CHAPTER 10

What Use Is Intelligence?

As of the end of the twentieth century, the United States is run by rules that are congenial to people with high IQ and that make life more difficult for everyone else.

Herrnstein & Murray, 1994, p. 541

The quotation from *The Bell Curve: Intelligence and Class Structure in American Life*, is a pretty strong statement about the importance of intelligence. When Herrnstein and Murray made it they were attacked as elitist and antidemocractic. Other people, with impeccable democratic credentials, had said similar things in a less contentious way. Just a few years before Herrnstein and Murray wrote, Robert Reich, a sociologist who had served as Secretary of Labor in the Clinton administration, wrote that work has shifted from emphasizing the manipulation of objects to the manipulation of abstract ideas, varying from programming a robot to analyzing a financial system.[1] It follows that skill in manipulating abstract concepts, intelligence, has become progressively more valuable over time. To what extent do cognitive tests predict such skill?

10.1. Problems in Investigating the Relationship between Intelligence and Success

This chapter examines the relation between intelligence and success in three broad regions; academics, the workplace, and personal life. These studies are not easy to do, for several reasons. First, we have to specify what we mean by success in each arena. Next, we have to select quantitative measures of success. Observable measures, such as grade point average or money earned, are frequently only partially satisfactory measures for our criteria. They often have undesirable statistical and measurement properties that hinder analysis and interpretation. Finally, there is the important problem of generality. We cannot study the totality of academia, the workplace, or, certainly, personal life. We have to study slices of them, where the necessary measures can be obtained. These slices are almost

1 Reich, 1991.

never random samples of the arenas, and in only a few cases can we obtain the ideal experimental group–control group contrast.

As is true in other areas of research on intelligence, we can learn something from imperfect studies. We just have to keep the imperfections in mind when we consider what has been learned. The rest of this section describes some of these imperfections. Do not lose sight of the magnificence of the forest because the trees have woodpecker holes in them! Quite a lot has been learned.

10.1.1. *The Conceptual Criterion Problem*

The biggest problem is defining success. In the academic arena a student is successful if he or she has learned. The commonest measure of academic success is a person's grade point average (GPA) across classes. However, grade point averages are not comparable across classes or institutions. A student with a GPA of 3.5 in English classes in a community college is not necessarily a better student than one with a GPA of 3.1 in Physics at Stanford. Merging gross measures of learning, such as GPA, across subjects or across schools introduces unwanted sources of variance. This will make the intelligence-GPA relation appear to be smaller than it is. But measuring the relation in one class or institution raises question about how the finding can be generalized.

An alternative measure of success is graduation or, in the K-12 system, its inverse, dropping out. Once again we have noncomparability across institutions; without naming names, not all our high schools, colleges, and universities are equivalent. Americans keep school records by district or state, not by a national register. If a student disappears from a K-12 system or fails to complete a postsecondary program, there is no record of where that student went. They may have dropped out, or they may have enrolled at another educational institution.

It is even harder to define success in the workplace. Within an industry or occupation income partially captures the idea of success, but incomes across occupations are hard to compare. Incomes are also often determined by variables unrelated to intelligence, such as seniority of employment. Some of our larger companies do keep records of periodic evaluations of employee performance, most commonly supervisors' ratings. Ratings are not reliable unless the raters are trained and the criteria for rating have been agreed upon. Objective measures of employee output are often hard to come by and generally capture only a part of a person's job. COSTCO, a giant warehouse sales company, tracks the number of check-outs per hour that each of their check-out clerks handles. It does not directly measure things like customers' reactions to a clerk's manner.

Defining success in life is even harder. We can measure extreme social adjustment, which can vary from achieving a civic prize to going to jail, but most people do neither. Success in life is a multifaceted thing. Informative studies have been conducted of the relation between intelligence and particular aspects of life success, such as health, but trying to relate intelligence, or virtually any other trait, to such a nebulous thing as "life success" is probably not a useful exercise.

Once we have defined our criteria we face the problem of actually getting the data. Several strategies have been followed. One is to conduct an experimental study, in which the investigator obtains measures of both intelligence and success from a selected set of participants. To take an example, one study related intelligence test scores to success as a race track gambler.[2] Such studies tend to be fairly small and to deal with unique situations. Because they are small, they can detect only large relationships. (Technically, they have *low statistical power.*) This brings us to a discussion of statistical issues.

10.1.2. *The Statistical Problems*

We measure the extent to which intelligence is related to some index of success by calculating *predictive validity*, which is defined as the correlation between a measure of

2 Ceci & Liker, 1986.

intelligence and the criterion measure. The process is sensitive to three statistical issues: reliability, range restriction, and generalization. In order to understand them we need a brief digression into statistical reasoning.

THE RELIABILITY ISSUE

Any measurement contains two elements, a "true value" and a residual term. While the residual term is frequently referred to as "error," it is not necessarily error in the sense of a mistake. It refers to the sum of all influences on the measured variable that are statistically independent of the true value. To take an example, consider the way in which weight is measured during the typical annual physical. Examinees are told to take off their shoes and stand on a scale. Measured weight is then shown on the scale. The measured weight has the following components:

$$\textit{Measured weight} = \textit{actual body weight} + (\textit{weight of clothes} + \textit{scale bias}),$$

where *scale bias* refers to any tendency of the scale to weigh high or low. The terms in parentheses, here *weight of clothes* and *scale bias*, are residual effects, uncorrelated with the examinee's actual weight. If an examinee were to be weighed on a different scale, wearing different clothes, measured weight might change even though actual body weight remained the same. Measured weight is said to be *reliable* to the extent that the same measure is obtained across comparable conditions. This reasoning applies to intelligence testing.

An intelligence test score x is determined by the examinee's "real" intelligence and a residual term that is unique to the examination of that person at that time. Exactly the same thing can be said of an academic grade, y. The grade is determined in part by what the student really knows about, say, English Literature and in part by a residual term unique to the examination and the person. Symbolically,

$$x = x_t + e_x; \quad y = y_t + e_y, \tag{10.1}$$

where the subscript t stands for "true" and e denotes the residual term. Now define the reliability of an intelligence test or a grade as the correlation between two measures, each assumed to be equally good, taken on the same person. Examples would be the correlation between two equivalent forms of the SAT, or the correlation between the grades assigned to the same set of English Literature examinations by two equally qualified graders.

What we can observe is the correlation between test scores and grades, r_{xy}. What we want to know is the correlation between intelligence and academic achievement, $r_{x_t y_y}$. This is

$$r_{x_t y_t} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}. \tag{10.2}$$

where r_{xx} and r_{yy} are the reliability correlations for the intelligence test, x, and the academic measure, y. The correlation $r_{x_t y_t}$ is sometimes referred to as the "true" correlation.

As reliability coefficients range between 0 and 1, the denominator, $\sqrt{r_{xx} r_{yy}}$, will also range between zero and one. Therefore, the correlation between the "true" variables, $r_{x_t y_t}$, will be at least as big, and generally larger, than the correlation between the observed variables, r_{xy}. Corrections for unreliability have to be treated with caution, as the reliabilities are themselves estimates, and if they are too low the estimated true correlation can exceed one, which is obviously not correct.

In both the academic and industrial cases the difference between the observed correlation and the (estimated) true correlation can be substantial. Professionally developed intelligence tests generally have reliabilities of .85 or above, but this is because of a great deal of careful item selection (Chapter 2). Large-scale academic achievement tests, such as those used in the United States to assess educational progress on a statewide basis,[3] have similar reliability

3 As of 2009 such tests were required by federal law – the No Child Left Behind Act.

coefficients. Within-class, teacher-assigned grades are quite another matter. I know of one case where thirty essays were graded, independently, by two university professors. The correlation between the two sets of grades was .3![4] Fortunately, this is an extreme. In most cases grades have reliabilities in the .6 to .8 range. This means that if a typical study of the relation between intelligence and grades within a class produces a value of r, that value should be multiplied by approximately 1.38 to estimate the true correlation between intelligence and academic achievement.[5] This substantial correction applies to studies of grades within a class. As the GPA is an average across classes, the reliability of the GPA is much higher than the reliability of a grade within a class, so the correction would be smaller.

Probably the most commonly used criterion for achievement in industrial settings is a supervisor's rating of performance. Unless these ratings are the result of carefully structured evaluations they are likely to have reliabilities in the .6 range or lower, considerably lower than the typical reliabilities of cognitive tests.

RESTRICTION OF RANGE

In most studies the variability of intelligence in the group actually studied will be smaller than the range in the population to which we wish to generalize. The problem of estimating intelligence-grade relations in elementary schools (K-5) illustrates the situation. Elementary schools generally draw students from the neighborhood immediately around them. In most industrially developed countries neighborhoods tend to have distinct socioeconomic and sometimes demographic characteristics. Therefore, the student demographics in a single elementary school will be more homogeneous than in the district or state. We say that the scores are subject to *range restriction*. In the case of the elementary schools, we would expect there to be less variability in a measure of intelligence within a school than across a state.

Range restriction influences the correlation coefficient. Let σ_s be the standard deviation of observed scores in the sample, and σ_p be the standard deviation in the population (in the school and in the district, in the elementary school illustration). The relation between the observed correlation in the sample, r_s, and the correlation to be estimated in the population, r_p, is

$$r_p = \frac{\frac{\sigma_p}{\sigma_s} r_s}{\sqrt{\left(\frac{\sigma_p}{\sigma_s}\right)^2 - 1) r_s + 1}}. \qquad (10.3)$$

In this equation r_p is greater than r_s if $\sigma_p > \sigma_s$. Note that this corrects only for attenuation on one of the two variables, either the intelligence variable or the criterion variable. Correction on both variables is also possible and often reasonable. For instance, suppose that a study were done in which we observed the correlation between intelligence test scores and scores on an achievement test in a school, and we wanted to estimate the correlation in the state. It would be appropriate to correct for range restriction on both the intelligence test and the achievement test.

Selection restriction is an important special case of range restriction. Selection restriction occurs whenever an applicant population takes a predictor test, here some sort of intelligence test, as part of application for a job or educational opportunity. All applicants above a given *cut score* are then accepted, and their performance on the job or in the school is recorded. For example, suppose a university uses an entrance examination, and admits the top 50% of the applicants. In order to validate the entrance examination, university officials would want to know if it was a good predictor of the grades that an applicant would obtain. However, grades are available only for the admitted students. Since, obviously, the top 50% of the applicants will have less variation in

4 The statement is based on personal observation.

5 This conclusion follows from the following argument. Assume that the reliability of the intelligence test is .88 and the reliability of the grade is .6. By application of equation 10.2, the observed correlation should be multiplied by $1/\sqrt{.528} = 1.38$.

examination scores than the entire group of applicants, the correlation between examination scores and grades in the applicant population can be estimated by computing the examination-grades correlation in the admitted group, and then correcting for range restriction.

Corrections for restriction in range can be substantial. A reasonable value for the hypothetical university example is $\sigma_s = .6\ \sigma_p$.[6] An observed correlation of .33 in the selected students would be corrected to .50 for the applicant population. As correlations are often squared and reinterpreted as representing "percentage of variance accounted for," this would change r_2 from 11% to 25%, a substantial change.

Because corrections for reliability and range restriction can be considerable, knowing when to use them is important. Here are some general rules.

1. Correction for reliability is appropriate when one's interest is in theoretical constructs underlying measures – for instance, whether intelligence as a concept is related to academic ability as a concept. The correction is not appropriate when one's interest is in whether one set of scores predicts the value of another set of scores – for instance, if you wanted to know whether the SAT predicts first-year college GPA.
2. Correction for selection restriction should be done whenever the purpose of the study is to determine the validity of a predictor, such as an entrance or hiring examination.
3. Correcting for range restriction is appropriate when the available observations are known to be a nonrandom sample in which scores are less variable than they are in the population. The case of using observations within a single school to estimate a population in the district is an example. However, in such cases correcting for range restriction will be possible only if an estimate of the population standard deviation is available.
4. If the sample is a true random sample of the population, the correction for range restriction should *not* be used.
5. Any correction for range restriction carries with it the assumption that the same (linear) relation holds between scores in the sample and in the population. This is not a trivial assumption. To take one example, there is evidence that the relation between IQ test scores and adult age is nonlinear. Scores decline more sharply with age beyond sixty than before. Therefore, it would not be appropriate to apply range restriction to estimate the age-IQ relation in adults from a sample of adults age sixty and older.

Rule 5 leads to a discussion of our last statistical issue, power.

Statistical power. To explain these issues, we need a bit of notation and a review of introductory statistics.

By tradition, scientific results are said to be "statistically significant" if they would be obtained by chance only in fewer than 1 out of 20 studies ($p < .05$) or fewer than 1 out of 100 studies ($p < .01$), on the assumption that the variables being studied in a sample are actually unrelated in the population (the "null hypothesis"). In research on intelligence, "unrelated" means that in the population there is no correlation between the predictor (an intelligence test score) and the criterion, $r_p = 0$. However, r_p cannot be observed directly. Instead it is estimated by an observed correlation, r_s, in a sample of N observations.

Assuming that the sample can be regarded as being chosen randomly from the population, there will be some critical value of the observed correlation, r^*, such that if the observed correlation, r_s, exceeds that value ($r_s > r^*$) we reject the null hypothesis that $r_p = 0$ at some level, p, where p refers to the probability of observing $r_s > r^*$ if the null hypothesis is true. The value of r^* increases if we lower the significance level (typically

6 In the case of selection restriction the percentage of applicants accepted determines the relationship between the variances in the sample and the population. In other cases of range restriction this has to be estimated.

from $p = .05$ to $p = .01$) and decreases as the size of the sample, N, increases. To take some examples, at the $p < .05$ level the critical value, r^*, is .36 for a study with 30 observations ($N = 30$), and .20 for a study with $N = 100$. At the $p < .01$ level the values are .46 and .26.

This much is taught in elementary statistics. The second point is taught but often not stressed. Suppose that the sample correlation is less than the critical value, $r_s < r^*$. This means that we cannot *reject* the null hypothesis. "Not rejecting" is not the same as accepting. What we have is what, in law, would be called a verdict of "not proven."[7]

Suppose that the population correlation is some value other than zero. (For simplicity, consider only positive values.) There would still be some probability that the sample correlation fell below the critical value – that we observe $r_s < r^*$ even though $r_p > 0$. This probability depends upon what the population value is, so the probability has to be specified given a population value and the size of the study, $Pr(r_s < r^* | r_p = k, N)$. The *power* of a study is the complement of this,

$$Power(r_p = k, N) = Pr(r_s \geq r^* | r_p = k, N). \quad (10.4)$$

In words, this is the probability that a sample of size N, drawn from a population in which the population correlation has value k, will have a sample correlation above the critical value. Going back to the earlier example, suppose that the population correlation is .50. If we set the significance level at $p < .05$, the power of a study with a sample size of 25 is .84. This means that 16 out of 100 samples will *not* reach a value reliably greater than zero even though the population correlation is a substantial .50.

Power increases with sample size. In the example just given, if the sample is increased to 100, the power is greater than .995.

The power problem becomes critical when it is combined with the problem of criterion reliability. Grades within a class and employer rating systems will often have a reliability of around .60, and intelligence tests will have a reliability of about .85. Suppose that the true correlation between intelligence and academic ability, the hypothetical variables underlying these measures, is .50 in the population. A bit of algebraic manipulation of equation 10.3 will show that the expected population correlation between test scores and grades is .26 after correction for attenuation. Setting the significance level at .05, the power of a study with 25 is participants is approximately .36.[8] About two out of three studies of this size would *not* provide strong enough evidence to reject the null hypothesis that $r_p = 0$, even though it is false. If the sample size were to be increased to 100, power would increase to about .75. In this case failure to reach statistical significance would be reasonable evidence against the hypothesis that a "true score" value of r_p was .50 or larger.

What these examples show is that power is produced by an interaction between the reliability of the measures and the size of the study. This interaction has to be taken into account in evaluating null results. Do people actually fail to do this? The answer is, stunningly, "yes." Panel 10.1 presents the case of a widely cited study in which no consideration was given to these issues.

10.1.3. *Drawing Conclusions in the Face of Statistical Uncertainties*

Given all these problems, can any conclusions at all be drawn? The answer is "yes," but only after careful consideration.

When evaluating empirical results we have to consider which statistic is appropriate. Are we interested in the observed correlation, or should the correlation be corrected for reliability and/or restriction in range? The rules given in the previous section apply.

