Journal of Vocational Behavior 29, 421–437 (1986)

Commentary

LLOYD G. HUMPHREYS

University of Illinois, Urbana–Champaign

I start my discussion with a brief description of my background and training and experience that is relevant to the topic of this commentary. This is followed by a discussion of the major issues raised by these papers. In my discussion I do not refer to *g*, but to general intelligence or the general factor. Spearman equated *g* with mental energy. Most psychologists following Spearman have discussed *g* as an entity, committing over and over again the error of reification without justification. © 1986 Academic Press, Inc.

I received my Ph.D. in the year that Thurstone's Primary Mental Abilities monograph (1938) was published. Three and a half years later I was commissioned in the Army Air Forces and assigned to the research unit of the Aviation Psychology Program commanded by J. P. Guilford. We were quickly able to obtain a substantial degree of differential validity of the Aircrew Classification Battery for pilots and navigators. Our tests were described in Thurstonian terms and were given stable differential weights for purposes of classification.

When I returned to the Air Force early in 1951 to direct the Personnel Research Laboratory, I was committed to the goal of qualifying nearly 100% of enlistees for at least one Air Force specialty at an above average level. This could be done presumably by capitalizing on the primary mental abilities of the recruits and would require nothing more than being clever and competent in constructing factor-pure tests. I was quickly disillusioned as we accumulated predictive validities and factor analyses of dozens of tests and thousands of examinees. We were able to define dependably many group factors, but found little differential validity for the multiple factors. (Differential validity is defined as stable differential regression weights for the prediction of multiple criteria.) I have quoted this experience in a number of publications without tabular documentation. Contrary to the practice of Joseph McCarthy, with whom I have been jocularly compared on one occasion, documentation was always available even though it was not in my personal possession.

Requests for reprints should be addressed to Dr. Lloyd G. Humphreys, Department of Psychology, University of Illinois, 603 East Daniel, Champaign, IL 61820.

0001-8791/86 $3.00

A model that described this experience soon became available. As a result of a debate with Ledyard Tucker on the respective virtues of orthogonal versus oblique rotations in which I defended the orthogonal approach that we had used in World War II, I became interested in the intercorrelations of factors. (My point had been that orthogonal and oblique factors were highly similar and that nothing was gained unless one interpreted the factor intercorrelations.) Thurstone (1947) had discussed factoring in orders beyond the first, but such factors were not readily interpretable. Factors defined by other factors do not seem as real as factors defined by measuring instruments. Then John Schmid and John Leiman (1957) of our laboratory developed a methodology that allowed transformation of oblique factors in two or more orders into a single orthogonal order that was appropriately termed hierarchical. Factors derived from all orders are defined by the original measures. Thus the English tradition stemming from Spearman through Burt to Vernon (1950) was integrated with the American multiple factor approach associated with Thurstone.

The hierarchical model does not force a general factor, but it does allow one to be defined when required by the data. The model also allows for factors of varying breadths beneath a general factor. Vernon's verbal–numerical–educational and mechanical–spatial–practical major group factors are broader than Thurstone's primaries, which in turn are broader than Guilford's (1985) structure-of-intellect factors. The near absence of differential regression weights in our data was alone sufficient evidence for the importance of the general factor, but consistent evidence for a limited degree of differentiation of Vernon's two major group factors was also obtained. Both Thorndike (1986) and Hunter (1986) describe this phenomenon, but it will be found more consistently in good data than Hunter believes and is more important than Thorndike believes.

THE GENERAL FACTOR

How General

A necessary and sufficient condition for a general factor to be defined by all of the measures whose intercorrelations form an R matrix is the absence of zero and negative correlations in that matrix. Tasks, whether called predictors or criteria, with appreciable cognitive content administered to a large sample of examinees from a population representing a wide range of talent are consistently positively correlated. One has to look long and hard to find exceptions. When exceptions are found, they have minimal cognitive content. Thus the number of attempts on highly speeded tests composed of very easy items does show small negative correlations with certain cognitive tests in Project Talent (Flanagan et al., 1964), but a formula score that weights wrong answers typically restores the positive manifold. Other tests in which cognitive content is low and group differences large and in opposite directions, such as hunting information and information about esoteric foods for males and females, may be negatively correlated in the general population but positively correlated within the subgroups composing the total population.

