free play periods. (When included in a factor analysis of the observers' ratings of activity level, according to various criteria, the activity recorder has a factor loading of .83 on the general factor of activity level.)

Measurements taken from the activity recorder and behavior ratings during free play at age $2\frac{1}{2}$ years show a highly significant correlation of $-.47$ with the Full Scale WISC IQ obtained at age $7\frac{1}{2}$ years. (Verbal and Performance IQs correlate with activity level $-.38$ and $-.40$, respectively.) These are remarkably high correlations, considering that the correlations of IQ with itself between the ages of $2\frac{1}{2}$ and $7\frac{1}{2}$ is only about .50. In other words, at age $2\frac{1}{2}$, objectively measured motoric activity level during free play predicts IQ at age $7\frac{1}{2}$ almost as well as an IQ test itself given at age $2\frac{1}{2}$. The correlation between activity level and later IQ is negative, which means that the young children who show the most fast-moving, vigorous, impulsive behavior during play turn out, on average, to have the lower IQs later on.

There was found to be considerable stability of individual differences of activity level over the five-year period between ages $2\frac{1}{2}$ and $7\frac{1}{2}$. Activity ratings at age $7\frac{1}{2}$ still correlate significantly ($r = -.31$) with IQ at age $7\frac{1}{2}$. The highly g-loaded Embedded Figures Test (given at age $7\frac{1}{2}$) also shows a significant negative correlation ($-.34$) with activity level at age $2\frac{1}{2}$.

Delinquency and Criminal Behavior. A number of studies show that IQ is associated with delinquency and criminality within the white population (Burt, 1925; Caplan, 1965; Glueck & Glueck, 1950; Gordon, 1975; Merrill, 1947; Siebert, 1962). Delinquents with court records average some 10 to 12 IQ points below the mean IQ of nondelinquents. Delinquents come preponderantly from the lower half of the IQ distribution and as a group average only about 3 IQ points higher than the mean IQ of 89 that would be obtained by excluding all IQs above the mean of the general population. Recent research by sociologist Robert A. Gordon (1975b, 1976) presents strong evidence that delinquency is related to IQ to much the same degree in the black as in the white population and that racial, ethnic, social-class, and regional differences in the prevalence of delinquency are highly predictable from the mean IQs of the persons comprising each of these groups. *Across various racial and social-class groups, the prevalence of delinquency is approximately the same at any given IQ level.* In other words, if one controls for IQ, the marked racial and social-class differences in delinquency rates disappear. Minority racial and ethnic groups with mean IQs at or above the general population mean, such as Orientals and Jews, show correspondingly lower rates of delinquency to the same extent that minorities with mean IQs below the population average show correspondingly higher rates. (It should be noted that even quite small but statistically significant correlations between two variables based on individual measurements can result in extremely high correlations between group means on the two variables.) From such findings, Gordon (1975b, 1976) argues that general cognitive ability, as indexed by IQ, must be regarded as a

central variable in the development of a scientific theory of delinquency and criminality. This viewpoint has recently been strongly reinforced by an excellent review of the research on intelligence and delinquency by Hirschi and Hindelang (1977), who show that IQ has an effect on delinquency independent of class and race.

Juvenile delinquency and adult criminality show a negative curvilinear relationship to IQ, with delinquency and crime rates diminishing markedly below IQ 50 and above IQ 100. (Just the simple within-race linear correlation (point-biserial r) between court-recorded delinquency and IQ is between $-.4$ and $-.5$.) The highest rates of delinquency and crime fall in the IQ range from 70 to 90. Apparently the majority of persons below IQ 50 are either under close enough supervision by parents and relatives to be kept out of serious trouble or are too incompetent or socially isolated to become involved in the kinds of serious delinquent activities that would come to legal attention. To steal an automobile, for example, one has to be at least smart enough to be able to break in, start the car without a key, and know how to drive—all skills that it would be rare to find in a person below IQ 50.