We must be aware of power considerations. We need to be especially wary of

7 Such verdicts are not allowed in US courts, but they are allowed in some countries.

8 Power estimates are based on Table 3.3.2 in Cohen, 1988.

Panel 10.1. A Day at the Races: A Failure to Consider Power and Reliability

In 1986 two Cornell University psychologists, Stephen Ceci and J. K. Liker, published an eye-catching article entitled "A Day at the Races."* They reported a four-year study of the expertise of a group of thirty habitual bettors on harness racing. Ceci and Liker did not study the accuracy with which these bettors predicted winners because, as they said and as many horse racing fans know, the winners are often determined by unpredictable events. Instead they studied the accuracy with which the bettors were able to predict the favorite and top three favorites at post time (the start of the race), given the extensive information about each horse that was contained in the daily racing form, which is available to bettors prior to a race. Although secondary references often misinterpret this, what Ceci and Liker actually studied was their participants' ability to predict how other bettors would place their bets, not which horse would win.

Ceci and Liker found that the participants' decision was a mathematically complex function of the information contained on the racing form, and that the accuracy of the predictions had a correlation of $-.03$, essentially zero, with the participants' IQs scores on the short form of the WAIS.† Ceci and Liker drew the following strong conclusion:

> *(a) IQ is unrelated to performance at the racetrack but, more important (b) IQ is unrelated to real-world forms of cognitive complexity.*
>
> *Ceci & Liker, 1986, p. 255*

These are strong words indeed. The null finding was claimed to be reliable, and the task, something that is related to but not the same as picking the winners in a race, was unhesitatingly generalized to the universe of complex tasks. Nowhere in the article was there any mention of reliability or power.

Douglas Detterman and his colleague Kathleen Spry wrote a detailed critique of the Ceci and Liker study.‡ Among other things, they observed that Ceci and Liker's criterion, the ability to predict the odds at post time, had a reliability of at most .41. What does this mean? Suppose that the correlation between the underlying abilities, intelligence and skill at setting the odds, is 1. In terms of the text, $r_p = 1$. The reliability of the short form of the WAIS is known to be .85. Therefore, the expected value of the correlation in the sample would be $.85 \times .41 = .35$. If $N = 30$, the power of the Ceci and Liker study would be approximately .5, which means that a study like theirs should fail to reach the convention .05 level of statistical significance five out of ten times even if the underlying correlation was one.

Of course, nobody thinks that the correlation between intelligence and race track betting is identically one. Based on meta-analysis, a widely quoted estimate of the correlation between intelligence and performance on a cognitively complex task is $r_p = .5$.§ To be generous, increase this to .6. Then, solely on the basis of reliability, the expected sample correlation would be $r_s = .21$. Using this estimation the power of the Ceci and Liker study was .20; studies like theirs should fail to reach the .05 level of significance four out of five times.

The Ceci and Liker study presents us with good news and bad news. The good news is that when a published study contains major flaws, other scientists point out the errors. The bad news is that almost no one notices the correction. The Ceci and Liker study has been cited ninety times, as evidence that intelligence, as measured by the IQ tests, is not related to real-world cognition. The Detterman and Spry study has been cited seven times.**

* Ceci & Liker, 1986.

† This figure is a correction to the original value, provided in Ceci & Liker, 1987.

‡ Detterman & Spry, 1988. This much-neglected article contains several other strong criticisms of the Ceci and Liker work.

§ Schmidt & Hunter, 1998.

** Data collected from an ISI Web of Knowledge citation search, July 2, 2009.

concluding that there is no relation between intelligence and some other variable when the study involved is small or uses a measure of unproven reliability.

One way to address the power issue is to do studies with a very large number of participants. The extreme case is to utilize large surveys, such as the Department of Labor longitudinal studies of US citizens described in panel 9.9. Sometimes surveys are "constructed" by analyzing records of intelligence, health, educational accomplishment, and occupational status that have been collected for other purposes.

Obviously, though, the larger the study and the more time required of the participants, the more costly the experiment or survey. In many cases the investigator has to accept less-than-ideal measures, such as using a brief vocabulary test as a proxy for a measurement of intelligence, or using place of residence as the sole measure of a participant's socioeconomic status. Such compromises are not fatal errors; they are things that have to be considered when evaluating results.

Another way to address the power problem is to use a statistical technique called *meta-analysis* to draw conclusions from multiple studies.[9] Special statistical methods are used to identify trends that may not be clear from focusing on the details of any one study. This technique can be quite revealing. However, there are reservations.

As is the case for any statistical technique, generalizations based upon any unjustifiable assumptions of random sampling are suspect. For example, many studies have been conducted of the relation between scores on college entrance examinations and student grades. Meta-analysis can, and has, been applied to these studies. The participating institutions tend to be the larger institutions, with budgets sufficient to support internal research organizations. Therefore, the result of a meta-analysis of such studies is a useful descriptive statement, but any appeal based upon a claim of random selection of institutions is questionable.

The individual studies reviewed in a meta-analysis will, inevitably, vary in the quality of the measurements taken and the procedures used. These considerations are judgments that have to be made by considering the details of each study. They do not lend themselves to statistical treatment. All reviewers will agree, for instance, on the number of participants in a study, and the effect this has on statistical power. They may not agree on the appropriateness of the measures used, or the way in which the measures were taken. Some meta-analyses have attempted to deal with this problem by classifying studies by their perceived quality, and then analyzing high- and low-quality studies separately, to see if this makes any difference. A finding that only appeared in low quality studies would certainly be treated with suspicion.

10.1.4. *Problems Related to Research Design*

The final problems to be considered have to do with research design, rather than statistics and measurement.

The ideal research design is a *prospective* study, in which the investigator obtains data on the intelligence of people at some point in their lives, ideally before they enter an academic program or the workforce, and then determines how well they succeed. This is by far the easiest kind of study to interpret. However, it is possible only if the investigator has some way of testing a large number of people, and then following them for a reasonably long period. There are a few studies that have done this. The Seattle Longitudinal Study[10] (panel 9.2) and the National Longitudinal Studies (panel 9.9) are important examples. However, such studies are expensive, and so are few and far between.

Prospective studies can sometimes be conducted by examination of government records. Studies of this sort have been carried out in those European countries in which eighteen- to twenty-year-old men have to register and be tested for potential

9 Hunter & Schmidt, 1990.

10 Schaie, 2005.

military enlistment. (As far as I know, Israel is the only country that requires registration for both men and women.) Valuable information can be gained if some of the registrants can be reinterviewed later in life, to determine how well they have fared. In some countries this can be done without actually interviewing the individuals, because the government keeps extensive records of the health, education, and income of its citizens. Legal and ethical issues concerning access to such data have to be resolved, but the important point is that the studies often can be done.

The alternative to a prospective study of the relation between intelligence and success is a *retrospective* study. In a retrospective study a group of people are identified who have, or do not have, varying degrees of social success. The investigator then attempts to determine their intelligence, either by direct testing or by examination of relevant records. Studies of eminence or genius often fall into this category. The investigator identifies a group of individuals who meet some criterion for accomplishment and then tries to identify the common characteristics of the group. Possibly one of the most ambitious of these studies was Simonton's determination of the correlation between a measure of intellectual capacity, reconstructed from historical records, and historians' ratings of the performance of the forty-two US presidents, from George Washington through George Bush.[11] The correlation was .56.

Studies of the relation between intelligence and success are prone to collinearity problems. To illustrate, intelligence test scores during adolescence are positively correlated with subsequent income.[12] Does this mean that high intelligence causes a rise in income? Perhaps. But it is also true that children's test scores are positively correlated with parental socioeconomic status, although the correlation (~.40) is not as high as many people assume. Is current income due to intelligence, or is it a legacy of the privilege of having come from a wealthy (or poor) background, with concomitant opportunities (or lack of opportunities) to get a foothold on the economic ladder? Or both?

10.2. The Relation between Intelligence and Academic Achievement

Binet's motivation for constructing the original intelligence test was to identify children who were at risk for failing in the standard academic system. Subsequent test developers generalized the goal to predicting degrees of success at all levels of education. How well has this worked?

10.2.1. *Intelligence in the K-12 System*

In 1972 the American clinical psychologist Joseph Matarazzo reviewed the evidence, and concluded that IQ, as measured by the Wechsler tests, was a good predictor of high school graduation.[13] In 1994 Herrnstein and Murray addressed the same question using the AFQT and the NLSY79 data. Figure 10.1 shows their results for graduation rates in the 1980–85 period. The finding is clearly robust. Different tests were used at different times, and with different definitions of "failure to graduate." Nevertheless, both studies found the same positive relation between test scores and probability of graduation.

Matarazzo also said that, based on "thousands" of studies, it had been shown that intelligence test scores correlate .50 with grades in the K-12 system.[14] This estimate has been widely accepted by subsequent reviewers.[15] Later reviewers do qualify the estimate, by saying that the correlations tend to be higher in elementary than in middle school, and drop to perhaps .40 in high school. As psychological studies are notorious for failures to replicate findings, the agreement among reviewers, over a considerable time span, is reassuring.

11 Simonton, 2006.
12 Herrnstein and Murray, 1994.
13 Matarazzo, 1972.
14 Matarazzo, 1972, p. 283.
15 Brody, 1992, pp. 251–254; Jensen, 1998; Macintosh, 1998, Chapter 2; Neisser et al., 1996.

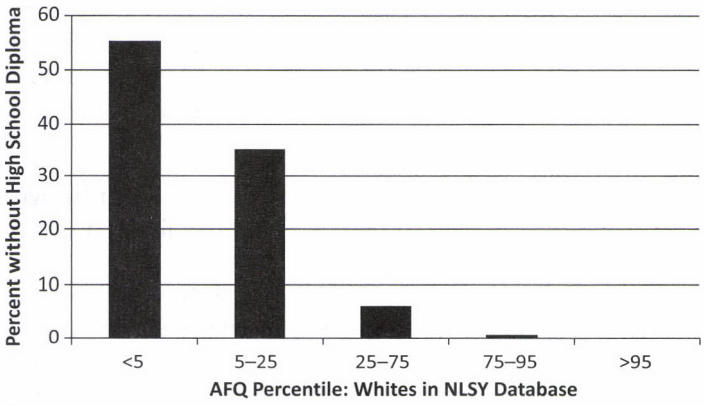

Figure 10.1. Percentage of White young adults in the NLSY79 survey who did not complete high school, plotted as a function of their percentile scores on the AFQT. Data from Herrnstein & Murray, 1994.

The .50 figure applies to measures of grades computed across classes, elementary, middle, or high school GPA. If the correlation is calculated for scores within a single class, it will drop due to range restriction and, often, due to the lowered reliability of locally produced tests. A study by researchers at the University of Pennsylvania found a correlation of just slightly over .30 between Otis-Lennon test scores (see Chapter 2) and academic achievement on tests at the end of the eighth grade.[16] This study was done in a "magnet" high school where the students had already been selected on the basis of test scores and previous grades, so range restriction was certainly a factor. Other studies confined to a single school or class have found correlations of about .5 between test scores and grade point averages in the early primary grades and in high school.[17]

Macintosh[18] has observed that although restriction of range is frequently appealed to as a mechanism that should reduce test-achievement correlations, the effect has never actually been observed, at least in studies of the K-12 system. Two large European studies come close to addressing Macintosh's concern.

The English system of education is much more centralized than the American. One year something over 78,000 students were given the Cognitive Abilities Test (CAT, described in Chapter 2). The test takers included almost all the eleven-year-olds in England, so range restriction is not relevant. At age sixteen all English students take nationwide examinations in a variety of subjects. The national examinations are subject to much more careful psychometric evaluation than is typically the case for locally generated (and certainly for teacher-generated) examinations, so reliability of the criterion variable was not a major concern.

Ian Deary and his colleagues[19] extracted a general intelligence (*g*) factor from the CAT scores and a general academic achievement factor, which I will call *a*, from the scholastic examination scores. The correlation between the two was $r_{ga} = .81$. This is a very high value. There was substantial variation between associations with the *g* factor and educational accomplishment within topics. Correlations ranged from a high of .77 for mathematics to .43 for Art and Design. In general, the topics usually considered the academic core courses – the Humanities, Mathematics, and the Sciences – had correlations in the .50–.75 range, while "practical" topics, such as Art and

16 Duckworth & Seligman, 2005.
17 Kaplan, 1996; Zwick & Green, 2007.
18 Macintosh, 1998.
19 Deary, Strand, et al., 2007.

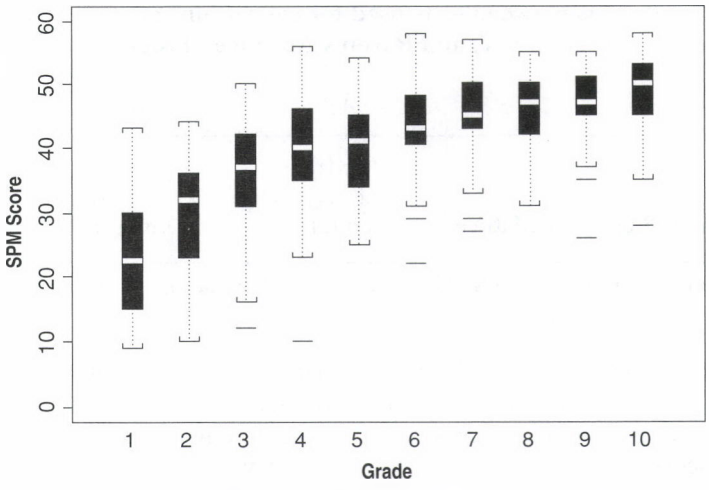

Figure 10.2. The distribution of RSPM scores across grades in a representative sample of 500 Icelandic schoolchildren. From Pind, Gunnarsdóttir, & Jóhannesson, 2003, Figure 1, with permission from Elsevier.

Design, Music, and Textiles, had correlations in the .4–.5 range. This very large study provides strong evidence for a robust relationship between intelligence, as assessed by a standard test, and general academic achievement.

Correlations between battery-type tests and academic achievement can be criticized on the grounds that some of the subtests in a test battery are close to a sample of academic tasks, so what we are measuring is the stability of academic aptitude, rather than a more general factor of intelligence. This interpretation cannot explain the fact that results somewhat similar to those in the English study have been obtained in Iceland, using Raven's Standard Progressive Matrices (RSPM), which is certainly not tied to the K-12 curriculum.[20] As was the case for the English study, this study used a very large sample, representative of the population of schoolchildren in Iceland. Figure 10.2 shows the results. Scores rise over the school-age years, agreeing with our intuition that children increase in their cognitive competence as they grow older. The RSPM scores within class levels were correlated with grades. At the seventh-grade level the correlations were .75 for overall mathematics grades and .64 for overall Icelandic (language arts) grades.

The British and Icelandic studies agree with each other well, even though the particular tests of intelligence used were quite different. They answer Macintosh's legitimate concern. Range restriction effects do operate in small, localized studies, and therefore corrections for range restriction are appropriate. The correlation between intelligence test scores and academic achievement in the K-12 system, calculated across large units, such as districts, is at least .50 for the core academic courses, such as language arts, science, and mathematics, and somewhat lower for courses in vocational topics and in the arts. Within smaller units, such as a school or a class, the correlation will drop to around .30, due to restriction of range.

In the K-12 system cognitive test scores are used to identify students whose low scores indicate that they may need to be assigned to special education classes. Tests are also used to assign students to accelerated programs for the gifted. In both cases other factors are also considered. The majority of students fall somewhere between these extremes, and for them the test scores do not matter, for no

20 Pind, Gunnarsdottir, & Johanesson, 2003.

Table 10.1. Correlations between tests used for college/university selection, the ASVAB general factor or AFQT, and Raven's Advanced Progressive Matrices

Test	*SAT*	*ACT*	*ASVAB*
ACT	.87 (1)		
ASVAB	.87 (3), .92 (1)	.77 (4), .90 (1)	
Raven's Advanced Progressive Matrices	.71 (2)	.61 (3)	***

Source: Data sources are (1) Coyle & Pillow, 2008 (NLSY97 data); (2) Frey & Detterman, 2004; (3) Koenig, Frey, & Detterman, 2008.

decisions are made on the basis of these scores. In college and university entrance decisions test scores matter, a lot, across the entire range of scores.

10.2.2. *Intelligence in the Post-Secondary System*

Since World War II American colleges and universities have incorporated two major testing programs, the SAT and the American College Testing Program (ACT), into the admissions process. These tests are validated regularly, by correlating test scores with first-year grade point average (GPA1), cumulative grade point average (GPAC), or probability of graduation within a specified period of time after matriculation, usually four to six years. The use of the tests is not without controversies, a point that was made earlier (section 2.7.3). We concentrate on technical rather than policy issues here.