A highly replicable mathematical dimension that can be defined under specified conditions (wide range of talent and cognitive content of the measures as evaluated by consensus among research workers) is real. Simple reaction time has small positive correlations with cognitive measures, but so do measures of physique. Height in a reasonably homogeneous population of Swedish conscripts has a correlation of .22 with a measure of intelligence (Husen, 1959). This correlation is considerably higher (.40) in prepubescent girls (Humphreys, Davey, & Park, 1985). Correlations with variables such as these do not make the general factor more real, but they do make it more general.

Given the correlations involving simple reaction time and physique with intelligence, it would be surprising if one did not find positive correlations involving various electrical potential recordings from the brain and intelligence. After an initial flurry of reports of large correlations, those with the repetitive potentials associated with ongoing states of the subject are now considered to be nonzero but small (Callaway, 1975). Interest has turned to the potentials evoked by specific stimuli, but the research is typically marred by small *N*s and post hoc selection among multiple dependent variables. The only reasonably acceptable data cited by Jensen (1986b) are those of the Hendricksons (1982) in which sample size was 219. Correlations were large with individual Wechsler subtests and the correlation with total IQ was .83. Nevertheless, I shall venture to predict that a consensus in 10 years will yield correlations closer to the small values for reaction time or adult height than to .80.

Neither is a high degree of heritability required for the reality of the general factor. A phenotypic trait with zero heritability can be real, that is, consistently observed. Phenotypic traits must be observable and measurable. The general factor of intelligence can be observed subjectively by friends, relatives, supervisors, etc., and reported in the form of ratings. A consensus from several judges, all of whom have had ample opportunity to observe, especially if they have been instructed concerning the kinds of behavior that best represent the construct, is a good measure of general intelligence. A good test, in contrast, is more economical and provides a zero and units of measurement that are comparable across conditions. Note that a putative unitary capacity or entity underlying the behavior measured is not a phenotypic trait. In an important sense the genotype does not represent a unitary capacity or entity. Not only is the inheritance of intelligence polygenic, but the many genes are involved in the development of many neural structures and many hormones.

There are also many ways of obtaining good measures of the general factor. If one can find a sufficient number of measures each with a moderate loading on the general factor but sufficiently diverse among themselves, the total score can be as highly correlated with the score on a standard test of intelligence as two such tests are with each other. Thus, a composite of academic achievement tests administered to students whose language at home is English (including black English) and who have consistently attended American schools is a good measure of general intelligence.

A more dramatic example is furnished by Roznowski (1986). She constructed two tests composed largely of subtests of narrow information. One was designed to produce a large sex difference favoring males, the other females. Correlations with a criterion of intelligence were ignored at this stage. The two composites were then correlated with the Project Talent intelligence composite. The male superiority composite had a slightly higher correlation with the criterion in the female than in the male sample, but both were greater than .80. The same correlations for the female composite were originally less than .80, but were attenuated by the presence of number-right scores on three highly speeded tests with very easy items. Omission of the three invalid components produced correlations greater than .80 again. In this case the correlation for females was slightly higher than the one for males. The correlation between the two composites is about .70, slightly lower for males, slightly higher for females. There is a significant gain in effectiveness as a measure of general intelligence when the two composites are combined, and the sex difference becomes trivial in size.

In the hierarchical model the two major group factors of Vernon can only be defined by controlling the variance of the general factor. If separate measures are constructed without this control, the resulting tests are good measures of the general factor that also emphasize content important in its own right. Now, if a better measure of the general factor is required, the sum of the two measures of the major group factors will furnish the desired distribution of scores.

Importance. The importance of the general factor depends only in part on its correlates with social criteria, but those correlates are the obvious ones with which to start. The correlations with education and occupation are so well-known, and have been discussed by Gottfredson (1986), that I shall concentrate on why they are not higher.

Amount of education and intelligence are reported by Gottfredson (1986) to be correlated to the extent of .60. In a meritocratic society that correlation should be higher. The explanation for its modest size starts with the wide ranges of mean levels of knowledge and skills that enroll in our institutions of secondary and postsecondary education and in the equally wide range among the means of their graduates in knowledge and skills. There is an institution for almost everyone. Coupled with this institutional variability is the pattern of college going as a function of both race and socioeconomic status. Socioeconomic status without regard to race is associated more highly with college going than its importance as a predictor of success in college merits. This is due in part to the requirement of financial support for college attendance, but it also arises in part from attitudes of families and friends toward the importance of college. In the Project Talent stratified random sample of the nation's high schools the correlation between the rate of college going in schools and the mean levels of achievement on the Talent intelligence composite was only a little larger than .40.