IQ does not predict delinquency rates across the sexes. Males and females do not differ in IQ, but males show much higher rates of delinquency and adult crime than females. Within each sex, separately, however, there is about the same degree of relationship between IQ and delinquency.

Is the association between IQ and delinquency explainable by the fact that both are correlated (in opposite directions) with social class, low income, and poverty? Apparently not, as research on full siblings reared together in the same families shows almost the same degree of association between IQ and delinquency as is found in the general population (Healy & Bronner, 1936; Shulman, 1929, 1951). Delinquents show lower IQs, on the average, than their nondelinquent siblings of the same sex. Hirschi and Hindelang (1977) have hypothesized that the child's school experience mediates the correlation between IQ and delinquent behavior. A similar theory, with some supporting evidence that delinquency is often a reaction to a learning disability in school, has been advanced by Berman (1978).

Obviously many factors in addition to IQ must be involved in delinquency and antisocial behavior, as the majority of persons at every level of IQ and in every race and social class are nondelinquent. A low IQ is neither a necessary nor a sufficient condition for delinquent behavior. But there is a heightened probability of delinquency in the low-IQ child, even as compared with his or her own siblings of higher IQ. The correlation is most likely mediated by the frustrations arising from possessing less than average general ability, with its consequences of more frequent failures in competition with age peers and in winning recognition and approval from significant persons in the environment. The school, as presently constituted, is generally a potent source of such frustration for children in the lower quarter of the IQ distribution, that is, IQs below 90.

Miscellaneous Behavioral Correlates of IQ. Intelligence test scores have been shown to have significant low to moderate positive correlations with a variety of other variables, such as *honesty* (Mussen, Harris, Rutherford, & Keasey, 1970); nonacademic attainment in *extra-curricular activities* (Kogan & Pankove, 1974); children's *appreciation of humor* as judged from response to cartoons of varying subtlety and sophistication (Zigler, Levine, & Gould, 1966); ability to solve *anagrams* (Gavurin, 1967); untrained *musical aptitude* (Wing, 1941); speed of learning a number of relatively complex (but not

simple) *motor skills* (Noble, 1974); susceptibility to certain *optical illusions* and various *perceptual phenomena* (studies reviewed by Honigfeld, 1962); and amount of specific *information retained* from viewing a television feature program, especially the incidental, unemphasized bits of information (Nias & Kay, 1954).

Physical Correlates of IQ

A number of anthropometric and physiological measurements show reliable small to moderate correlations with measured intelligence.

Brain Size. A most thorough and methodologically sophisticated recent review of all the evidence relevant to human brain size and intelligence concludes that the best estimate of the within-sex correlation between brain size and IQ is about 0.30, taking proper account of physical stature, birthweight, and other correlated variables (VanValen, 1974). Such a correlation is considered quite important from a biological and evolutionary standpoint, considering that much of the brain is devoted to noncognitive functions. The author argues that there has been a direct causal effect, through natural selection in the course of human evolution, between intelligence and brain size. The evolutionary selective advantage of greater brain size was the greater capacity for more complex intellectual functioning. "Natural selection on intelligence at a current estimated intensity suffices to explain the rapid rate of increase of brain size in human evolution" (VanValen, 1974, p. 417).

Brain size is correlated with head size, and thus it is noteworthy that the Harvard anthropologist Ernest Hooton (1939) found that the head circumferences of Boston whites in various occupational levels are in about the same rank order as is usually found when occupations are ranked according to their average IQs, as shown in Table 8.8. A chi squared test shows that the means of the eight occupational categories differ significantly ($\chi^2 = 84.4$, $df = 7$, $p < .001$).

Brain Waves. IQ is correlated with various indices involving the speed and amplitude of electrical potentials in the brain, evoked by visual and auditory stimuli, and measured by the electroencephalogram (Callaway, 1975). This topic is considered in greater detail in Chapter 14, pp. 707–710.