Recall, from the discussion in Chapter 2, that the first portion of the current SAT, referred to officially as the SAT-I, contains sections stressing verbal comprehension and logical reasoning. By tradition (and officially, in earlier versions), these two sections have been referred to as the SAT-V and SAT-M. I will continue this usage. Both tests represent attempts to evaluate comprehension and reasoning without tying questions to specific high school curricula.

The American College Testing program takes a different approach. It develops tests that are specifically tied to curricular material, such as history and mathematics. The idea is to predict what a student will learn in college by determining how much he or she has learned in high school. The second part of the SAT program, the SAT-II, does the same thing.

Although there have been arguments about which approach is better,[21] the tests could be interchanged in an academic selection program without changing acceptance and rejection decisions very much. Table 10.1 presents estimates of the correlations between the SAT, ACT (summary score), and the general factor derived from the ASVAB, which is closely approximated by the Armed Forces Qualifying Test (AFQT). The correlations with Raven's Advanced Progressive Matrices are also included, in order to show the relation between the educational tests and an avowed marker for general intelligence, *g*.

The correlations are quite high. The correlation between the SAT and the ACT approaches the reliabilities of the two tests. This suggests that the true correlation between the two tests is one! A study using NLSY97 data found that both academic aptitude tests had loadings of about .9 on a general factor derived from the ASVAB.[22] The finding is important because the ASVAB general factor is a measure of crystallized intelligence (Gc), rather than of Gf.[23]

The need to distinguish between Gf and Gc in college students is shown by the fact that the correlations between matrix tests and the SAT and the ACT are in the .6–.7 range.[24] This is about what one would

21 Lemann (1999) discusses the dispute in some detail. It has been carried forward to this day.

22 Coyle & Pillow, 2008.

23 Roberts et al., 2000.

24 Frey & Detterman, 2004; Koenig, Fry, & Detterman, 2008.

expect, because the fact that Gc and Gf are themselves correlated in the .5–.7 range, depending upon the sample. Because aspiring and attending college students represent roughly the upper two-thirds of the general population, in terms of cognitive skills, one expects the general factor to be somewhat weaker among this group than among the population at large. (See Chapter 3 for elaboration.)

Do the tests work? Several appropriately designed large studies of the SAT have produced consistent results. The correlation between SAT scores and GPA1 is approximately .35 in students who have been admitted, and who therefore have both SAT and GPA1's available.[25] This is the uncorrected correlation in the selected population, whereas what is needed is predictive correlation in the applicant population. An extensive study by Paul Sackett and his colleagues at the University of Minnesota, in which they conducted a meta-analysis of previous studies, shows quite clearly what the situation is.

Sackett and his colleagues analyzed data provided by the College Board for 41 colleges and universities where the SAT was used in 1995–97. Over 155,000 test takers were involved. The researchers calculated three SAT-grade correlations. They were:

1. r_s – the correlation between SAT and GPA1 in admitted students, calculated within institutions and then averaged. $r_s = .35$.
2. $r_{p1} - r_s$ corrected for restriction of range within the applicant population for each institution, and then averaged. This is the predictive correlation that would be of interest to admission officers. $r_{p1} = .47$.
3. $r_{p2} - r_s$ corrected for restriction of range of SAT scores across all institutions. This can be thought of as the predictive correlation to be used to determine the benefit of using the test across all participating institutions. $r_{p2} = .53$.

25 Geiser & Studley, 2002; Kobrin et al., 2008; Sackett et al., 2009.

Freshman grade point averages indicate a student's initial reaction to college. What about predicting later performance or graduation? Beyond the first year there is great variation in the courses college students take, and there are also substantial differences in grading practices across disciplines. This muddies the situation.

There is a negative correlation between the SATs of students within an academic program and the mean grade point assigned by that program. This is because mathematics and science programs, which assign relatively low grades, tend to draw the students with the highest SATs, while humanities and education programs, which assign high grades, draw students with lower SATs. The effect is quite large. A study involving over 200,000 students from 38 public universities during the 1990s[26] found that the difference in SAT scores between the discipline with the highest entering scores, engineering, and the one with the lowest scores, education, was .92 standard deviation units.[27] The negative correlation between the rigorousness of grading within a discipline and the SATs of the entering students will reduce the correlation between overall GPAs and entering SATs, calculated over the institution as a whole.

The probability of graduation behaves much like, but not exactly like, GPA1. Herrnstein and Murray's analysis of the NLSY79 database showed that, as of the 1980s, approximately 70% of the survey participants in the top decile of AFQT scores obtained bachelor's degrees. This fell to 30% in the eighth decile, and to 10% in the fifth decile.[28] A detailed report from the College Board[29] found that graduation

26 Kroc et al., 1997.

27 This is a conservative estimate, based on the assumption that the within-discipline standard deviation is equal to the population standard deviations. The assumption is very unlikely to be valid. The effect of interdisciplinary variation would be to reduce the variance (and hence the standard deviation) within disciplines. The upshot would be that *less than* 18% of the entering education students would be expected to have SAT scores above the engineering mean.

28 Herrnstein & Murray, 1994, p. 37.

29 Kroc et al., 1997.

rates are nonlinearly (logistic) related to an index composed of SAT, High School Grade Point Average (HSGPA), and several demographic variables, including gender and race. People with relatively low scores on the index generally were unlikely to graduate; people with high scores were highly likely to graduate; and the probability of graduation changed markedly between "low average" and "high average" scores.

As was the case for the K-12 system, the findings on the correlation between test scores and college/university success are strikingly consistent. The SAT, the most widely used test, has a predictive validity of about .5. This is probably an underestimate of the correlation between the SAT and an abstract measure of academic ability. Because students with high SAT scores are more likely to enroll in "tough-grading" courses than students with low SATs the SAT-GPA1 correlations will be depressed below what they would have been if all students took the same courses.

Are these correlations really enough to justify test use? Answering that question requires a brief discussion of the statistics of personnel selection.

10.2.3. *Cognitive Tests and Selection Decisions*

The college/university admissions process is an example of a personnel selection decision. How useful are entrance examinations, such as the SAT, in making such decisions? This raises the question of how high a correlation has to be in order to be useful in practice, whether or not it is "statistically significant." This depends upon how the correlation is to be used.

A widely cited way of evaluating the size of a correlation is to square it, and then to report it as the proportion of variance accounted for in either variable by predictions using the other variable. In the admissions case, r^2_{p1} would be the proportion of variance in grades that could be associated with variance in an admissions test – approximately $.5^2 = .25$. Multiplied by 100, one could say that 25% of the variance in grades is accounted for by variance in the examination. If, as is often (and erroneously) done, this logic is applied to the uncorrected correlation between grades and test scores, in the population of admitted students, $.35^2 = .11$, so 11% of the variance of grades is associated with variance in test scores – which does not seem high. However, that is misleading.

If the selector uses a screening examination, it is possible to predict aptitude (grades or workplace performance) and accept people in order of predicted performance. Unless prediction is perfect (predictive validity = 1) people with the same predicted performance usually turn out to have different actual performances. Students with identical SATs do not all have identical grades. In statistical terms, there is variance around the predicted performance level, and the greater the variance, the less accurate the prediction. However, variance around the predicted performance can never be greater than the variance in the applicant population. So variance in the *applicant* population can be used to scale the extent to which the prediction is *not* accurate. The ratio $I =$ (variance around predicted value of aptitude)/(variance of aptitude in applicant population) represents an "inaccuracy" index, relative to the inaccuracy that would be achieved without using a selection examination. It follows that the complement of I, $1 - I$, is an index of accuracy. It can be interpreted as the relative reduction in inaccuracy achieved by using a predictor test. The I index is related to predictive validity, as defined in section 10.1, by the equation

$$r^2_p = 1 - I \qquad (10.5)$$

where $p = 1$ or 2, depending upon whether you are interested in within-institution or across-institution predictivity. Multiplied by 100, r^2_p is the percent increase in efficiency achieved by using a screening examination. If, as is the case, $r_p = .5$, the increase in efficiency is 25%.

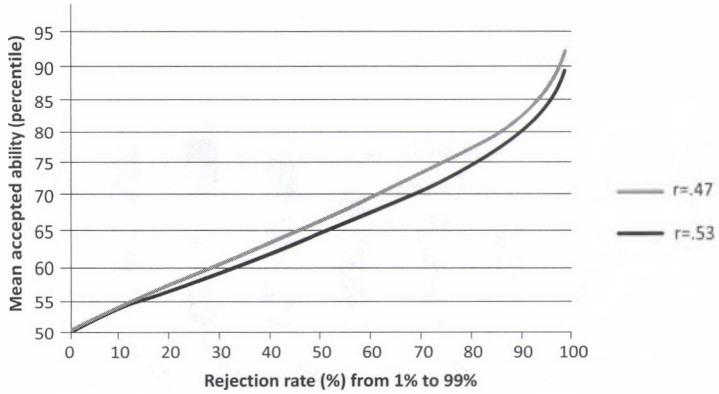

Figure 10.3. Test accuracy and rejection rate interact to produce quality acceptances. The expected value of aptitude for an accepted candidate (student or worker), measured in terms of the percentile of aptitude in the applicant population and shown as a function of the rejection rate and the predictive validity of a screening examination.

At this point we can see an argument brewing between the admissions committee and the rejected applicants. Suppose an applicant is rejected, and then learns that among accepted applicants (students) the correlation between SAT and grades is only .35. How dare the committee reject an applicant on the basis of a test that is only 11% better than chance?

The committee's first reply can be that the correlation is not really .35; it is .5. The applicant's rejoinder is that a 25% improvement over chance still is not good enough. But this is not the admissions committee's real argument.

The admissions committee is not interested in the accuracy of individual predictions; it is interested in selecting the best possible entering class. Suppose that the institution has room for only 10% of its applicants (a rejection rate of 90%). Insofar as is possible, the committee wants to select the top 10% of the applicants in terms of academic aptitude. However, the committee knows only the top 10% of the test scorers. If $r_p = 1$, the two "top ten percents" will be the same people; to the extent that r_p is less than 1, there will be some disagreement.

The success of the selection process will be determined by both the accuracy of the test, r_p, and the rejection rate. If the rejection rate is zero, everyone who wants to enter gets to enter. The accuracy of the test does not matter, because no decision is going to be made using the test score. At the other extreme, suppose there is just room for one person. The person accepted will be the one with the highest test score, and the probability of that person being the person with the highest aptitude in the applicant pool will depend upon the accuracy of the test.

Between these two extremes, the expected quality of accepted applicants is determined by an interaction between r_p and the rejection rate. The form of this interaction is shown in Figure 10.3, using Sackett's two estimates of the predictive correlation as examples. If the rejection rate is low, the value of the predictive correlation matters very little. If the rejection rate is high, it matters a lot. For example, if the rejection rate is 90%, as it is for some of our elite universities, the use of an entrance examination with a predictive correlation of .47 can improve the mean level of aptitude in the entering class from the fiftieth percentile of the aptitude in the applicant population (no test used) to about the seventy-seventh percentile.

Exactly the same reasoning applies to industrial hiring. If the rejection rate is high, a screening examination with predictive validity in the .4–.5 range can substantially

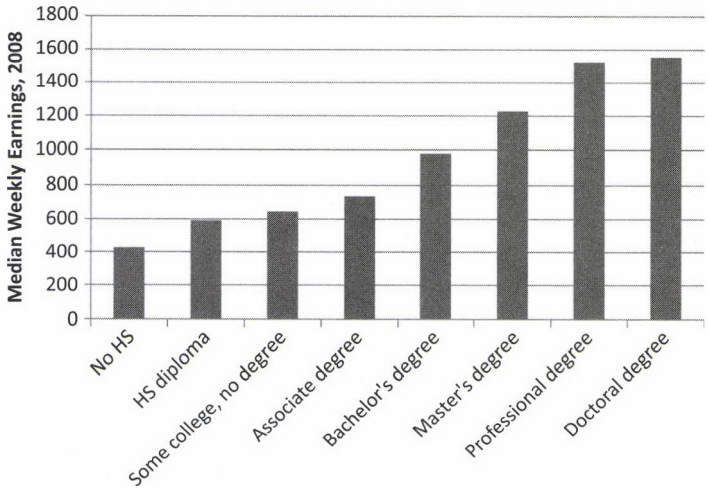

Figure 10.4. Median weekly earnings in 2008 as a function of level of education. *Source:* US Bureau of Labor Statistics. HS – high school diploma.

improve the selection process, as seen by the employer.

Note that quality has been defined in terms of the quality available in the applicant population. Personnel selection has to operate with this constraint; you cannot select people who do not apply. Any recruitment technique that improves or diminishes the distribution of aptitudes in the applicant population will affect the quality of the selected applicants – either the student body or the workforce. What this effect will be will depend upon the amount of change in applicant aptitude and upon the effect of added or reduced recruitment upon the rejection rate.

10.2.4. *Alternatives and Augmentations to the Use of Test Scores in College Entrance Decisions*

Chapter 9, section 5, contained a century-old quotation from Theodore Roosevelt about the economic value of education. Figure 10.4 shows income figures, as of 2008, as a function of level of education completed. Roosevelt's remark rings true. How we decide who gets to go to college makes a tremendous difference in who has economic and social opportunity. Therefore, it is understandable that college entrance examinations such as the SAT and the ACT have received considerable scrutiny.

SAT scores are positively correlated with parental SES. This has led some to fear that using the SAT simply identifies applicants who have the social and financial resources to complete the undergraduate program. Giving these applicants preference in college admission will therefore exacerbate inherited social advantages, something that is generally not considered a good thing in a democracy. (Paradoxically, the SAT was originally designed to reduce these advantages! See the discussion of the SAT in Chapter 2.)

To what extent is this concern warranted? The way to investigate the question is to examine the partial correlation between grades and test scores, equating for SES. If the SAT is a proxy for SES, the partial correlation should approach zero. However, it does not. An analysis of the forty-one-institution data collected by the College Board found that the partial correlation, based on an analysis of over 155,000 students, is .44 – very little different from the predictive correlation without considering SES, .47. The analysis can be reversed, to see if SES is associated with first-year grades, after equating for test scores. When this is done the correlation between SES and

GPA1 drops from .31 to .05. Similar results were found in the meta-analysis of previous studies.[30]

These values are consistent with the assumption that parental SES does influence undergraduate performance, but that it does so as a distal variable. SES exerts its influence by influencing traits that are important for success as an undergraduate and that are measured by the SAT. Presumably these traits are what we mean by intelligence.

High school grade point average (HSGPA) has also been used as a predictor of GPA1 and college graduation. A study conducted in the University of California system found that, averaged over the incoming freshman classes of 1996–99, the SAT-I could predict 13% of the variance in GPA1 ($r = .36$), HSGPA could predict 15.4% ($r = .39$), and the two of them together could predict 20.8% of the variance ($R = .46$).[31] These correlations, which are consistent with the data from the forty-one-institution study, have not been corrected for range restriction. As far as accuracy of prediction is concerned, the appropriate thing to do is to combine the entrance examination and HSGPA into a single index.

As is true of all cognitive tests, the SAT has been designed to measure "can do" aspects of cognition. Cumulative indices of performance, such as HSGPA and GPA1, also tap "will do" aspects of performance, such as study habits and perseverance. We have seen this already for HSGPA; the same thing is true for GPA1.[32] The fact that HSGPA and the SAT, combined, do better than either alone is further support for an expanded definition of intelligence, to include skill in allocation of effort over the long haul, outside of the conventional testing paradigm.

We have been concerned here solely with the relation between intelligence and cognitive performance after matriculation. We need to remember that this is a limited view of only one aspect of the admissions decision. Admissions policy makers also consider other ways in which prospective students may contribute to the university. These vary from maintaining family ties to an institution (the "legacies" that produce endowments on which some universities rely) to a student's athletic abilities. Policy makers are also influenced by a desire to balance male-female ratios and to promote racial and ethnic diversity in the student body. Considering the appropriate role of these goals in student admissions would take us far beyond a discussion of the value of intelligence in education.

10.2.5. *Post-Graduate Education*

In 1837 a twenty-three-year-old named Edward Cree received a license to practice as an apothecary from the University of Edinburgh. Having a desire to go to sea, he was appointed a Surgeon in the Royal Navy. Ten years later he took some time off from the Navy to complete his M.D.[33] The same sort of thing happened in the United States. The *Greenfield Village* "living museum," part of the Henry Ford Museum, contains an account of a mid-nineteenth-century physician who "studied medicine awhile" at the University of Michigan and Case-Western Reserve University, decided he had learned all about medicine that he needed to know, and set up his practice on the northwestern frontier.