It is also true that blacks obtain more years of education than their measured academic talent at the beginning of either high school or college would indicate. They also obtain more diplomas and degrees than their levels of achievement at the time the credential is awarded would indicate. Largely segregated education at all levels brought this about, and it continues today with only a little change.

There are also ready explanations for the small amount that measured intelligence adds to amount of education in occupational placement in the economic and sociological literature. This is due to the undue weight our society gives to credentials, which stems in turn from the widespread American belief that education can accomplish almost anything. Because credentials are easier to acquire than generally acknowledged, making occupational entry, or selection and promotion, dependent on credentials attenuates the importance of general intelligence in occupational placement. During World War II we routinely obtained multiple regression weights for amount of education in our aircrew research. The standardized weight was consistently trivial in size. Later we followed the same practice in airman classification research with the same result. Appropriate tests allow substantially more accurate prediction of job performance than does amount of education.

A more subtle contributor to the absence of a significant path from general intelligence to occupational placement arises from a problem associated with the common belief that intelligence is assumed to be stable during all but infant development. Research workers use intelligence test scores from school or clinic records that were obtained at any time during development, but intelligence at age 6 is a highly fallible predictor of intelligence at 18. When intelligence and amount of education are compared in their contributions to occupational placement, intelligence should be measured near the time of entry.

General intelligence is important not because it is a capacity that is fixed and unchanging, but because it is relatively stable and because we know very little about how to produce change. The intercorrelations of intelligence measured on multiple occasions during development take the form of a quasi-simplex matrix (Humphreys, 1985). Between adjacent occasions correlations are high, but they shrink in size with the amount of elapsed time between occasions. The observed correlations can be fit by a model that requires a large but imperfect true score correlation from year to year and independence of true score gain and true score base at the beginning of the year. This hardly describes a fixed capacity, potential, or unitary mental power, but it does describe the high degree of stability over shorter intervals of time and the much lower stability over longer intervals. The change in individual differences relative to the age group occurs without explicit intervention, and no dependable information is available concerning effects of explicit intervention.

It is entirely possible that ways could be found to increase the stability of individual differences on the general factor or, on the other hand, to decrease the stability. Neither of these outcomes is socially desirable as such. The goal of educational intervention is to develop the intelligence of all children, with any change in the stability of individual differences over time being completely secondary.

We do know that there was a large gain in the mean scores of 18-year-old American males on the general factor between the two world wars (Tuddenhan, 1948). Presumably there was little change in genotype, but the gain in level on the phenotype was associated with an increase in the ease of acquisition and accurate utilization of military knowledge and skills. The most attractive explanation for the gain is the increase in the amount of education made available to young people during the 24-year interval between the wars. It seems reasonable that changes in individuals will also occur as a function of staying in or dropping out of educational programs. Perhaps there is a reciprocal relation between intelligence and the amount of education. The effect of the latter on the former is trivial over short intervals of time, but may well have a measurable effect over long intervals.

Limitations. One limitation on the overriding importance of the general factor in prediction has already been mentioned. One can find evidence for differential validity of measures of the broad major group factors in a wide range of talent among persons with little differential formal training. Military enlistees are prime examples. In a wide range of talent the incremental validity is not large, but it is replicable. The utility of small increments in validity for tests that are widely used, whether for selection or for guidance, can be large.

The preceding generalization cannot be rejected on the basis of any present meta-analyses. This methodological tool of Hunter (1986) and his co-workers is important and necessary as long as correlations are computed on small samples of examinees, but it is still a relatively crude technique. There is a continuing problem of apples and oranges in all meta-analyses. In prediction studies the problems relate to populations sampled, tests used, and criteria available. Given the heterogeneity among the many studies to be aggregated, corrections for measurement error and restriction of range of talent are rough estimates at best.