Stature. In American and European Caucasian populations there is a significant low within-sex correlation (ranging in various studies from about .1 to .3, with an average

Table 8.8. Head circumferences (in millimeters) of Boston whites in various occupational categories. (From Hooton, 1939)

Occupational Category	*N*	Mean	SE_M
Professionals	25	569.9	1.9
Semiprofessionals	61	566.5	1.5
Clerical	107	566.2	1.1
Trades	194	565.7	0.8
Public service	25	564.1	2.5
Skilled trades	351	562.9	0.6
Personal services	262	562.7	0.7
Laborers	647	560.7	0.3

of about .25) between IQ and physical stature (Stoddard, 1943, p. 200; Paterson, 1930). This correlation almost certainly involves no causal or functional relationship between stature and intelligence but is a result of the common assortment of the genetic factors for both height and intelligence. These are both perceived in our society as desirable characteristics, and there is a fairly high degree of assortative mating for both characteristics. This results in a between-families genetic correlation between the traits. There appears to be no within-families correlation, as indicated by the fact that, on the average, there are no differences in height or other physical characteristics between gifted children (average IQ 141) and their nongifted siblings (average IQ 109), yet gifted children, on the average, are taller for their age and have generally better physiques than the average child (Laycock & Caylor, 1965). This finding is precisely what we should expect if the correlation between stature and intelligence is a between-families correlation, due to common genetic assortment of the two traits, rather than a correlation due to pleiotropy, genetic linkage, or functional relationship. (See Chapter 6, pp. 193–195.)

Basal Metabolic Rate and Obesity. The evidence is somewhat equivocal on BMR, showing correlations with IQ running from close to zero to as high as .80 in various studies. It appears that there may be significant correlations in childhood, during the most rapid growth period, and that the significant correlations diminish and finally disappear from adolescence to adulthood (Stoddard, 1943, pp. 206–407; Tyler, 1965, p. 429).

The diagnosis of obesity (defined as 20 percent or more overweight for age, height, and build) has been found to have a quite marked inverse relationship to IQ in women (Kreze, Zelina, Juhas, & Garbara, 1974). The percentages of women in the lower and upper quartiles of IQ who were classified as obese are 41.4 percent and 10.7 percent, respectively. (The corresponding percentages for men are 17.0 percent and 9.3 percent.) The negative relationship between IQ and obesity is more likely mediated in large part by the variable of social class, which was uncontrolled in this study, but which is known from other studies to be correlated (in opposite directions) with IQ and obesity.

Myopia. Near-sightedness or myopia is believed to be attributable to genetic factors, most probably recessive inheritance with full penetrance. Myopia is quite markedly associated with higher IQ. Myopes average about 8 IQ points higher than nonmyopes. Because no purely environmental explanation for this striking relationship (a point-biserial correlation of about .25 between IQ and the diagnosis of myopia) has been found to withstand critical scrutiny in light of evidence, it is suggested that there is a pleiotropic effect of the myopia gene on intelligence. The evidence has been reviewed by Karlsson (1978, Chs. 9 and 10), who concludes that "the myopia gene has an important stimulant effect on brain activity. It thus becomes the first identified specific gene which appears to contribute significantly to intelligence" (p. 78).

SUMMARY

Test validity is the extent to which scientifically or practically useful inferences can be drawn from test scores to behaviors outside performance on the test itself. The four main types of validity, known as the *four C*'s, are content validity, criterion validity, concurrent validity, and construct validity. Each is appropriate for a particular purpose. The methods for determining a test's validity, of course, depend on the type of validity.

For the practical use of tests in educational and employment selection, criterion (or predictive) validity is the most important; it is indexed by the correlation coefficient between test scores and some measure of the criterion performance.

The validity coefficient is open to several different statistically correct interpretations, depending on the purpose for which the test is used. The practical efficiency or utility of a test used for selection depends on more than just its validity coefficient; the test's utility depends also on the selection ratio, that is, the proportion on the total number of applicants who can be selected and the success–failure ratio of randomly selected applicants. Under certain realistic conditions, even tests with only moderate criterion validity can have great utility.