Today we are a bit more formal. We demand completion of programs for entry into many professions. No credit is given for attendance. The rewards for completion can be considerable. Just as there is a 50% increase in income for going from the High School degree to the Bachelor's, there is another 50% by going from the Bachelor's to the Doctorate (Figure 10.4).

A variety of cognitive tests are used as screening examinations for post-graduate educational programs. It is difficult to say anything comprehensive about their validity, for the importance of grades in graduate education varies tremendously with the

30 Sackett et al., 2009.
31 Geiser & Studley, 2002.
32 Credé & Kuncel, 2008.
33 Cree, 1982.

Table 10.2. Odds ratios comparing probability of graduation for the top and bottom halves of admission test scores in an entering population of post-graduate students

Test	*Typical Manner of Use*	*Odds Ratio*
Graduate Record Examination	Entrance into Ph.D. programs in many fields of study	2.3:1
Miller Analogies Test	Entrance into Ph.D. programs in many fields of study	2.2:1
Law School Admissions Test	Entrance into Law School	1.4:1
Graduate Management Admissions Test	Entrance into MBA programs in business and management	1.6:1
Medical College Admissions Test	Entrance into Medical School	1.7:1

Source: Data excerpted from the supporting online material for Kuncel & Hezlett, 2007, Table S1. Reprinted with permission from AAAS.

program. It is my impression (but no more than that) that grades are taken fairly seriously in professional programs such as Law, Business, and Medicine, and are regarded as incidental to research participation in most science programs. A validity measure that does not require equating grades across programs is the accuracy with which high scores predict program completion. This is measured by the *odds ratio* for program completion, which is defined as

$$\text{Odds Ratio} = \frac{\text{Completion rate for students whose entrance scores are in the top half}}{\text{Completion rate for students whose entrance scores are in the bottom half}}.$$

Table 10.2 shows the odds ratios for a variety of entrance examinations used as part of the screening examinations for entry into various graduate schools. The odds ratios vary from a low of 1.4 (for the Law School examination) to a high of 2.3 (Graduate Record Examination). Having a test score in the top half of applicants is associated with at least a 40% improvement in probability of graduation, compared to test scorers in the bottom half.

Although having a post-graduate degree clearly pays off, completing a post-graduate course often entails considerable financial and personal sacrifice in the short term. The information in Table 10.2 is of as much use to an accepted applicant, trying to decide whether to enter graduate school, as it is to an admissions officer.

10.3. The Workplace

Do tests of intelligence predict performance in the workplace? Here is a claim by three industrial-organizational psychologists.

> *Many laypeople, as well as social scientists, subscribe to the belief that the abilities required for success in the real world differ substantially from what is needed to achieve success in the classroom. Yet, this belief is not empirically or theoretically supported. A century of scientific research has shown that general cognitive ability, or g, predicts a broad spectrum of important life outcomes, behaviors, and performances.*
>
> *Kuncel, Hezlett & Ones, 2004, p. 148*

Putting Kuncel and colleagues' proposition more argumentatively, this is a case where the public (and many social scientists) have made up their mind, so please do not confuse them with facts.

Linda Gottfredson is ready to plunge ahead, whether she is believed or not:

> *In no realm of life g is all that matters, but neither does it seem irrelevant in any. In the vast toolkit of human abilities, none has been found as broadly useful – as general – as g.*
>
> *Gottfredson, 2002, p. 332*

Gottfredson is right, but Kuncel and colleagues are right to be concerned that the facts are a hard sell. Some of the reasons why are captured in a third quote, this time by J. Raven, the son of the J. C. Raven who developed progressive matrix testing and himself a prolific researcher on intelligence. Given his pedigree, one might expect J. Raven to take Gottfredson's position, but he is rather hesitant.

> *In the workplace and in the educational system numerous other qualities are important but remain invisible if one utilizes only tools developed within the traditional measurement paradigm, focuses mainly on conventional criteria of job performance, and accepts assumptions about the functionality of hierarchical organization of workplaces and society.*
>
> *J. Raven, 2008c, p. 432*

J. Raven further argues that the important things determining job performance are not general cognitive power, but rather the specific skills and the motivation that a person brings to work. He also points out that evaluations of both job performance and academic success take place in constrained situations. The constraints of the situation may be just as important as cognitive capabilities in determining behavior. Constraints on job performance vary widely across the workplace, while academic constraints are more uniform. This argument is worth developing.

In the academic setting there is a reasonably clear-cut criterion for success – how well does a student know the material stated in the curriculum? When implicit objectives like "teaching a student how to think" are introduced, agreement over the criteria for success vanishes.

10.3.1. *Some Evidence from Studies of Military Enlisted Performance*

In the 1980s the United States Department of Defense conducted extensive studies of the prediction and assessment of the job performance of enlisted personnel.[34] The predictive measurements taken included cognitive and personality tests and biographical statements of interests. Occupational assessments were similarly varied. They included examination of service record books (which contain job performance ratings) and records of promotions, commendations, and disciplinary actions. In addition, both pencil-and-paper and hands-on performance tests were given. Examinees had to demonstrate their general skills and knowledge as soldiers, sailors, marines, or airmen and their proficiency in their specific occupations. The occupations chosen varied from strictly military positions, such as infantrymen and artillerymen, to jobs with exact counterparts in the civilian world, such as automobile mechanics, clerks, and cooks.

Five dimensions of job performance were identified. Two, *general military proficiency* and *technical proficiency* in one's specialty, were "can do" measures. They evaluated how well a person could do his or her job, when they knew that they were being evaluated. The next three factors were "will do" measures. *Discipline* referred to whether or not the individual followed regulations and could be relied upon to be ready to do his or her job. *Leadership* referred to the ability to encourage others and to take initiative. *Fitness* referred to personal bearing, appearance, and physical fitness. With the possible exception of fitness these dimensions apply to both the military and civilian workplaces.

Figure 10.5 shows the relation between the five factors and measures of personality, biographical interests, and cognitive performance (including scores derived from the ASVAB). The cognitive measures were the best predictors, by far, of the two "can do"

34 Campbell & Knapp, 2001.

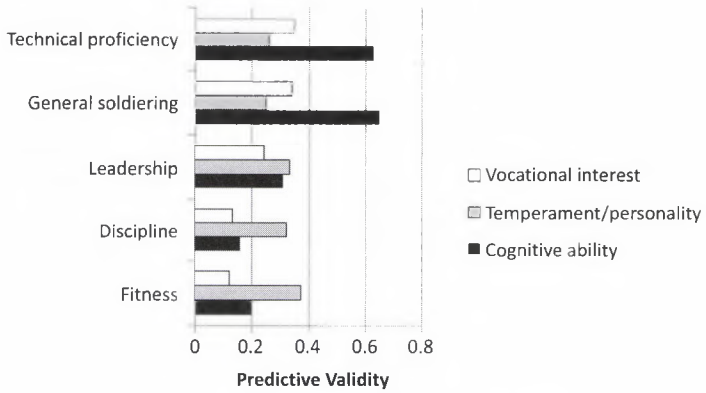

Figure 10.5. Correlations between predictors and criterion measures in the U.S. Army study of enlisted performance. Data from McHenry et al., 1990, Table 4.

factors. Interest and personality measures were the best predictors of the "will do" aspects of job performance.

Steven Hunt, an industrial and organizational psychologist, has pointed out that the first two steps in developing an assessment program in industry are to define the job that you expect employees to do and to determine how you are going to decide whether their performance measures up to these expectations.[35] It is not reasonable to expect anyone to excel in all aspects of performance. To the extent that the required job skills are themselves not correlated, it is impossible for one predictor to predict them all. The results shown in Figure 10.5 illustrate Steven Hunt's point. "Can do" is useless without "will do."

Job performance is highly dependent upon experience, because the more one practices something, the better one becomes at it. Expertise in complex tasks can take years to acquire.[36] Expertise implies the ability to learn from experience.

Further military studies showed that job performance was a joint function of experience on the job and intelligence. Soldiers at all intelligence levels took about eighteen months to approach their top levels of performance, with much slower improvement in the next two years. Soldiers with the highest cognitive scores (AFQT Level I and II) performed better after six months on the job than soldiers with lower scores after forty-two months.[37]

The military provides a highly structured workplace, and the workforce is younger than the civilian workforce. What are the relationships between intelligence and performance in the civilian workplace?

10.3.2. *Evidence from the Civilian Workplace*

Literally hundreds of studies have been done of the relation between test scores and job performance in the civilian sector, using tests ranging from the ASVAB, which takes several hours, to the Wonderlic and the Raven tests. Two American industrial-organizational psychologists, John Hunter and Frank Schmidt, have conducted a number of widely cited meta-analyses of these results. Figure 10.6, taken from one of their best-known studies,[38] shows that in the blue-collar, clerical, and administrative occupations the predictive validity of general intelligence, averaged over all studies, is .51 (corrected for range restriction and unreliability in the job performance criterion). The validity coefficient can be increased by combining a measure of general mental ability with various other assessment methods. Predictive ability can be raised to its

35 S. Hunt, 2007, Chapter 5.
36 Ericsson, 2003.
37 Wigdor & Green, 1991.
38 Schmidt & Hunter, 1998.

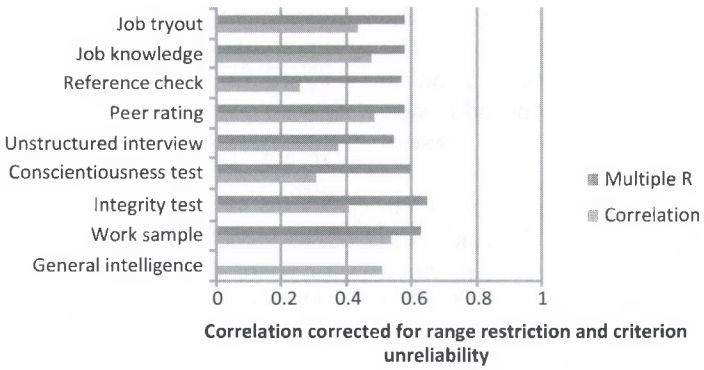

Figure 10.6. The correlations between measures of job performance, measures of general intelligence, and a variety of other assessment measures. The values shown are for the correlation between job performance and the assessment (Correlation) and the correlation between job performance and an optimum weighting of the assessment and the assessment of general intelligence (Multiple R). Data from Schmidt & Hunter, 1998, Table 1.

maximum validity, .65, by combining a measure of general intelligence with a test of integrity. (Conscientiousness is a close second, at .60.) This illustrates the combined importance of "can do" and "will do" traits.

The correspondence between the military and civilian data shows that the findings are robust over different situations and different methods of evaluation. The military data was gathered by direct observation of young adults; the civilian figures were based on a meta-analysis of dozens of small studies, covering all age ranges, but none as comprehensive or rigorous as the military studies.

The fact is clear. General intelligence has a predictive validity of about .50 in the workplace, just as it does in academia. No other method of assessment does any better. Nevertheless, people persist in using other techniques for predicting workplace performance. An examination of these alternatives is in order.

The only type of test with predictive ability greater than a test of general intelligence is a work sample (correlation of .54 compared to .51). This can be used in some situations. For instance, when musicians audition for places in major symphony orchestras they are often asked to play their instruments behind a curtain, so that the judges do not know who the candidate is. Work samples have two highly desirable qualities: they are statistically valid, and they are easily justified when assessment methods are challenged. Their drawbacks are that they can be rather expensive and that they can be used only if the candidates for a job have already been trained to do the job.

Combining a work sample and a general intelligence measure increases predictive validity to .63. The increase is not surprising, for by combining a general intelligence measure with a work sample the employer is simultaneously informed about the prospective employee's general reasoning powers and specific job knowledge.

In personnel selection situations a test this accurate, combined with a high rejection rate, can greatly increase the quality of the employed workforce. Recall that if no screening test is used, the average person hired should have an ability level equal to the fiftieth percentile (median) of the applicant population, regardless of the rejection rate. If a predictive validity of .63 is combined with a rejection rate of 50% (half the applicants are hired), the average ability level of a person hired will be at the sixty-ninth percentile of the applicant population.

An *unstructured interview* is an interview in which the recruiter and the candidate "just chat," so that the recruiter can get a

feel for the candidate. This is probably the most widely used selection procedure. The unstructured interview is not very good on its own (r = .38, corrected) and adds very little to the information gained in a test of general intelligence.

A *structured interview* (not shown in the figure, but included in Schmidt and Hunter's analyses) is an interview in which the recruiter has decided, beforehand, what topics are to be discussed in the interview, and what information must be provided. The technique requires a careful analysis of the requirements of the position to be filled, before searching for candidates. Structured interviews have good predictive validity, both on their own and when combined with a test of general intelligence (r = .51, R = .63).

Job knowledge is usually assessed by performance on a written test, where the questions are chosen to reflect what a job holder should know. This is a face-valid measure; we can reasonably expect bus drivers to know the rules of the road, and firefighters to know how to use various pieces of equipment. Job knowledge is not quite as good a predictor as is general intelligence, but it does add to predictive validity beyond that provided by a general intelligence score (r = .48, R = .58). In terms of the Gf-Gc model of intelligence, what a job knowledge test does is assess what the applicant knows about the particular situation in which he or she will be working. The same idea is captured by Robert Sternberg's emphasis on practical intelligence, which would include job knowledge. Sternberg and his colleagues have provided such tests, and they have on occasion shown some incremental validity.[39]

The practical intelligence tests Sternberg and his colleagues have described are very close to job knowledge tests. For instance, one practical intelligence test, designed for Alaskan hunters, asked what different pieces of evidence mean as indicators of coming weather.[40] Such questions measure crystallized intelligence (Gc) within a specialized context.[41]

10.3.3. *Upper-Level Managerial and Professional Positions*

The data presented so far is based largely on data from studies of blue-collar and white-collar jobs, up to the lower managerial level. In this population of occupations the correlation between general intelligence test scores and job performance generally rises with increasing job complexity.[42] Given this fact, it would be reasonable to expect the correlation to be still higher for high-level managerial, executive, and professional positions. However, there are reasons not to assume a straightforward extrapolation of the results to the managerial/professional class.

Many studies of high-level occupations report the observed correlation between test scores and measures of job performance, but cannot correct for selection restriction because there is no data on the applicant population. This is serious, because the selection effects are likely to be large. High-level positions are quite competitive, and are virtually always filled by people in the upper quartile of the intelligence range, IQ 110 and above. It is also difficult to find a measure of how well a professional or executive is doing, beyond gross judgments of satisfactory or unsatisfactory performance. As *Fortune* magazine repeatedly shows in its annual survey of executive salaries, the correlations between executive compensation and objective measures of company performance are close to zero. Physicians, attorneys, and other professionals are evaluated periodically, but the ratings are often limited to certification of competence without any further differentiation.

It is also often hard to acquire the required data on intelligence. People who

39 See Sternberg, 2003, for a general discussion, and Sternberg et al., 2000, for a compendium of many of the studies.

40 Grigorenko et al., 2004.

41 See Gottfredson, 2003a, and Hunt, 2008, for expansions on this point.

42 Gottfredson 1997, 2002.

occupy high-level positions are busy, and usually see no need to have their cognitive skills evaluated. As a substitute for direct observation, many studies look at professional training rather than on-the-job performance. As was shown in the section on post-graduate education, cognitive tests do surprisingly well in predicting completion of professional and managerial training.

There is also the problem of the multidimensionality of the criterion. People who occupy high-level positions are typically asked to do a number of different tasks. These range from high-level planning to public relations, face-to-face leadership, and negotiations. The relative importance of different tasks varies greatly across occupations and even from time to time within an occupation. It is not surprising that there has been a good deal of resistance to the idea that any unidimensional measure could predict performance at high levels. This has led to three different approaches.

In order to evaluate highly intelligent people using the conventional psychometric paradigm we have to have harder tests. Examples are the advanced version of the Raven tests, the Raven Advanced Progressive Matrices (RAMP), and the Miller Analogy Test (MAT), which contains difficult verbal analogy problems. These tests do predict job performance, with observed correlations in the .15–.30 range, depending upon the criterion used to evaluate performance, and with a predictive ability on the order of .40.[43] This is somewhat below the level of prediction obtained for higher-level skilled work, but still a reasonable figure.