Military data are better in several ways. Samples tend to be larger, predictor tests are constant over many criteria, and the standardization of technical training and grading ensures reasonable comparability of criteria over multiple instructors. Another source of more dependable data is subsequent occupational placement of the members of nationwide samples of students measured early in their educational careers. Of course, differences among occupational groups on tests are not necessarily related to the predictive validities of the same tests within groups. A relationship is more likely, however, for measures of abilities as distinguished from interests. If the tests are not ordinarily used as hurdles during educational preparation for the occupation, it is not reasonable to believe that large group differences would occur unless the test was related to early training success. In the sample of 10th-grade students in Project Talent whom I have studied, there is not the slightest doubt that the first discriminant function among occupational groups identified 13 years following test administration would involve level on the general factor. There is also not the slightest doubt that dependable differences exist for a second discriminant function defined by the two major group factors (Humphreys, Davey, & Kashima, in press). We were surprised to find that physics majors were almost as high on mechanical–spatial–practical ability as on the general factor. They differed substantially from attorneys in their level on the group factor, but not in general intelligence.

A second limitation is imposed by restriction of range of talent. As students move up the educational ladder, restriction on the general factor occurs at each rung. Thorndike (1986) discussed the importance of this parameter with respect to the experimental group of pilot trainees, but did not provide the full story. The experimental group was more variable than selected trainees, but it also had a mean of 113 and a standard deviation of 14 on the Army General Classification Test in place of the 100 and 20 expected in the unselected population. Although more variable than the cadets who followed the usual selection and classification procedures, the experimental group was highly self-selected. In the full range of World War II talent a measure of the general factor would have been even more predictive of success in pilot training.

Later in our enlisted research we equated the officer quality stanine (basically an intelligence composite) with the Armed Forces Qualification Test. This research documented dramatically the amount of restriction of range in the officer population. A stanine of 1, representing the lowest 4% of officer applicants in the standardization sample, was the equivalent of a percentile in the low 70s on the AFQT. The range of stanines from either 3 or 4 was included in the highest 10% of the enlisted population.

In retrospect we were quickly able to obtain differential validity for pilots and navigators because the restriction of range on the general factor made other differences relatively more important. As a matter of fact, we were able to use negative weights for certain kinds of information in predicting success in pilot training, thereby reducing the overlap between the pilot and navigator stanines. Information about pilots, planes, engines, cars, motorcycles, etc., was highly predictive of success in pilot training and was, of course, moderately loaded on the general factor. When information about art, music, dance, and literature was added with small negative weight, validity and differential validity were increased, and the correlation with the general factor was reduced.

A third limitation on the sufficiency of prediction based on the general factor is the importance of acquired information and skills that are relevant to training or the job. Differential prediction is increased not only by restriction of range in general ability among college students, but also by specialized training. Differences of these kinds at a given level of general intelligence do have predictive significance. An example is provided by college students, for whom the advanced tests of the Graduate Record Examination typically have higher validities than the Verbal or Quantitative tests in predicting graduate grades.

The advantage provided by transfer of training may frequently be somewhat limited in time, but this limitation does not always occur. The general factor does not seem to become more prominent later in the flying officer's performance as a pilot. We validated tests administered in 1942–1945 against the criterion of becoming a jet ace later during the Korean War. The information test used for initial classification had the highest correlation with the criterion. This test was one of 20-odd tests that included measures of reading comprehension and arithmetic reasoning, both highly correlated with the general factor. John Flanagan (personal communication, 1986) reports that a more advanced information test was the single best predictor of success in the postwar training program of United Air Lines. All candidates were experienced pilots, with most of them being veterans of the Army Air Forces of World War II. As noted earlier, they were highly selected on the general factor.

Any evaluation of past experience should not depend on self-reports of experience. Tests of knowledge and skills are more valid. A recent report of selection research in the Federal Aviation Administration contains a useful example. The agency had used a self-report of experience in military flight controlling and in military and civilian flying in its selection program over a number of years. When an appropriate information test was constructed and administered, there was a substantial gain in validity of traditional measures of aptitude in comparison to the evaluation of reported experience (Dailey & Pickrel, 1984). There was also an increment in the validity of the selection battery in spite of the increased overlap between tested information and the other selection tests.

The obvious moral is not to become too doctrinaire about the general factor in selection. Hunter and his principal collaborator, Schmidt, have provided an extremely useful service in destroying the myth that selection tests are sensitive to every little nuance in test content, training, jobs, populations of people, etc., but a new myth should not be established in place of the old one. Hawk's (1986) paper suggests that the United States Employment Service has accepted the new doctrine too uncritically. I recommend that they remain more open and be aware of the parameters that affect the possibilities for differential prediction.