Because much of the debate concerning bias in mental tests involves the concept of *differential validity* of a test for minority and majority groups, it is essential fully to understand the meanings of validity in its technical sense.

Validity coefficients, and particularly their practical interpretations, have some degree of situational specificity, involving not only the test itself but also the reliability and validity of the criterion and the nature of the population and the circumstances in which the test is used. However, the situational specificity of test validity has been exaggerated and overemphasized in the past. It is now realized that much of the apparent specificity of test validity, as indicated by fluctuations of the raw validity coefficient from one study to another, is due to statistical artifacts; and, when these are taken into account, test validity is quite generalizable across situations (Schmidt & Hunter, 1977).

A review of the correlates of highly *g*-loaded tests, such as standard IQ tests, reveals that *g* has many correlates with variables outside the realm of the tests themselves, probably more correlates with more far-reaching personal and social significance than any other psychological construct. IQ alone predicts scholastic performance better than any other single variable or combination of variables that psychologists can measure. This is especially true of performance in the more academic school subjects, such as reading comprehension, mathematics, and written composition. The high predictive validity of *g*-loaded tests in this sphere is not at all due to common learned content between the tests and the school subjects, but to essential mental processes common to both spheres. Test performance and scholastic performance both involve essentially the same *g* factor of mental ability of acquired scholastic knowledge and skills. Thus there is a clear conceptual distinction between mental ability and scholastic attainments, even though these variables are highly correlated with one another. The validity coefficients of IQ tests for predicting school grades or achievement as measured by tests decrease from elementary school to high school and from high school to college, for reasons extraneous to the tests themselves. It is important to understand the several conditions that spuriously lower the obtained validity coefficients of IQ tests at each successively higher stage of the educational ladder.

IQ is correlated only slightly with simple rote learning and memory abilities but shows higher correlations with forms of learning that permit transfer from previous learning, insight, seeing relationships, spontaneous organization of complex material, and the like.

IQ also has important nonscholastic correlates: occupational level attained in adulthood and income level (within occupations). IQ has a threshold property for success in many occupations. There is some point on the IQ scale below which the probability of success in a particular occupation is practically nil. For persons above the threshold of IQ

needed for a given occupation, other personal factors besides IQ become relatively more important determiners of success. Amount of education, which is highly related to IQ, mediates much of the correlation between IQ and occupational status, but this cannot be validly interpreted to imply that amount of education per se is the *cause* of occupational status. Educational attainment, however, is causally dependent on IQ.

IQ is also correlated with a host of other traits and behaviors in which often the causal connections are still quite obscure: leadership qualities, socially recognized creativity, personality adjustment, social competence, and adaptive behavior. IQ is negatively correlated with proneness to delinquency. IQ also has a number of physical correlates, including brain size, amplitudes and latency of evoked brain electrical potentials, stature, metabolic rate in childhood, obesity, and myopia. It is clear that IQ tests and other highly *g*-loaded tests measure something considerably more profound and far-reaching than merely knowledge and skills acquired in school or in a cultured home.

NOTES

1. In comparing a multiple correlation R with a simple correlation r, it is proper to correct the multiple R for "shrinkage." Multiple R capitalizes on mere chance association, the more the greater the number of predictor variables k and the smaller the number of subjects N. When the number of variables is as large as the number of subjects, the multiple correlation can be perfect ($R = 1$), even when the true correlation between the predictors and the criterion is zero. The correction for shrinkage is