It is somewhat more enlightening to look at a single large prospective study. During the 1960s the Bell Telephone System, at the time a near-monopoly covering telephone services in the United States, used the assessment center technique to select beginning managers. Management trainees spent several days in an assessment center, where they were rated for their ability to solve complicated problems both individually and in groups, and also given a cognitive test similar to the SAT reasoning tests, along with several personality tests. The results of these assessments were carefully shielded from their superiors in the company, so that the test scores could not influence supervisors' judgments or hiring decisions. (By contrast, in the US military a service person's scores on entry tests are part of his or her service record, and hence are available to commanders and promotion boards.) Twenty years later the assessment center results were validated by determining whether they predicted the level of management the candidate had achieved. The cognitive test ($r = .38$) was by far the best predictor.[44] As would be suggested by Schmidt and Hunter's analyses, personality tests had lower validity than cognitive tests, but did add substantially to predictivity.

Because high-level performance is said to be so multidimensional, some interesting alternatives to the conventional testing methods have been developed. One of the more popular of these is the *situational judgment test*. In this test an examinee is asked what he or she would do in a realistic, difficult situation. An example I particularly like, and that has appeared in a number of guises, is asking an applicant for a middle management position how they would inform their own supervisor that the supervisor's pet project was not working. As the example illustrates, an attempt is made to design situational judgment tests that draw on both cognitive skill narrowly defined and the examinee's social skills. Situational judgment tests add an additional .06 to the predictive validity that can be achieved by a cognitive test alone – not a large amount, but enough to be worthwhile in a large scale assessment program.[45] It is worth noting, though, that a situational judgment test asks the examinee what he or she would do in a hypothetical situation. It does not immediately follow that that is what the examinee would do, if placed in an actual, possibly emotional situation.

43 Kuncel, Hezlett, & Ones, 2004; Raven, 2008b.

44 Howard & Bray, 1988.

45 McDaniel et al., 2001.

To summarize, general cognitive ability is the best single predictor of executive/professional-level performance, just as it is of performance in the middle to high-end range of the general workforce. Prediction of executive/professional performance is somewhat less accurate than prediction of general workplace performance. There are several reasons why this might be so. They include difficulties in defining and obtaining measures of job performance, the extreme restriction in range of intelligence among applicants for high-level positions, and, possibly, the fact that general cognitive ability is a less dominating factor, compared to other dimensions of intelligence, in the upper ranges of cognitive competence than in the lower (see Chapter 4). Nevertheless, it is reassuring to know that among the movers and shakers in our society intelligence does count.

10.3.4. *The Rewards for Cognitive Skills in the Workplace*

The previous sections have shown that there is a positive relation between intelligence and workplace performance, within both military and civilian occupations. This is the sort of information employers want to have. From the viewpoint of an individual entering the workforce, the question is rather different. The individual wants to know what sort of economic niche he or she is likely to occupy, given a certain level of intelligence. As an ancillary question, what sort of rewards can the intelligent person look forward to, in terms of either money or occupational prestige?

Determining the statistical relationship between rewards and prestige is straightforward. You look at the correlation between test scores and some index of rewards. This can be done on an individual basis, or, as is sometimes easier to do, researchers can determine the typical level of intelligence of people in different occupations, and then look at the prestige and economic rewards offered by those occupations. However, as always, correlation does not necessarily mean causation.

In Chapter 1 I introduced the challenge hypothesis, the idea that within genetically prescribed limits people will increase their intelligence in response to a cognitively challenging environment. It has been shown, for instance, that there are qualitative differences in the reasoning skills of psychologists, physical scientists, and lawyers. Psychologists, who receive substantial training in statistics, are more sensitive to arguments based on probabilities than are people in the other two fields.[46] Were the psychologists, physical scientists, and lawyers attracted to their fields because the demands of the field matched their preferred styles of reasoning, or were the styles of reasoning determined by their experiences? The best way to answer this question is by a prospective study, where a person's intelligence is determined before he or she enters the workforce, and then related to the person's subsequent work history.

During World War II the United States Army Air Force (USAAF, the predecessor of today's Air Force, USAF) tested large numbers of young men who had applied to serve as aviation officers.[47] Two Columbia University psychologists, Robert Thorndike and Elizabeth Hagen, located approximately 10,000 of the men about twelve years after they had been tested.[48] At that time most of the men were in their early to mid-thirties.

Table 10.3 shows the mean test scores on general reasoning, verbal, and perceptual-motor scales of the original test, for men in selected occupations. These estimates were then converted to IQ ranges. The cadets who eventually entered those occupations that we consider more generally intellectually challenging, or that require a considerable amount of education, were also the cadets who had, at age twenty-one, scored high on the general reasoning tests. As we scan down the general reasoning scale we begin to encounter white-collar office jobs, and then various blue-collar jobs that, although they

46 Amsel, Langer, & Loutzenhiser, 1991.

47 Women were not accepted for aviation cadet training. Some women who had already qualified as aviators in the civilian sector did serve.

48 Thorndike & Hagen, 1959.

Table 10.3. Cognitive skills assessed in USAAF aviation cadets, shown by occupation followed after WW II. The far left-hand column shows estimated general intelligence measures using the conventional IQ scale. The estimate is based on a conversion of the general reasoning score, assuming that 0 on that scale corresponds to 105 on the IQ scale. Right-hand columns show mean scores achieved at approximately age twenty-one on three different composites of a testing battery. All three scales have a mean of 0 and a standard deviation of 100 in the sample of cadets.

Estimated IQ Range	*Occupation*	*General Reasoning*	*Numerical*	*Visual-Perceptual*
≥115	Chemical engineer	106	42	30
	Mechanical engineer	93	34	44
	Physical scientist	80	22	23
	College professor	75	38	38
	Civil engineer	75	31	56
	Electrical engineer	65	6	9
110≤ – <115	Physician	59	20	18
	Treasurer/comptroller	55	96	23
	Industrial engineer	44	31	34
	Lawyer	39	22	−7
	Personnel manager	33	18	13
	Pharmacist	29	39	−9
105≤ – <110	Dentist	28	20	15
	Accountant/auditor	28	54	−4
	Optometrist	14	34	−4
	Clergyman	13	1	−17
	Airplane pilot	13	10	−1
	Real estate salesman	6	17	6
100≤ – <105	Office manager	4	33	9
	Insurance underwriter	3	2	−9
	Veterinarian	−8	−2	−20
	Insurance claims adjuster	−13	−5	−9
	Bricklayer	−24	−5	−38
	Radio/TV repairman	−33	−37	21
95-<-<100	Hardware Salesman	−36	−12	−9
	Sales clerk	−40	−22	−28
	Plumber	−42	−21	−31
	Carpenter	−44	−17	−4
	Police detective	−50	−26	−20
	House painter	−63	−12	−24
<95	Crane operator	−66	−84	−37
	Vehicle mechanic	−72	−65	−7
	Assembler (in factories)	−83	−76	−40

Source: Data selected from Thorndike and Hagen, 1959.

may require considerable skill, are less intellectually demanding and can be learned on the job rather than via formal education.

Some pairwise comparisons are interesting. Cadets who became physicians had general reasoning skills similar to those of cadets who became treasurers/comptrollers (i.e., high-level financial managers), but the treasurer/comptrollers had higher numerical skills than the physicians. Cadets who became college professors had the same general reasoning skills as those who became civil engineers, but the engineers had higher perceptual skills. This illustrates the point

Table 10.4. Correlations between AFQT scores obtained at age sixteen to eighteen and measures of social outcomes when the participants were in their early thirties

Demographic Group	*Educational Attainment*	*Square Root Income*	*Occupational Status Index*
White men	.67	.39	.54
Black men	.48	.29	.45
White women	.59	.31	.45
Black women	.53	.44	.47

Source: Data from Scullin et al., 2000, Table 2.

that in many occupations general reasoning skills have to be augmented by more specific cognitive skills.

Other pairwise comparisons show that occupations can be the same in terms of the stress they put on different special abilities, but differ in the level of associated general reasoning skills. Treasurers and comptrollers (high-level financial officers) and accountants and auditors (financial technicians) were both characterized by high reasoning and numerical skills, but the cadets who became treasurer/comptrollers had higher levels of these skills.

While the Thorndike and Hagen study is certainly informative, there are some aspects that limit sweeping conclusions. The participants, aviation officers, were an intellectually select group, with an estimated mean IQ of 105. The occupations these young officers entered, following the war, were fairly high on the occupational scale. And the test battery was designed for aviation cadets. Accordingly, compared to the typical battery-type IQ test the aviation battery was biased toward the evaluation of visual-perceptual skills – so much so that verbal skills were incorporated within the general reasoning factor. Even given these limits, a coherent picture emerged: intelligence is worth quite a bit.

Thorndike and Hagen studied the workplace of the mid twentieth century. Compared to that workplace, today's workplace places much more emphasis on the manipulation of data and abstract representations, rather than things.[49] Have these changes in the workplace changed the requirements for cognitive skills?

To address that question we look at a second prospective study using the NLSY79 data (see panel 9.9). Recall that in this study a nationally representative sample of adolescents and young adults took the AFQT. In 2000 a research group at Cornell University investigated the occupational and income status, as of 1995–96, for over 2,400 of the panelists who had been born in 1963–64.[50] The men and women in this study were of approximately the same age as the former aviation cadets studied by Thorndike and Hagen, but in a cohort born roughly forty years later.

In general, the Cornell group found that there was a moderate positive correlation between AFQT score, educational attainment, income, and occupational prestige (SEI). The correlations are shown in Table 10.4, reported separately for four different demographic groups, White and Black men and women. Although the correlations are all positive, and never negligible, they do vary markedly across demographic groups. As a rough generalization, intelligence seems to be a more useful predictor of future success if you are a White man or a Black woman than if you are a White woman or a Black man.

Table 10.4 treats educational attainment as an outcome. Education can also be thought of as something to be achieved en route to further social and economic success, rather than as an end in itself.

49 Hunt, 1995; Reich, 1991; Zuboff, 1988.

50 Scullin et al., 2000.

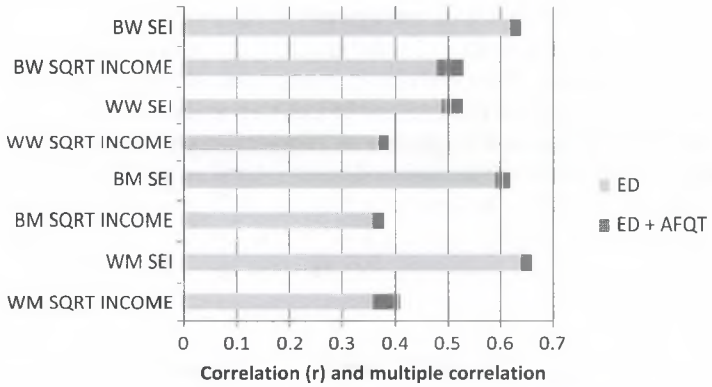

Figure 10.7. Predicting square root income (SQRT INCOME) and occupational prestige (SEI) in 1995 from educational attainment (in years) and AFQT score obtained in 1980. Respondents were sixteen to eighteen years old in 1980. Codes: WM – White men, BM – Black men, WW – White women, BW – Black women. Data for calculations from Scullin et al., 2000, Table 2.

Not surprisingly, educational attainment and AFQT score are substantially correlated in this sample, ranging from .67 (White men) to .48 (Black men). This is consistent with the data reported in section 10.2, relating intelligence to academic achievement. Figure 10.7 shows that there is little added predictive value of knowing a person's AFQT score, once his or her educational attainment is known.

There are three possible interpretations of this finding. One is that education is the proximal variable that determines socioeconomic outcome and income, while intelligence acts as a distal variable, determining education but then playing no further role. The other is that intelligence acts as a proximal variable, and that education serves as an additional statistical marker for intelligence beyond the AFQT score. These explanations can be discriminated by contrasting the partial correlations between AFQT and the outcome variable, occupational prestige or income, holding education constant, or between education and the outcome variable, holding AFQT score constant. I calculated these and found that for income the partial correlations for education are generally larger than for AFQT score, but that the differences are not great. The results for occupational prestige are striking. The partial correlations for education given AFQT score are .43 for Blacks (both men and women), and .38 and .28 for White men and women, respectively. The corresponding values for AFQT given education range from .16 (White men) to .20 (White women). Evidently education and intelligence are collinear predictors of income. Intelligence acts as a distal variable, exerting its influence through education, which then permits entry into prestigious occupations. The educational effect seems to be stronger for Blacks than for Whites.

Another way to determine what the workplace is willing to pay for intelligence is to examine the test scores of people who apply for jobs in various occupations. This carries with it the defensible assumption that the applicants exert self-selection. People without college degrees generally do not apply for entry-level executive positions.

Gottfredson used a number of studies of intelligence test scores in job applicants to construct a "life's chances" chart for various occupations.[51] She associated IQ equivalents with five classes of occupation, ranging from what she regarded as "slow, supervised" work to "gathers own information." Table 10.5 lists occupations cited by Gottfredson, along with her estimates of the typical IQ score for an applicant.

51 Gottfredson, 1997.

Table 10.5. Gottfredson's examples of intelligence levels (on IQ scale) associated with different occupations and techniques of information processing. The columns to the right of Gottfredson's figures show the intelligence-level estimates obtained from Thorndike and Hagen's study of the careers of USAAF aviation cadets, approximately fifty years earlier (Table 10.4) and the range of Wonderlic scores of applicants reported in the Wonderlic Corporation's Normative Report (April 2007) for the WPT-R, revised in 2003. As occupations used in the norming for the WPT and WPT-R, a comparable profession to Gottfredson's "typical profession" was used. Wonderlic scores have been converted to IQ units using the conversion 2 * Wonderlic score + 60 (Dodrill, 1981; Dodrill & Warner, 1988).

Training and Qualification Method	*Typical Position*	*Gottfredson's Estimated Typical IQ Score*	*Estimates Based on Thorndike & Hagen Data*	*WONDERLIC with Comparison Occupation*
Explicit, hands-on training	Assembler	80–90	<95	88–104 Electro-mechanical assembler
Mastery learning, hands-on training	Police officer	95–105	95–100	100–114 Police and sheriff officer
College formal instruction	Accountant	110–120	105–110	102–120 Accountant
Graduate instruction, gathers own information	Attorney	120 +	110–115	110–124 Executive

The table also includes data for comparable occupations, based on the Thorndike and Hagen data, taken forty years earlier, and for the 2003 revision of the WONDERLIC test, the WPT-R. There is a striking similarity between Gottfredson's estimates, the WONDERLIC estimates, and the estimates that Thorndike and Hagen had made forty years earlier. The data was gathered using different sampling methods, in different workplaces separated by over half a century, and the tests used were quite different.[52] Nevertheless, the estimates are basically the same. Comparing Tables 10.4 and 10.5, it appears that the role of intelligence in today's workplace is very much the same as it was before the term "information technology" was invented.

A second retrospective study provides a more comprehensive look at general levels of occupational accomplishment, but does not provide data on individual occupations.

In the late 1980s the Center for Disease Control conducted a follow-up study of the health of people (primarily men) who had served in the military during the Vietnam War era (1967–76). A large number of veterans were contacted and asked to participate in extensive physical and mental testing. Helmut Nyborg, a Danish psychologist, and Arthur Jensen related the veterans' current occupational status to their educational attainment and to a measure of *g* extracted from their test scores, as of the late 1980s.[53] Figure 10.8 shows the results

52 Thorndike and Hagen's general reasoning factor was extracted from a battery of subtests that took several hours to complete. Gottfredson's estimates were based on data from the Wonderlic Personnel Test (WPT) The current Wonderlic estimates are based on the revised version, WPT-R, with data from 2003. See Chapter 2 for a description of the WPT.