It is highly relevant to the issues facing the USES that their test battery was constructed on the belief that there were large differences between so-called aptitude and achievement tests. Their battery includes no measures of information. It seems highly probable that job-relevant information acquired by a person on his/her own, including the sources of that information, such as the acquisition of information about flying by prospective pilots, functions differently than job information acquired on a job. It is suggestive along these lines that the most valid information for pilot success was also current information. The most successful pilots did *not* know the names of the men who were first to fly across the English channel. Their information about flying was not garnered from history books.

One is also entitled to be somewhat skeptical of the very high correlations reported by Hunter (1986) for the relationship between job knowledge and objective, hands-on measures of job performance. I have tried to visualize, for example, the nature of the hands-on measure of cooking that he refers to. It was probably not objective. It is very desirable to have a measure of the number of production units per unit of time, but such criteria are not generally available. Neither are high correlations with hands-on measures inevitable. One can imagine designing a test course that would allow a reasonably objective measure of driving skill, but I would guess that the correlation with information about driving would be moderate. There are many problems in the design of hands-on measures of proficiency of which Hunter does not seem to be aware. If he were to undertake a hands-on job of constructing hands-on tests in industry or in the military, he would find it both enlightening and humbling.

CROUSE'S PROPOSAL

The proposal that Crouse described and Gottfredson now places in perspective (Gottfredson & Crouse, 1986) has a good deal of merit and a trivial fault. It is entirely compatible with my views of long standing concerning the desirability of achievement tests for selection purposes.

In the early 1950s I was largely responsible for inserting mechanical information into the AFQT along with verbal, arithmetical, and spatial subtests. I considered then, and still do, that there were no important differences between traditional components of intelligence tests and other ways of measuring the general factor. The addition of mechanical information provided a small edge in the selection of military personnel qualified in areas in which the demand was greater than the supply. The content of the AFQT subsequently remained constant for a good many years, but was ultimately changed by the deletion not only of mechanical information but of spatial visualization as well. I know neither the stated nor unstated reasons for the deletions, but I suspect that the decision was not based on good predictive data. There is generally too much a priori thinking about the content of an intelligence test and the supposed gulf between intelligence and achievement. Neither is the use of "aptitude" tests of verbal comprehension and quantitative reasoning for college admission adequate, but it is more appropriate in the SAT than in the AFQT. College courses, after all, are more verbal and quantitative than are military technical training courses.

In the 1960s, and back in a university, I wrote to the Graduate Record Office of ETS asking if our applicants in psychology could take three advanced tests in place of the two "aptitude" and one advanced test that were then and still are the only possible combination. I quickly received a summary rejection. Three advanced tests would, I am confident, provide all of the useful information now provided by V and Q. This would still occur if applicants were allowed to elect the three tests. Such a program would require somewhat more accurate information about the equivalence of scores across tests than is now available, but the gains in other information would outweigh the disadvantage of imprecise equivalents.

At the time of college entrance one might wish to have five achievement tests administered, of which two would be common for all applicants (English and mathematics) and three would be elected. An aggregate score of the five would be a very good estimate of the student's standing on the general factor, but differential information would also be possible. The choice among possible electives should alone furnish significant differential information.

The conclusion in Gottfredson & Crouse (1986) is correct in that the Scholastic Aptitude V and M generally furnish little information to improve prediction over and beyond high school rank in class. For high schools that do not fit the average pattern of quality of intake and quality of graduate, however, the test information is crucial. A university high school, for example, tends to be highly selective. Rank in class is highly ambiguous for low-ranking students. Mean grades are little better because teachers tend to give a range of grades no matter how much information they have about their students. By the same line of reasoning, rank in class as well as mean grades for high-ranking students in black inner city high schools are highly ambiguous. When the average high school graduate has levels of reading and aural comprehension of the English language at the ninth-grade level of white norms, only test scores can resolve doubts about the meaning of the high school record.

The preceding problems are met as well or better technically by a requirement for five achievement tests, and there is a pronounced advantage of the change in terms of public understanding. The advantages in terms of students' motivation to study and learn that Crouse (Gottfredson & Crouse, 1986) adduces may also be important. He is not alarmed and neither am I by the common objection that the tests would determine the curriculum. If we start with good tests, any change in the curriculum would be an improvement. Also important is the measurement of change in academic quality over time. If scores improved from 1917 to 1942, they can also decline. We have seen a decline in recent years, and it is noteworthy that the grades awarded by high school teachers increased year by year during the same time period.