$$R_c = \sqrt{1 - (1 - R_o^2)\frac{N - 1}{N - k - 1}},$$

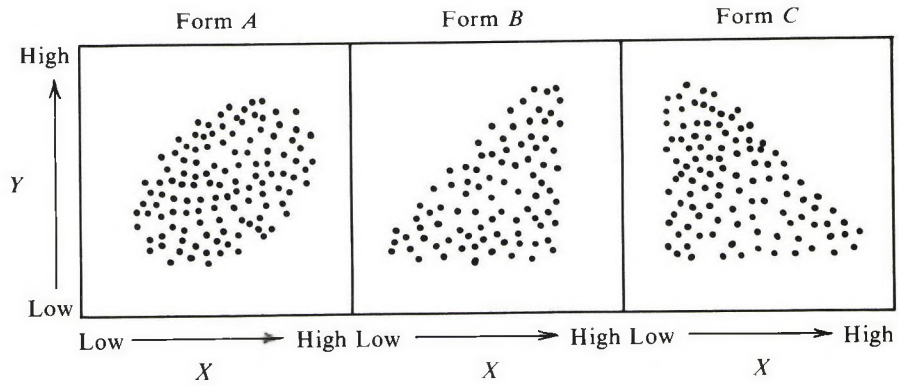
where

R_c = the shrunken (or corrected) multiple correlation coefficient,
R_o = the observed multiple correlation,
N = the sample size, and
k = the number of predictor variables used in the multiple correlation.

2. The correlation scatter diagram can take many forms other than the most common form, an ellipse. Three of the forms are shown in Figure 8N.1. (All represent positive correlations.) In each case the form of the correlation between variables X and Y can be described in terms of "necessary/sufficient." In form A (the usual ellipsoid scatter diagram), X is *necessary and sufficient* for Y. In form B, X is *necessary but not sufficient* for Y. In form C, X is *sufficient but not necessary* for Y. (In the case of zero correlation, i.e., a circular scatter diagram, it would be said that X is *neither necessary nor sufficient* for Y.) Notice that in form A all values of X are equally predictive of Y (a condition known in statistics as homoscedasticity), whereas in form B low values of X are more highly predictive of Y than are high values of X, and in form C high values of X are more predictive of Y than are low values of X. (Forms B and C are called heteroscedastic.)

The scatter diagram of forms B and C have been called "twisted pear" correlations, because the shape of the scatter diagram resembles a silhouetted twisted pear. For

Figure 8N.1. Types of scatter diagrams. Form *A* is bivariate normal. Forms *B* and *C* are "twisted pear" correlations and are heteroscedastic. (See text.)



further discussion of the interpretation of "twisted pear" correlations in psychological research, the reader is referred to Fisher (1959) and Storms (1960).

3. Given the Pearson correlations r_{ab}, r_{ac}, and r_{bc} among three variables a, b, c, the *partial correlations* (i.e., the correlation between each pair of variables, holding the third variable constant or "partialing out" its effect) are as follows:

$$r_{ab\cdot c} = \frac{r_{ab} - r_{ac}r_{bc}}{\sqrt{(1 - r_{ac}^2)(1 - r_{bc}^2)}},$$

$$r_{ac\cdot b} = \frac{r_{ac} - r_{ab}r_{bc}}{\sqrt{(1 - r_{ab}^2)(1 - r_{bc}^2)}},$$

$$r_{bc\cdot a} = \frac{r_{bc} - r_{ab}r_{ac}}{\sqrt{(1 - r_{ab}^2)(1 - r_{ac}^2)}}.$$

The *multiple correlation* R between the combination of any two variables and the third variable (called the dependent variable or the criterion) is of the following form:

$$R_{c\cdot ab} = \frac{r_{ac}^2 + r_{bc}^2 - 2r_{ab}r_{ac}r_{bc}}{1 - r_{ab}^2},$$

where $R_{c\cdot ab}$ is the multiple correlation between the independent variables a and b and the dependent variable c. $R_{c\cdot ab}^2$ is the proportion of the total variance in the dependent variable c accounted for jointly by the independent variables a and b.

4. One of the soundest, objective research-based discussions of the personality characteristics associated with creativity is in *The Prediction of Achievement and Creativity* (1968) by Cattell and Butcher, particularly Chapters 14 and 15.