53 Nyborg & Jensen, 2001.

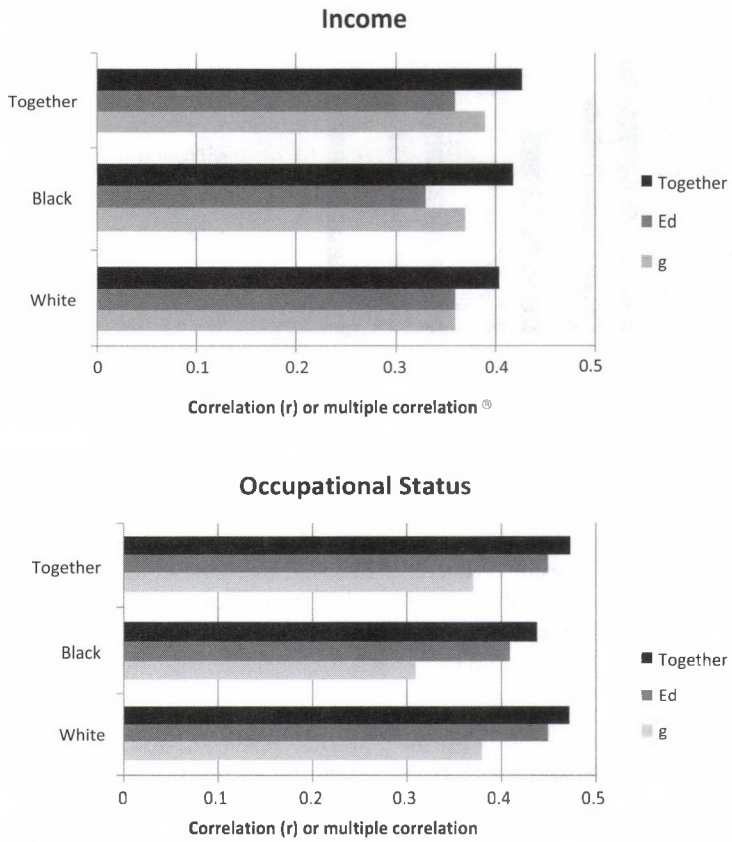

Figure 10.8. Correlations and multiple correlations between income (top) and occupational status (bottom) and general intelligence and educational level. The calculations are based on Table 3 of Nyborg & Jensen, 2001.

of Nyborg and Jensen's analysis. Data are shown separately for black and white veterans, as they differed on educational, intelligence, occupational, and income measures.

In this sample the intelligence test score was a slightly better predictor of income than was educational achievement, while the reverse was true for the NLSY data. It would be unwise to make very much of this. The samples were different, the tests were different, and the NLSY analysis attempted to predict future accomplishment from measures taken in adolescence, while Nyborg and Jensen related current test performance to current accomplishment. In any case, the similarities are far greater than the differences.

Both educational level and intelligence are substantial predictors of accomplishment in the workplace. The two are highly correlated. This is hardly surprising. Both measures are based on the development and display of cognitive skills.

10.3.5. *What the Jobs Demand*

Another way of determining the value of intelligence in the workplace is to analyze the job requirements of a large number of occupations, and infer what this means in terms of the demands on intelligence. The first step is to make an analysis of the relative value of cognitive skills for different jobs.

The US Department of Labor (DOL) maintains an elaborate job counseling service, in which it describes over 12,000 jobs and rates the extent to which they require certain skills. The skills rated range from general reasoning ability to finger dexterity. The rating system was originally

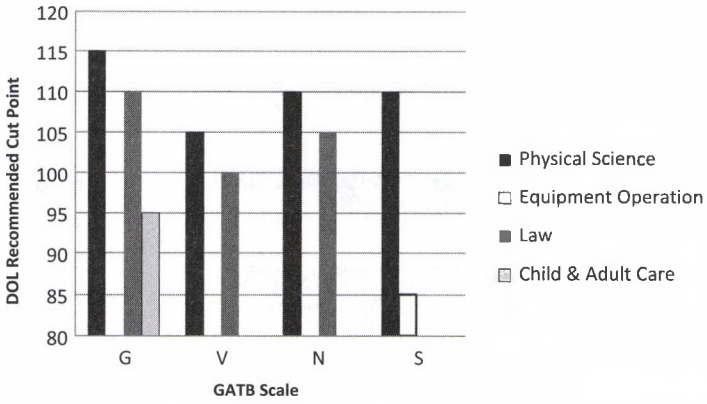

Figure 10.9. Recommended GATB cut points for four occupations, taken from high-status and lower-status patterns, within the class of occupations dealing with physical or with social relationships. GATB scales are G – general reasoning, V – verbal, N – numerical, and S – spatial. Data selected from Gottfredson, 1986, Table 2. In cases where no value is specified the DOL analysts had made the judgment that virtually anyone would have sufficient ability to do the job. For instance, there is no cut point for spatial reasoning required for the law, and no cut point for general reasoning required for heavy equipment operation.

incorporated into a descriptive volume called the *Dictionary of Occupational Titles* (DOT). The DOT has been superseded by an on-line, interactive system called O*NET. O*NET is a considerable expansion over the DOT, designed primarily to help job seekers. As a side benefit, it contains a massive amount of data available to researchers interested in issues involving workforce skills.

Gottfredson utilized the original DOT system to construct a "space" of jobs.[54] Her analysis coordinated job ratings with data on the test performance of people who had applied for, or were occupying, a variety of jobs. She identified five classes of occupations – those dealing with physical relations, social and economic relations, maintaining bureaucratic order, and performing. (She also had a small class of "leftover" occupational patterns that will not be dealt with further.) Within each of these classes she identified the patterns of aptitudes required. For example, within the class of occupations dealing with physical relations there was a cluster of occupations that dealt with research and design (e.g., physicist, engineer) and a cluster that dealt with building, maintaining, or operating physical objects (e.g., equipment operators, craftsmen). Within the class dealing with social and economic systems one cluster dealt with research and design (including social research, law, and finance), while another dealt with providing service to individuals (including hospitality services, and child and adult care).

Individual jobs within a cluster could be associated with a pattern of abilities, as defined by the DOL's General Aptitude Test Battery (GATB), which was in use at the time. The DOL used these values to recommend minimum values (cut points) for a job along each of four GATB dimensions: general reasoning, verbal reasoning, numerical skills, and spatial skills. Figure 10.9 provides examples for four occupations, two from Gottfredsons's class of occupations dealing with physical relations and two from the class dealing with social relations. Within each class one occupation was selected from Gottfredson's high-status cluster, and the other from a lower-status cluster. Physical scientists and lawyers,

54 Gottfredson, 1986.

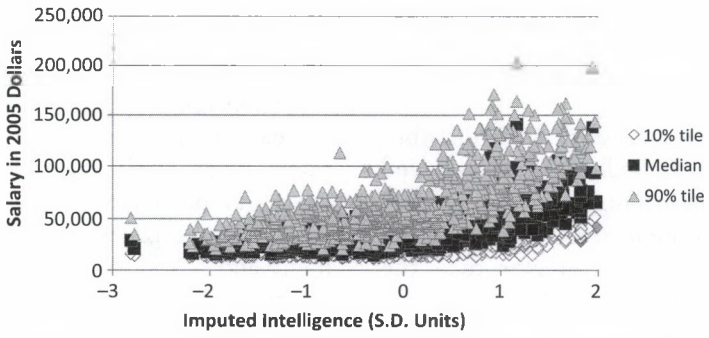

Figure 10.10. The ninetieth, fiftieth (median) and tenth decile of incomes in various occupations, plotted as a function of the imputed intelligence demands for the occupation. Data from US Census Bureau and from Hunt and Madhyastha's analysis of job demands.

high-status occupations from the physical relations and social relations clusters, have similar high-level patterns except that lawyers are not required to have spatial reasoning skills, and do not have quite as high scores on other scales as the physical scientists do. Equipment operators have to have a certain minimal level of spatial reasoning skills, while personal caretakers have to have a minimal level of general reasoning skills.

Gottfredson found that general intelligence was by far the biggest driver of variations in cognitive skills across occupations. Verbal, numerical, and spatial skills were important in some occupations, but they accounted for much less of the variation in the descriptions of occupational requirements than did differences in requirements for general reasoning.

Tara Madhyastha and I have analyzed the modern O*NET data (as of 2008) using different techniques than Gottfredson had used, but intended to answer basically the same questions. In general, our analysis of cognitive demands agreed well with Gottfredson's.[55] We found that the ratings of skills could be described by a two-dimensional space, where one dimension was a general reasoning factor and the other was a *bipolar* factor, indicating whether the job emphasized verbal or perceptual-motor skills. We also found a smaller factor indicating the extent to which a job required numerical skills. The factors were not statistically independent, because most jobs that require a high degree of general reasoning also require fairly high verbal skills, just as Gottfredson had noticed. This is not surprising; people who hold intellectually demanding jobs usually have to communicate with other people. In spite of comic strip stereotypes, real computer programmers spend a great deal of time describing what their programs do, and how they fit into suites of programs developed by other people.

Our analysis also allowed us to relate the incomes associated with different occupations to the cognitive demands of those occupations. The relations to general intelligence are shown in Figure 10.10. The income associated with occupations increased as the occupations demanded higher levels of intelligence. However, there appeared to be a nonlinear increase; incomes are fairly flat over occupations that require slightly more than average intelligence. Incomes are more closely related to demands on intelligence if the occupation requires an IQ of more than about 105 ($g \geq .3$ on the standard deviation scale). What is even more striking is the range of incomes associated with a given level of cognitive demand. This is clearly shown in the figure. The range in annual income between the 10% best-paid and 10% worst-paid occupations with a cognitive demand of 100 (0 in standard deviation

55 Unfortunately it is not possible to compare our analysis to the data obtained by Gottfredson or by Thorndike and Hagen, because our analysis was at a much finer level of detail.

units) was about $65,000 (2005 dollars). The range for jobs with a cognitive demand of 116 (1 in standard deviation units) was roughly $140,000.

Figure 10.10 also shows that the distribution of income is markedly positively skewed at all levels of intelligence. The absolute difference between the median and the ninetieth percentile is always greater than the absolute difference between the median and the tenth percentile. The extent of the skew increases with increasing imputed intelligence, a trend that begins much sooner than the rise in median income as a function of imputed intelligence.

10.3.6. *Summary: The Role of Intelligence in the Workplace*

The quotations that began this section, from Gottfredson and from Kuncel and colleagues, were accurate. Intelligence predicts a person's job status and income better than any other trait that has been studied. This leaves us with two questions: why does the association exist, and why is it that the association is so widely denied by people who have not studied the topic?

The influence of intelligence is undoubtedly mediated in part by education. This is especially true across fields, for the vast majority of the more lucrative occupations have, as an entry requirement, at least a college education. The best-paid professional occupations require substantial graduate-level training. It is increasingly the case that the entry routes to skilled trades, such as auto mechanic, involve academic certification through community college or other professional training programs. Because educational attainment is strongly related to intelligence, any variable that is correlated with educational attainment will also correlate with intelligence.

In the case of the professions education is essential; no one wants to have a surgeon (or an airline pilot) who is learning basic skills on the job. To the extent that intelligence is required to get through a rigorous training program, intelligence and education are entwined.

Charles Murray has argued that in many cases education is not really needed for success in a field, but that education is used as a (socioeconomic?) screening device.[56] To the extent that this is true a spurious relation between intelligence and economic success could be created, via a real relationship between intelligence and educational attainment and a spurious one between educational attainment and economic success, calculated across occupations. However, this is not completely the case. Both intelligence and education are important, because there are nonzero partial correlations between indices of workplace success and either intelligence or education, after the other has been held constant.

The fact that intelligence is correlated with on-the-job performance in a wide range of military and civilian occupations, including ones that do not have high educational requirements, provides further evidence that intelligence is important in itself, not just as a facilitator of education.

10.4. The Social and Economic Prospects at the Ends of the Bell Curve

If intelligence is an important trait in our society, then the lifetime prospects of people on the two extremes of the distribution of intelligence should be very different. And they are. In this section we take a look at the careers of some people who are at the upper and lower ends of the intelligence distribution.

It is important to be clear that we will *not* be looking at people who are conventionally labeled "geniuses" or at people who are mentally handicapped to the point that they cannot function in our society without special help. There are reasons for avoiding these extremes.

The term "genius" is usually applied to people who have accomplished great things. The consensus of people who have studied genius is that extraordinary accomplishment in any field requires some talent, but

56 Murray, 2008.

also a great deal of motivation, and very hard work. The social support network must be right. Howard Gardner makes this point in his excellent study of extreme creators, in such varied fields as physics (Einstein), writing (T. S. Eliot), politics (Gandhi), and art (Picasso).[57] Gardner's subjects were geniuses in any sense of the word. They were all very bright. They also had a single-minded sense of purpose *and* a social network of people who were willing to support their single-minded efforts. In addition, the times must be right. An unknown Sumerian genius invented the wheel around 3500 BCE. His or her invention spread rapidly through the ancient world, for it greatly improved the utility of oxen and horses. Perhaps 2,500 years later, and completely independently, an equally unknown Aztec invented wheels for children's toys. The idea never went further, for the Aztecs had no beasts of burden. Cognitive traits undoubtedly are important in the creation of genius, so are noncognitive traits and features of the situation. Studying acknowledged geniuses is important in itself,[58] but it is not a good way to determine how high intelligence affects one's progress through life.

At the other end of the scale, there is little point in studying the lives of the extremely mentally disabled, who simply cannot cope with our society. Determining the sorts of social support these unfortunate individuals require is an important topic, but one that is far beyond the scope of this book. What we can do is to examine the lives of people who fall in the "low normal" range, roughly IQ scores of 70–85. Most of these individuals are productive members of the society, but seldom maximally productive members.

Once again, we have to balance the relative costs and benefits of retrospective and prospective studies. There have been numerous retrospective studies of the characteristics of high achievers, ranging from artists to politicians.[59] There have been even more studies of various low-achieving groups, such as welfare recipients and criminals. In both cases it is possible to find some traits that seem to characterize the target group. However, as is almost always the case with retrospective studies, it is hard to interpret these findings. For instance, eighteen of the first forty-four presidents of the United States received earned degrees from one or more of just five colleges.[60] Does this mean that if you attend one of these colleges you have a good chance of becoming president? Hardly. Only a miniscule fraction of the graduates of these colleges attained the presidency.

At the other end of the social scale, it has been estimated that about one in every six homeless persons has either schizophrenia or manic-depressive psychosis. Does this mean that a person who suffers from either of these diseases has roughly one chance in six of becoming homeless? Hardly. Approximately 3% of the US population, roughly nine million people, suffers from one of these two diseases. Some 200,000 are both homeless and are either schizophrenic or suffer from bipolar disorder. The chances are roughly one in forty-five, not one in six, of a mentally ill person becoming homeless.

The best way to determine what happens to intellectually gifted or below-normal individuals is to start with a group of (gifted) (below-normal) persons and follow that group through some portion of their lives, in a prospective manner, rather than attempting a retrospective study of people who have had a particular life outcome.

We will look first at a study of the gifted, and then examine the low-normal group. In each case my discussion will use as illustration the results of one or two large studies.

57 Gardner, 1993b.

58 See, especially, Simonton, 1984.

59 See, Simonton, 1984, and Gardner, 1993b.

60 The colleges are Harvard (six), the College of William and Mary (four), Yale (four), Princeton and the US Military Academy (two each). The list includes George Washington, who received a surveyor's certificate from William and Mary. This was his only post-secondary education. I have assigned George W. Bush to Yale, where he received a B.A. He also received an MBA from Harvard.

Table 10.6. Percentages of men and women in the Terman study who attained various levels of education, compared to educational achievements of their cohort

Group	*Men in Terman Study*	*Women in Terman Study*	*General Public, Men*	*General Public, Women*
College graduates	70	67	10	6
Graduate studies given graduation from college	56	33	19	Unknown
Doctorates	14	4	2	2

Source: Data from Terman & Oden, 1959.

10.4.1. *The Gifted I: The Quiz Kids and the Termites*

During the 1940s and 1950s there was a popular radio program called *The Quiz Kids*. A panel of very knowledgeable six- to sixteen-year-old children answered questions that required anything from an ability to do rapid mental calculations to knowing rather obscure scientific and historical facts. Their performance was impressive. Some were said to have IQs of 200, but that was apparently a score based on the old mental age/chronological age calculation. A more realistic estimate is that the IQ scores ranged in the 140–160+ range, roughly the top one in one thousand.

About thirty years later one of the former Quiz Kids, who had become a professor, located a number of them to see how they were doing.[61] The commonest answer was generally quite well, thank you. An inordinate number of them had followed academic professions. The others were mostly in professional fields. One had received the Nobel Prize in Medicine and Physiology.[62] To be sure, not every one of the Quiz Kids had done well, and some had rather unhappy lives. But, on the whole, they were successes.

The Quiz Kids hardly represents a scientific study. Candidates were recruited rather informally from the general Chicago area. I am sure the selection of candidates depended on both the child's apparent intelligence and the radio show producers' judgment about how appealing the child would be on the radio stage. Fortunately, more formal studies have been done.