RACE DIFFERENCES

The Spearman Hypothesis

Jensen's (1986b) fixation on the Spearman hypothesis is based on inadequate data and is also counterproductive. My argument starts with important qualifications that he overlooks. In the first place, the aggregate correlation he reports for the relationship between the size of the race difference and the size of the general factor loading on samples of tests is quite modest—about .5–.6 (Jensen, 1985). Squaring this correlation, it means that there is only about 30% of common variance. In contrast, for the same variables on a much larger sample of tests the common variance for low and high socioeconomic groups in the white population is much greater (Humphreys, 1986a, 1986b). Second, when a black group and the previous low SES white group are compared, the common variance is almost zero. Apparently the principal basis for the correlation reported by Jensen is not race but socioeconomic status.

His fixation is counterproductive because it tends to direct attention to a genetic basis for the race difference. Jensen's estimate of heritability of individual differences is on the high side in the literature. Even with high estimates there is room for substantial environmental effects, but both critics and supporters neglect this. His fixation also results in underestimating the problem of race differences. Differences are even broader than loadings on the general factor suggest. Because they are broader, they are ipso facto more important. The black deficit in mechanical comprehension, for example, is much larger than loadings of mechanical tests on the general factor would indicate. Not only is mechanical comprehension in short supply in the military services, but it is in short supply in the civilian job market as well. High-tech societies need people who can operate, troubleshoot, and repair complex equipment.

By looking at the content of the tests in the high–low and low–high quadrants of the plots that relates the total sample of whites with blacks and the low SES white sample with blacks, it should be possible to create hypotheses about causation. In the complete racial comparison the size of the race difference on some tests is considerably smaller than the expectation based on the general factor loading, just as mechanical tests have the opposite pattern.

An earlier article (Humphreys, Fleishman, & Lin, 1977) along the same lines reached the same conclusions as the more recent ones, but we used a methodology that was superficially very different. Jensen (1986b) has been familiar with the earlier research for 10 years. He must criticize more to the point than he has to date (Jensen, 1986a) or report that the Spearman hypothesis has been modified.

Importance of the Race Difference

I welcome Gottfredson's evaluation of the black–white difference on the general factor. If anything she has been too restrained. Social scientists have quite generally been delinquent in their roles as scientists in their treatment of the problem. They have ignored the many dependable correlates of the deficits, preferring to debate causation. They have suggested excuses in place of documented reasons. They have recommended policies for which they had no support other than wishful thinking. In total they have acted like the three little monkeys who shielded their eyes, ears, and lips. Politicians, editors, and writers have followed the same examples.

The cognitive deficits that are measured do not exist in isolation. They are a part of a complex that includes the development of the underclass, teen pregnancy, female-headed families, crime, drugs, and AFDC. Perhaps the most constructive social action that has been taken in recent years was to liberalize policy on abortion, but a near majority of the American population with the strong dogmatic support of several religious denominations wishes to reverse abortion policy. For example, the number of children per family in AFDC households has decreased markedly since the early 1970s. In short, I believe that we have a social problem that is at least the equivalent of the AIDS epidemic, but there is a good deal more constructive action being taken about the latter.

Causation versus Importance

For the sake of argument let us suppose that individual differences and race differences in intelligence are largely if not entirely environmental in origin. These assumptions, incidentally, are quite commonly believed to be true by social scientists, politicians, and opinion leaders, but many of these environmentalists are really closet hereditarians. They assume that there is a natural biological capacity not measured by intelligence tests that only needs to be unleashed by opportunity in order to become quickly manifest. I have called this the water-under-pressure theory of human ability. If water in the system is under pressure, one needs only to find the tap and water immediately gushes forth. Instead evidence points to the slow development of abilities over time and with effort.

The construct of general intelligence, whatever the genetic and environmental influences may be, is basically a mathematical dimension. Even a high heritability coefficient, or correlations with reaction time, brain waves, and physique, cannot establish intelligence as a fixed capacity or any other kind of entity. The construct is behavioral and has associated functional correlates that alone support the personal and social importance of existing individual and group differences. A mathematical dimension that has many important correlates, that is relatively stable, especially late in development, and that we do not know how to modify, can stand on its own.

Although a broad set of achievement tests measures this mathematical dimension well under most circumstances, public understanding and acceptance would be advanced if we stopped using the terms intelligence and general factor. It would also be helpful to translate mean IQs and the standard deviations about the means into mental age or grade equivalent units. I say this in spite of the technical defects in those scales, because age and grade units make the problem more concrete and the defects in the scales produce trivial inaccuracies in the translation.

At the time of high school graduation or age 18 years, the average black student is at about the 9th-grade or 15-year level in aural comprehension of English, visual comprehension of English, writing skills, arithmetic, elementary mathematics, information about physical and biological science, information about social studies, and information about the industrial arts. Standard deviations of blacks and whites are about the same so that only 15–20% of blacks are at or above 12th-grade or 18-year norms. By the same token 15–20% of black students are at or below 6th-grade or 12-year norms. These differences are not artifacts. When blacks and whites compete in cognitive learning situations and are judged by the same standard, blacks perform at best at the level predicted by their test scores.

One really has to believe in the water-under-pressure theory of ability to expect anything approaching proportionate representation of blacks in higher education, the professions, or skilled or semiskilled occupations in the light of these data. Unfortunately, the evidence available to date indicates that acceptance of affirmative action goals or use of affirmative action quotas does not reduce the deficits. The water does not gush forth.

Perhaps the most extreme example of wishful thinking as a solution to a problem is the belief that an athlete whose level of aural and visual comprehension of English is at the grade school level can succeed in largely white universities. Even if such disabilities are acquired, they are overcome slowly if at all by students who have spent 12 years in American schools and are now 18 or older.

Stability of Race Differences

Although a statement that there has been no appreciable change in the size of the race difference in intelligence since World War II is largely true, it does overlook the data gathered by the National Assessment of Educational Progress (1985). Reading comprehension of 9-, 13-, and 17-year-old students has been assessed four times since 1971 at approximately 4-year intervals. Tests of reading comprehension have high loadings on the general factor, so that there is presumptive evidence for a gain on the general factor when there is a gain in reading.

There are numerous publications available that describe assessment results. It would seem logical to refer readers to the most recent one (National Assessment of Educational Progress, 1985), but there is a problem in evaluating the information in that report. Previous reports included only means of correct responses to common items. The most recent report used a new metric based on item response theory methodology. This change, although theoretically highly desirable, produced some anomalies. For example, the race difference for 9-year-olds during the first cycle is reported to be significantly smaller in the same birth cohort 4 years later. At 17, however, the race difference in the same birth cohort has increased to the same size as the earliest one. Reading comprehension is too robust a skill for the means of large, random samples of children born and reared during the same time period to change in this fashion. My account of the data attempts to integrate the information from both metrics and may well be considered a "broad brush" treatment.

Between the first and second cycles of assessment, black 9-year-olds made a substantial gain in reading and partially closed the gap with their white counterparts. There was little change this time for 13- and 17-year-olds. Between the second and third cycles, 9-year-old blacks made another gain, and 13-year-olds made a small gain as well. For the second time, however, the oldest blacks failed to gain appreciably. Finally, between the third and fourth cycles the 13- and 17-year-olds gained. During this period the 9-year-olds lost slightly, but by an insignificant amount.

Changes occurred first in the youngest cohort and moved through the age groups as that cohort matured, but the youngest blacks in the assessment have now stopped gaining. No one knows how the gains occurred nor why they stopped. There are two popular explanations for the gains:

racial integration, supported indirectly by the fact that the largest gains have been in the southeast, and the Title I remedial educational programs. There is only one popular explanation for the recent failure to gain: the budget cuts of the Reagan administration. Less popular but a likely candidate is the increasing demoralization of the black family. Considering the importance of the problem, it is shameful that we do not know achievement levels of black children in knowledge and skills related to later reading comprehension at the time they enter school. The black 9-year-olds who showed the substantial gain in 1975 may have scored higher on a test of general intelligence at the time they started the first grade. Whatever the genetic contribution to such scores may be, an intelligence test can legitimately be considered an achievement test covering important knowledge and skills acquired in the home and neighborhood.

In conclusion, it is much too early to evaluate the black deficit on the general factor as fixed. Race differences are no more fixed than are individual IQs. On the other hand, recognition that a major problem exists is coming much too late if it is coming at all. That recognition should have come at the time that civil rights legislation was being passed. Data available at that time were sufficient to define the problem.