The Quiz Kid idea was a popularized version of a well-known study that had been (and was being) carried out by Louis Terman, the Stanford University professor who introduced the Binet tests into the United States. In the early 1920s Terman asked teachers in California schools to nominate exceptionally bright children for a long-term study. The children were then given IQ tests, and those who scored 140 or above (one in a thousand) were invited to participate. Eventually 1,528 students were enrolled. The study continued after Terman's death in 1956. Eventually the "termites," as they were sometimes called, were followed into their seventies.[63]

The results were clear in one way, and difficult to interpret in another. Terman's participants were born around 1910–20, so the cultural aspects of their time have to be kept in mind. In spite of living through a depression and a world war, they did exceptionally well. A few statistics from the last study in which Terman himself participated, at which time the "termites" were in their fifties, shows what had happened.[64]

By the 1950s virtually all of the people in the study had completed their education. Table 10.6 provides a comparison of their

61 Feldman, 1982.
62 James Watson, for the discovery of the structure of DNA.
63 Holahan & Sears, 1995.
64 Terman & Oden, 1959.

Table 10.7. Family income distribution of Terman study participants in the 1950s, compared to "urban white families" at that time. Income figures are in 1950s dollars. The study participants earned much more than the base rate established for similar families.

Group	*"Urban White Families"*	*Terman Participants*
Income > $15,000	1%	30%
$15,000 ≥ income > $5000	36%	64%
$5000 > income	63%	6%

Source: Data from Terman & Oden, 1959.

achievements compared to those born in roughly the same cohort. Clearly the gifted had much greater educational attainment than was typical of the time. The college graduation rate for the gifted, who went to college during the Great Depression, was higher than the general graduation rates at the start of the twenty-first century! This was true for both men and women.

Over 80% of the men in the study followed professional or business careers. As was typical of the times, many of the women became homemakers. Table 10.7 contrasts the family incomes of the study group to a group that Terman referred to as "urban white families." This was an appropriate comparison group, for the participants were themselves predominantly urban and white.

The final follow-up of this group, when they were in their seventies, reinforced the picture. The "termites" had achieved high educational levels and had had successful careers. They were healthy and satisfied. Their marriage rates were high, compared to their cohorts, and their divorce rates were low. The incidence of severe mental illnesses, alcoholism, and other types of social dysfunction was similarly low. The study gave no support whatsoever to the stereotype of the sickly, neurotic genius. Nor, I add, has any other study of the gifted.

To what extent was the success of the termites due to their intelligence? Here the answer is not so clear. Terman's work has been criticized on three grounds. The most serious is that reliance on teacher reports and personal contacts biased the study toward the selection of upper-middle-class urban Whites. The second is that Terman actively interfered in the lives of the participants, through interviews and mail contacts. The third is that the participants were not really geniuses.

Terman's selection methods were biased toward selecting children from high SES, White homes. On the average, although not in all cases, Terman's participants benefited from strong social support, such as having families who could support them during their college years, and having the social contacts that facilitated success. In hindsight, it would have been nice if Terman had also followed children from a comparable SES group who did not have unusually high IQ scores. This would have greatly increased the cost of the study. The disparity between success rates in Terman's group, compared to base rates in similar social groups, is so great that I do not think that the bias toward upper SES participants could have entirely accounted for the results.

Terman's selection methods were biased against identifying highly intelligent children in low SES families or in minority groups. The bias against minorities was a side effect of the bias toward selecting children from families with relatively high SES, because in California in the 1920s the difference in SES between Whites and Blacks or Latinos was much greater than it is today. There may have been many talented individuals who should have been in Terman's group but were not selected. However, Terman never set out to study all, or even a representative group, of gifted schoolchildren in California. His intention

was to find some gifted students and follow them literally throughout life. He did this. What the distribution of gifted is in the general population is a different question.

Would the conclusions have been any different had there been an aggressive attempt to recruit bright Black and Latino children? We will never know.

10.4.2. *The Gifted II: The Study of Mathematically Precocious Youth and Related Studies*

We now turn to a more modern, much larger study that is every bit as ambitious as Terman's was, but has a more clearly defined recruitment procedure.

In 1971 Julian Stanley, a professor at Johns Hopkins University, began the Study of Mathematically Precocious Youth (SMPY).[65] Students in middle schools (then called "junior high schools") were urged to take the SAT when they were twelve or thirteen years old. Recall that the SAT is designed for students in the third or fourth year of high school, at age sixteen to eighteen. Stanley initially focused on mathematical precocity, but in 1973 he began to study verbal precocity as well.

A two-tiered selection procedure was used. The SMPY researchers identified those twelve- to thirteen-year-old students who had scored in the top 3% on standardized tests that had already been given, as part of their school's normal assessment program, and asked them to take the SAT. Students who scored 500 or higher on either the mathematics or verbal (SAT-M or SAT-V) portion of the test were asked to participate in the main study. This score would put a twelve- or thirteen-year-old student in the top half of seventeen-year-olds. Such a level of accomplishment on the SAT-M is an impressive accomplishment, because this test evaluates proficiency in mathematical topics that are often not covered until late middle or high school. The middle school students either had to have studied these topics outside of school or had to apply general reasoning to solve the problems on the test. The researchers regarded the students who scored 500 or above to be in the top one-half of one percent (1 in 200) in their age group, while the students who scored 700 or higher were believed to be in the top one-hundredth of one percent (1 in 10,000). The design of the study thus permits comparison between the bright and the extremely bright.

Stanley died in 2005. The SMPY study has been carried on by Stanley's colleagues, Camilla Benbow and David Lubinski, at Vanderbilt University. Their intent is to make SMPY a fifty-year study of the careers of people identified as talented in their early adolescence. (Stanley's first participants were born in the late 1950s and early 1960s.) In addition to observing the behavior of talented students, Stanley and his colleagues were interested in the nurturing of exceptional talent, especially in mathematics. Therefore, the SMPY has included a teaching component, in which participants are enrolled in intensive summer programs.

The data that has been gathered at this time (2010), which roughly carries the participants through their thirties, provides insight into several aspects of the development of the gifted. These include reactions to academic environments, academic achievement, and accomplishments by early middle age.[66]

The students took to intensive academic instruction like ducks to water. The SMPY students could assimilate a year's high school course work in about three weeks of intensive study. This is consistent with other reports, indicating that when special instruction is offered it helps everyone, but it helps the talented students the most.[67] I would add a caveat to this. The statement is true unless the special instruction is highly structured and specifically developed for a low-ability group. In this case the high-ability group's accomplishments may

65 Stanley, 1996; Brody & Blackburn, 1996.

66 The data cited here is largely taken from Lubinski and Benbow's (2006) review of the status of the project thirty-five years after the initial enrollment of students.

67 Ceci & Papierno, 2005.

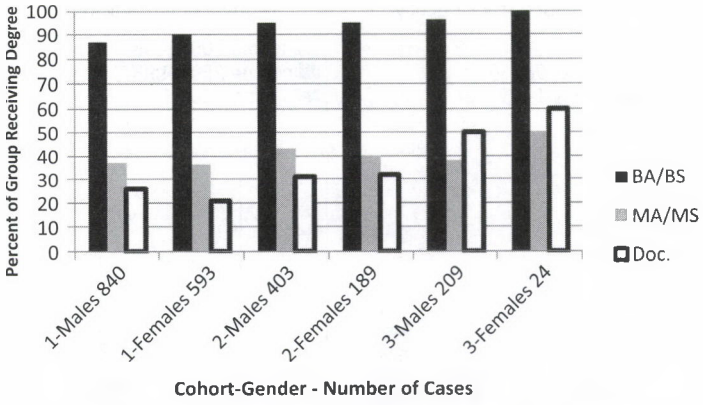

Figure 10.11. Educational attainment of participants in the SMPY, by age thirty-three. Members of cohorts 1 and 2 attained scores of 500 or better on the SAT-M by age thirteen. Members of cohort 3 attained scores of 700 or better. For reference, in the general population in the US as of 2004, 27% of the population twenty-five or older had received bachelor's degrees, 9% had master's degrees, and 3% held a doctorate of some type, including the Ph.D., M.D., L.L.D., and E.D. degrees. Data from Lubinski & Benbow, 2006.

actually deteriorate due to loss of interest and motivation.[68]

The academic achievements of the group are staggering. Figure 10.11 shows the level of academic achievement obtained by SMPY participants who had achieved scores of either 500 or greater or 700 or greater on the SAT-M. Lubinski and Benbow have pointed out that it is, to say the least, interesting that a two-hour test taken at age thirteen or younger can predict that the "risk," if you will, of obtaining a doctorate has risen from three in one hundred, the U.S. national rate, to one in two (cohort 3 in Figure 10.11).

What about on beyond schooling? At the time of this writing (2010) the SMPY participants are less than fifty years old. Nevertheless, their accomplishments are spectacular. Because many of the participants went into science, engineering, and related careers, a particularly compelling comparison contrasts the SMPY participants to graduate students of a similar age who had attended highly ranked university programs in science or engineering. Figure 10.12 shows this comparison for a distinctly academic criterion, employment at highly ranked universities, and a distinctly nonacademic criterion, how much money the individual was making. By age thirty-three approximately 8% of the "over 700" men were earning more than $250,000 as individuals. In 2005 (the time of the survey) approximately 1.5% of all US *households* earned more than $250,000.

The SMPY was initially motivated by a desire to study the life histories of adolescents who showed exceptional promise in mathematics. As the study progressed it extended its reach to the evaluation of people with high SAT-V scores. This made it possible to study people whose talents were tilted toward either mathematical or verbal skills. It should be remembered, though, that people with very high verbal scores will probably have above-average mathematics scores, and vice versa.

The tilt definitely affected the type of contribution that was made. SMPY participants whose strongest test scores were on the SAT-V tended to major in the social sciences and humanities; those with high SAT-M scores chose the sciences and engineering fields. This difference held true both for

68 Snow, 1982, 1996.

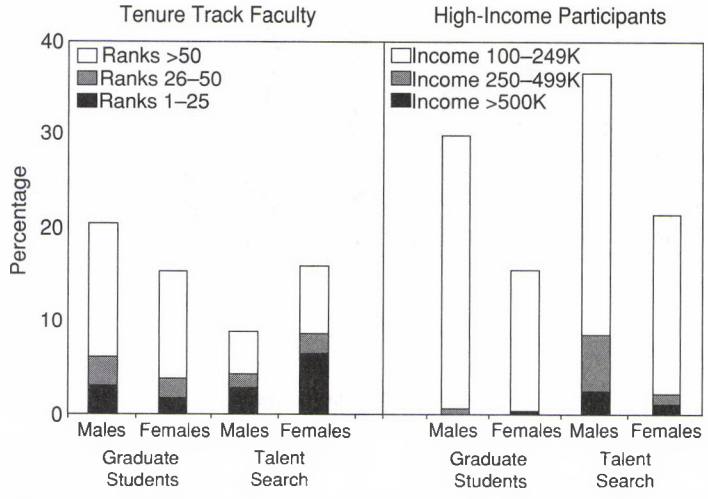

Figure 10.12. The SMPY participants compared to graduate students of the same age. A comparison of the accomplishments, by age thirty-three, of SMPY participants who had scored 700 or better on either the SAT-V or SAT-M, at age thirteen, to the accomplishments of graduate students from highly ranked U.S. university programs. From Lubinski and Benbow, 2006, Figure 3.

choice of major and for later professional work.[69]

The mathematically precocious youth did not rest on their laurels. They were willing to work more hours than the graduate students to whom they were compared – even though mathematics and science graduate students, all in highly rated programs, are a notoriously hard-working group. Somewhat more men than women qualified for the program. The disparity between men and women increased sharply as a function of the SAT-M score. This can be seen from an examination of Figure 10.11. The male/female ratio in cohorts I and II, which were in the top 1/200 group, is 1.6 to one. The male/female ratio in cohort III, the 1 in 10,000 group, was 8.7 to one.

The SMPY participants tended to come from small families of relatively high SES. Their parents were much more likely to have college degrees, including advanced degrees, than would be the case for randomly selected students. These two relationships are not independent, for family size is negatively correlated with SES.

Only about 1% of the SMPY participants were African American or Latino. This is comparable to the figure in Terman's study, although the percentage of African American and Latino students in the schools had increased markedly between the SMPY recruitment and Terman's recruitment, fifty years earlier.

Thirty-three percent of the participants in the SMPY and related programs are Asian.[70] Nationally, Asians constitute about 4% of the population. It has been claimed that Asians have a greater genetic potential for intelligence, and especially for mathematical reasoning. This claim is somewhat controversial.[71] However, genetic potential is unlikely to have been a cause for the overrepresentations of Asians. Eighty percent of the Asian participants had parents who had been born and educated outside of the United States. Were the contrast to be due to genetics, one would have expected a much higher percentage of Asians whose

69 Park, Lubinski, & Benbow, 2007.

70 Brody & Blackburn, 1996. Unfortunately these authors did not provide figures for the various Asian groups.

71 See Flynn, 1991. This topic is discussed in more detail in Chapter 11.

parents had been born in the US. Nor is the contrast likely to be due to some great deficiency in US schooling, for most of the non-Asian participants had been educated in the US school system. This strongly suggests that environmental factors influencing the home life of recent Asian immigrants were a major factor in the children's developing interests and skills in academic pursuits.

The Terman study and the SMPY are not the only studies of the gifted, but they are perhaps the largest, and their results are representative. People who do well on cognitive tests in their late childhood or early teens have quite bright prospects. High scorers tend to come from fairly high SES families, so some of their success may be due to the advantages of privilege. However, the accomplishments of the gifted are too substantial to support a claim that advantage is all they have. Their own ability counts for a great deal.

10.4.3. *Developing the Gifted*

A variety of acceleration programs have been developed to assist the gifted in reaching their potential. These vary from special summer courses, as in the case of the SMPY, to provision of accelerated tracks in public high schools and offering children as young as fourteen early admissions to universities. Gifted children perform very well in such programs. They also report enjoying them. There is little evidence for widespread social maladjustment, although participants in the early college entrance programs report preferring the company of their equally gifted age-mates to that of the considerably older college students. One particularly telling comment was made by Nancy Robinson, a professor at the University of Washington who, together with her husband, Halbert, developed an early entrance program at that university. She noted that gifted students who came from the public schools benefitted from a pretraining program that prepared them for the pace of instruction at a major university. Why? Because they needed to develop study habits! They had been able to get by in their regular schools without exercising the discipline needed when they were thrown into the far less supportive atmosphere of university instruction.[72]

10.4.4. *A Comment on Criticisms of the Concept of the Gifted*

On occasion the results of the studies of the gifted have been criticized in ways that I think are unfair or irrelevant. For instance, I have heard the Terman study criticized for not having identified a Nobel laureate! There are also widespread, although as far as I know undocumented, stories that two men who did subsequently win Nobel Prizes in science were overlooked during the selection of participants. Such criticisms set an unrealistic standard, by asking the researchers to predict a one in a million event, while ignoring major trends in the data.

Another criticism is a claim that very high scores on cognitive tests do not predict anything. This belief is extremely widespread, even among psychologists.[73] It is false. In the SMPY there is a substantial difference between the accomplishments of the top 1 in 200 among test scorers and the accomplishments of the top 1 in 10,000. Similar findings were noticed by Terman for people with IQs over 170 as children.

10.4.5. *The Prospects for Individuals with Low Test Scores*

What about the other side of the coin, people in the low normal intelligence range, which I shall define as those whose IQs lie between 70 and 90?

People in the low intelligence range are not automatically candidates for assisted living or other institutional programs. However, more people in the low intelligence range are found in welfare and prison/jail

72 See Cronbach, 1996, and Robinson, 1996, for elaboration on these points.

73 See Muller et al., 2005, and Vasquez & Jones, 2006, for examples of such assertions.

populations than would be found if the welfare and prison/jail populations were selected randomly from the population. This does not mean that the majority of low intelligence individuals are headed for some form of institutional control. It simply means that they are at a greater risk of having these bad things happen than is the case for a randomly chosen member of the population.

As in the case of the gifted, the best way to understand the issues facing people in the low intelligence range is to examine prospective studies, in which individuals in the low intelligence range are first identified, and then, hopefully along with a control group of average and above-average individuals, are followed for some time. Here are three such studies.

In Herrnstein and Murray's "Bell Curve" study[74] of the NLSY79 panel, teenage men and women took the AFQT in their mid to late teens.[75] The Department of Defense uses these scores to classify people into categories I–V, with categories IV and V covering the range of scores below 90. About 45% of the NLSY79 men and women in categories IV and V failed to obtain either a high school diploma or a general education degree (GED). The base rate for the entire NLSY79 panel was 9%. Measures of workplace performance were similarly low. Herrnstein and Murray concluded from this that intelligence, as measured as a teenager, is a major predictor of income as a young adult. Other analyses of the same data[76] have agreed that intelligence is indeed a predictor of income, but that ethnic status, gender, and location in the country also have to be considered.