Merit versus Equity

The problem of balancing merit and equity as described by Gottfredson is indeed complex and without an easy solution. As a matter of fact it is more complex and less tractable than she believes. An analysis made by Jencks et al. (1972) describes these issues. His data appear to be as sound today, perhaps even firmer, as they were 15–20 years ago.

I shall paraphrase his generalizations. If our society made entrance into higher education and occupations high on the prestige scale depend solely on academic grades and test scores, the present advantage of the children of middle-class whites over those of working-class whites would be reduced by about a third. Thus dependence on measures of individual merit would, by democratic standards, increase equity in the majority segment of our society. The children of middle-class whites profit from our present system that gives high weight to educational credentials. Unfortunately, dependence on grades and tests also hurts black children. The latter also profit from dependence on credentials, even though this dependence does not produce proportionate representation in occupations.

A question naturally arises about the motives of white middle-class liberals who criticize the use of tests in education. For reasons discussed earlier, test scores have a more limited impact on occupational entrance, but the impact is real. It is surely not coincidental that the children of middle-class parents are, on average, less intelligent than their parents. The difference is not a trivial one being, in standard score units, six-tenths of the distance back to the population mean. Middle-class children frequently have to attend less prestigious institutions of higher education. The typical background of students in foreign medical schools is a dramatic case in point. Middle-class parents have a selfish interest in minimizing the use of measures of the general factor and in dismissing the importance of its correlates.

REFERENCES

Callaway, E. (1975). *Brain electrical potentials and individual differences*. New York/London: Grune & Stratton.

Dailey, J. T., & Pickrel, E. W. (1984). Development of the air traffic controller occupational knowledge test. In S. B. Sells, J. T. Dailey, & E. W. Pickrel (Eds.), *Selection of air traffic controllers*. Washington, DC: National Technical Information Service.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., & Neyman, C. A., Jr. (1964). *The American high school student*. Pittsburgh: University of Pittsburg.

Gottfredson, L. S. (1986). Societal consequences of the *g* factor in employment. *Journal of Vocational Behavior*, **29**, 379–400.

Gottfredson, L. S., & Crouse, J. (1986). Validity versus utility of mental tests: Example of the SAT. *Journal of Vocational Behavior*, **29**, 363–378.

Guilford, J. P. (1985). The structure-of-intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence*. New York: Wiley.

Hawk, J. (1986). Real world implications of *g*. *Journal of Vocational Behavior*, **29**, 411–414.

Hendrickson, D. E., & Hendrickson, A. E. (1980). The biological basis for individual differences in intelligence. *Personality and Individual Differences*, **1**, 3–33.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence*. New York: Wiley.

Humphreys, L. G. (1986a). Race differences and the Spearman hypothesis. *Intelligence*, **9**, 275–283.

Humphreys, L. G. (1986b). Attenuated hypothesis or attenuated test of hypothesis? *Intelligence*, **9**, 291–295.

Humphreys, L. G., Davey, T. C., & Kashima, E. S. (in press). Experimental measures of cognitive privilege/deprivation and some of their correlates. *Intelligence*.

Humphreys, L. G., Davey, T. C., & Park, R. K. (1985). Longitudinal correlation analysis of standing height and intelligence. *Child Development*, **56**, 1465–1478.

Humphreys, L. G., Fleishman, A., & Lin, Pang-chi (1977). Causes of racial and socioeconomic differences in cognitive tests. *Journal of Research in Personality*, **11**, 191–208.

Hunter, J. E. (1986). Cognitive Ability, Cognitive Aptitudes, Job Knowledge, and Job Performance. *Journal of Vocational Behavior*, **29**, 340–362.

Husen, T. (1959). *Psychological twin research*. New York: Free Press.

Jencks, C., Smith, M., Acland, H., Bane, M., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Harper & Row.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences*, **8**, 193–264.

Jensen, A. R. (1986a). Humphreys' attenuated test of Spearman's hypothesis. *Intelligence*, **9**, 285–290.

Jensen, A. R. (1986b). g: Artifact or reality? *Journal of Vocational Behavior*, **29**, 301–331.

National Assessment of Educational Progress (1985). *The reading report card*. Princeton, NJ: Educational Testing Service.

Roznowski, M. (1986). *The use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias*. Manuscript submitted for publication.

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

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior*, **29**, 332–339.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: Univ. of Chicago Press.

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

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist*, **3**, 54–56.

Vernon, P. E. (1950). *The structure of human abilities*. New York: Wiley.

Received: August 11, 1986.