If we look at analyses of the intellectual demands of the workplace, this is not surprising. Look back to Figure 10.10, which plots wages against imputed intelligence requirements for over 800 jobs. Income is fairly flat over the wide range of occupations that have low imputed intelligence requirements. The figure also shows that there is a considerable spread between the mean and ninetieth percentile of earnings within occupations. To the extent that intelligence determines within-occupation earnings, as we have seen that it does (section 10.3), one would expect people with low scores to earn less. And they do.

The picture is somewhat bleaker with respect to welfare. In this case we have to make a distinction by ethnicity, for welfare rates vary markedly with ethnic status. Herrnstein and Murray found that the rate at which White women in the low intelligence range had received Aid for Dependent Children support was better than four times the rate in the NLSY79 sample as a whole.

A more detailed picture of what happens to the low intelligence group in the workplace can be obtained by examining a Department of Defense study of how low intelligence males fared in military service, in a setting where working conditions are more precisely defined and recorded than they are in the civilian workforce. The study, *Project 100,000*, conducted in the late 1960s at the behest of Secretary of Defense Robert McNamara, was motivated by an important policy issue.

The US military normally does not recruit Category V individuals, and is limited, by law, to recruiting a fixed percentage of Category IV soldiers. The argument for the policy is that it does not make sense to go to the added expense of training and supervising personnel who perform at a low level. On the other hand, the military forces have to have a certain number of recruits each year. If the category IV designation is a false indicator of military performance, excluding these men and women would amount to ignoring a potential recruiting population.

In Project 100,000 approximately 100,000 Category IV men (IQ range roughly 80–90) were enlisted outside of the normal channels. Their military careers were compared to those of a control group of enlisted servicemen who matched them in age and educational status prior to entry, and who had met normal recruitment standards. The control group underrepresented the higher levels of AFQT scores (I and II) compared to

74 Herrnstein & Murray, 1994.

75 See Chapter 2 for a discussion of the ASVAB and the AFQT.

76 See, e.g. the analyses in Devlin et al., 1998.

Table 10.8. Attrition rates during basic training (percentages) for Project 100,000 participants and for the service as a whole, broken down by military branch.

Branch of Service	*Project 100,000 Enlistees*	*Overall Service Rate*
Army	3.7	2.0
Marines	11.1	4.4
Navy	8.6	2.8
Air Force	9.2	3.0

Note: Data from Sticht et al., 1987. Figures are for the 1969–72 period, during the Vietnam War.

the percentages found in the general population, for there were no officers in the study.[77] In civilian terms, Project 100,000 examined the workplace performance of people involved in blue-collar and lower-level white-collar occupations, excluding managerial positions above the foreman level, and excluding professional occupations.

In all military services the first thing that enlistees do is to go through recruit training or, as it is known in the Navy and Marines, boot camp. The ostensible goal of recruit training is to inform the enlistees about service customs and to provide them with a taste of the life they can expect in the future. This taste (and the service life that follows) varies greatly depending upon the service; the Army and Marines envisage a different life for recruits than do the Navy and the Air Force, and the Navy and Air Force differ from each other. Boot camp has a second, less announced purpose. It serves as a screening device to see which enlistees have the adaptability to change from civilian to a more disciplined military life.

Table 10.8 shows the attrition rate from boot camp during the period of the Project 100,000 study, roughly 1969–71. Attrition rates were higher for Project 100,000 than for the control group in every service. Attrition rates also varied considerably across services. This may be because of both differential recruitment by the services and differences in basic training itself. Basic military training, strictly construed, requires roughly a fourth-grade reading ability. Because the Navy and the Air Force have more technical billets than the other services, higher reading and mathematics requirements may have been in force in those services. These requirements would be especially hard on the Project 100,000 servicemen, compared to members of the control group. By contrast, the Army and the Marines place more stress on determining a recruit's ability to follow instructions in physically demanding situations.

Looking at the other side of the coin, the vast majority of the Project 100,000 servicemen did complete basic training. From the viewpoint of the services, a rigid insistence on conventional standards would have excluded over 90,000 trainable enlistees. That would have been a serious loss.

Modern military services are a microcosm of society. They contain some positions, such as infantryman, that are clearly unique to warfare. They also contain many positions that either have exact civilian counterparts, such as clerk or electronic technician, or have close analogs in civilian life, especially outside of combat. We have already seen, from the analysis of civilian occupations in section 10.3, that occupations vary a good deal in their demands for intelligence. Where did the Project 100,000 servicemen wind up during their first term of service?

Figure 10.13 shows that Project 100,000 servicemen were much more likely to be assigned to nontechnical, nonspecialized jobs than were normally enlisted servicemen matched for pre-service education and ethnic status. However, the evidence that this is

77 Sticht et al., 1987.

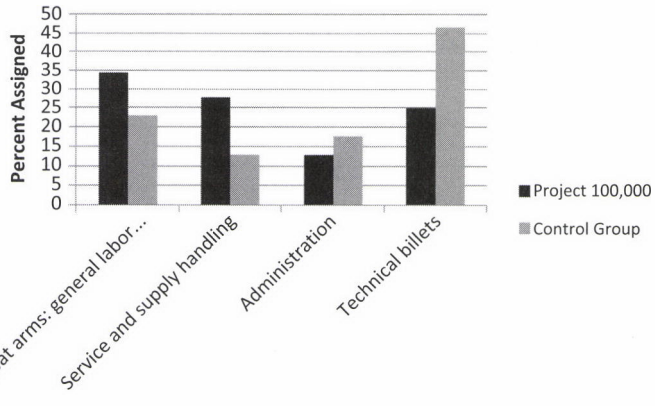

Figure 10.13. Assignment of Project 100,000 and a matched control group to different classes of military occupations during their first term of enlistment. Based on data reported by Sticht et al., 1987.

due to their demonstrated abilities (or lack of them) after testing is not as strong as it could be. Assignments to technical billets are made based partly upon observations and interviews during recruit training, and partly upon a serviceman's ASVAB scores, obtained prior to entry. This means that the low scores of the Project 100,000 servicemen may have influenced their assignment.

From the viewpoint of a scientific study, this was a flaw in the design. However, one can hardly fault the services for using the scores to make assignments in ways believed to minimize training costs.

As in the case of the data on attrition, one can see two different things in the data on occupations. On the one hand, it shows (not surprisingly) that low intelligence men tended not to be assigned to technical and administrative occupations. On the other, they filled roles that are vital to the military mission.

What happened after the servicemen completed recruit training? Figure 10.14 compares the career progress of Project 100,000 and control servicemen, in terms of rank attained through their first enlistment and, for those that stayed in the service, their status in 1983, which would have been from twelve to fourteen years after initial enlistment. The picture is similar to the earlier comparisons. The Project 100,000 servicemen were generally not "top of the line," rapidly promoted military personnel. On the other hand, they did show progress in their careers.

Project 100,000 was replicated . . . accidentally. During the 1980s a technical mistake was made in determining the norms for a revised version of the ASVAB. Before the mistake was discovered several thousand normally unqualified Category IV soldiers were enlisted into the Army. The careers of these soldiers have been followed, and by and large the results are the same as those obtained in the better-controlled Project 100,000 study.[78] There would be little point in repeating the statistics. However, I will recount a hopefully informative anecdote to enliven the statistics.

While serving as a consultant to a research project studying the accidentally recruited Category IV soldiers, I interviewed a lieutenant colonel who had commanded an armored battalion containing a high percentage of Category IV soldiers. The colonel believed that category IV soldiers performed well when doing clearly defined tasks, even if they were quite detailed. He offered as an example the task of replacing the power unit on a tank. This task is done in over a dozen well-defined steps, always done in the same sequence. After training, Category IV soldiers could replace the power unit.

78 Sticht et al., 1987, discusses the renorming problem and the resulting performance of the soldiers involved in some detail.

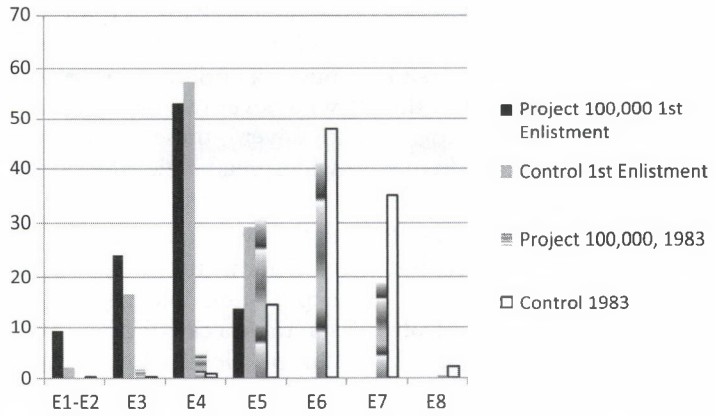

Figure 10.14. The advancement in rank of Project 100,000 personnel. The percentage of servicemen in the Project 100,000 and control groups who had reached different enlisted pay grades either during their first enlistment or after approximately thirteen years of service. Pay grades E1 and E2 were the grades typically assigned immediately upon completion of recruit service (e.g., private or private first class in the Army and Marines). Pay grade E4 (sergeant in the Army or Marines) was considered satisfactory service during the first enlistment. The higher pay grades represent staff noncommissioned officers (NCOs) or, in the Navy, Petty Officers. Grade E8, Master Sergeant or Chief Petty Officer, was a position with considerable prestige and responsibility.

More generally, the colonel believed that Category IV soldiers could be trained to do tasks where the instructions were "do this, then do that, then the other thing . . . "

The colonel thought that Category IV soldiers had trouble with tasks that are defined by the end to be accomplished, rather than by the steps to be taken. He offered as an example the task of recalibrating a gun sight that has been knocked out of alignment with the barrel of a tank's cannon. The instructions for this task began, "Find a way to fix the sight rigidly on the tank's hull." This instruction defines the goal to be accomplished and leaves it up to the soldier to find a way of accomplishing the goal.

If we place this in a more psychological framework – for the colonel certainly did not use such words – the Category IV soldier could learn what to do in a well-defined situation but had trouble anticipating what would happen if he took a particular sort of action in a more poorly defined situation. This sort of deficit appears to be characteristic of low intelligence people in schools, the military, and the civilian workplace. Depending upon whether you approach intelligence from the viewpoint of the psychometrician, the information-processing psychologist, or the cognitive neuroscientist, you can say that people with low intelligence have trouble with tasks that have a high demand for *g*, tasks that involve the working memory/executive function class of behaviors, or tasks that place demands on the forebrain–cingulate cortex circuit. All three statements amount to the same thing.

10.5. A Concluding Comment on Intelligence and the Workplace

Someone who says that intelligence, as assessed by standard cognitive tests, is irrelevant to performance in either academia or the workplace is simply wrong. So is someone who says that intelligence does not amount to very much, compared to a variety of personality characteristics. The data

both within and across occupations show that measures of intelligence are among the best, and most often *the* best, predictors of academic and occupational success. On the other hand, no one has claimed that cognitive tests are perfect predictors. Predictive validity correlations are in the .4–.6 range, which is much better prediction than can be achieved with any personality measure, but is still far from perfect.

The same message can be extracted from studies of extremes. On the whole, people who have high intelligence test scores do quite well. *On the average* the gifted get better jobs and make more money. The Terman study, and all studies of the gifted afterward, gave the lie to the stereotype of a gifted person as a neurotic introvert with health problems. The contrast between the SMPY 1 in 200 and 1 in 10,000 groups shows that the tests have predictive power at very high levels, directly contradicting statements to the contrary by psychologists who do not, themselves, study individual differences.

Here are some documented statistics, but on a very small sample, so the report falls somewhere between a scientific fact and an anecdote. In the 1960s the National Aeronautic and Space Agency conducted an intensive screening of volunteers to become the first American astronauts, the MERCURY and GEMENI programs. The selected candidates had WAIS IQs averaging 135. Unselected candidates had IQs averaging 131. A control group of comparably aged aviators averaged 118. Being intelligent is part of having the right stuff.[79]

At the opposite end of the distribution, people in the low intelligence range generally have difficulty with school, especially if they are placed on an academic track, and usually take occupations that do not make high cognitive demands. They earn less than people in the normal-high intelligence ranges and are more likely to require some form of welfare assistance.

It cannot be stressed too strongly, though, that these are trends. Every study of the extremes comes up with exceptions. There are people who do quite well although they had modest test scores, and there are stunning examples of people with high scores who never live up to their promise.

Given these facts, why is there a widespread belief that intelligence does not count for very much in life? I think, but cannot prove, that several factors are involved.

One factor is a failure of unrealistic expectation. We may expect the gifted person to be a casual genius who can solve difficult problems without much care or effort. That is not the case. Two of the personality characteristics of the gifted are that they generally enjoy their work, and that they work very hard at it. The brilliant genius who, without training, knows at a glance the answer to difficult problems in mathematics, physics, or what have you is a very rare bird.[80]

People without statistical training have a hard time grasping the concept of something that increases the probability of an event but does not establish its certainty. Thus if we can think of examples of people with high test scores doing stupid things, or people with low test scores doing good things, that is taken as proof that the predictors do not work. In the newspaper business "Man bites dog" is news, "Dog bites man" is not. The unusual is publicized and sticks in our minds. The prosaic does not.

We may have an unrealistic idea about the extent to which personal characteristics determine success. Large-scale social and economic forces, and idiosyncratic impersonal events, can play a great part. The various reports of the Terman group stress how much these highly intelligent people were influenced by having come of age during the Great Depression and then, especially for the men, having had to deal with World War II. The SMPY group has grown up in times of relative peace and economic expansion, at least until 2010. Such things

79 Santy, 1994, p. 276.

80 It is possible that a very few such individuals exist. The best-documented case study is of the Indian mathematician Ramanujam, who made major contributions to mathematics even though he was self-taught. He also spent a great deal of time working on mathematical problems.

influence one's success, quite outside of personal traits. While it is true that we partly make our own environments, we are partly stuck with them.

People are heavily influenced by their personal experiences. Charles Murray has pointed out that we live in a society that is sharply stratified by intelligence.[81] College-educated people, by and large, deal with other college-educated people, and people with high school educations deal with each other. Within the restricted range of intelligence that people can observe directly, other variables may account for more variation in performance than intelligence does. It is only when we step back and look at the big picture that the importance of intelligence becomes clear.

My final speculation is that some people, for understandable reasons, do not want to find that intelligence has an effect upon success.

Many people in post-industrial society, and especially in post-industrial American society, hold a belief and have two attitudes that, combined, provide a motivation for rejecting intelligence as a (partial) explanation of workplace success. The belief is that intelligence is something that is more or less fixed for life. As was discussed in Chapter 9, and will be discussed further in the section on aging in Chapter 11, this is a false belief – especially about intelligence in the conceptual sense of the problems that people can solve, rather than in the narrower sense of a test score. But that is not what people think. When one combines a belief in the permanence of intelligence with an attitude of distrust of elites, it becomes almost necessary to argue that intelligence is of little relevance in life, for to do otherwise can be seen as an affirmation of the appropriateness of rewarding an elite class of thinkers.

The second attitude is quite different. It has to do with a sincere desire for equal opportunity for all.

There are marked differences in the typical cognitive test scores obtained by members of different racial and ethnic groups. There are much smaller, rather complex differences in test scores between men and women. Up to the middle of the twentieth century these differences were used to justify varying degrees of segregation of minority groups, in areas including admission to universities and the granting of specialized degrees, and to place less strict, but still important, restrictions on women's opportunities. Since that time overt discrimination has virtually stopped, but there are people who have used group differences in test scores as evidence for the proposition that group differences in educational and professional achievement are largely due to group differences in intelligence.[82] To the extent that this conclusion is correct, provision of equal opportunity for all groups in society will not produce an equal distribution of social and economic rewards across groups.

Many people who are deeply committed to social equality find such a conclusion offensive. It is difficult for them to argue about the fact of differential distribution of test scores. Therefore, they deny the relevance of scores to social outcomes. This denial cannot be maintained.

The discussion of racial and ethnic differences in intelligence, and male:female differences in intelligence, raises extremely complex issues. We take them up in the next chapter.

81 Murray, 2008.

82 See, for instance, Murray, 2005; Rushton & Jensen, 2005; and Lynn and Vanhanen, 2002, 2006.