10/26/88
Draft--Not for Quotation

THE ECONOMICS OF EMPLOYMENT TESTING

John Bishop
Cornell University
Working Paper # 88-14

Center for Advanced Human Resource Studies
New York State School of Industrial and Labor Relations
Cornell University
Ithaca, New York 14851-0925
607-255-2742

This is a draft of a paper that is to appear in Testing and Public Policy edited by Bernard Gifford, Chair of the Commission on Testing and Dean of the Graduate school of Education at the University of California-Berkeley. The research that has culminated in this paper was sponsored by the Center for Advanced Human Resource Studies, the National Center for Research in Vocational Education and the Commission on Testing and Public Policy. I would like to thank Peter Mueser for helpful comments on earlier versions of the paper. The opinions and conclusions expressed herein are solely those of the author and should not be construed as representing the opinions or policies of any agency of the United States Government. This paper has not undergone formal review or approval of the faculty of the ILR school. It is intended to make results of Center research available to others interested in human resource management in preliminary form to encourage discussion and suggestions.

ABSTRACT

Greater use of employment tests for selecting workers will have important effects on the economy. First, the rewards for developing the competencies measured by the tests will rise and this will increase the supply of workers with these competencies. Employment tests predict job performance because they measure or are correlated with a large set of developed abilities which are causally related to productivity and not because they are correlated with an inherited ability to learn. Our economy currently under-rewards the achievements that are measured by these tests and the resulting weak incentives for hard study have contributed to the low levels of achievement in math and science.

Greater use of tests to select workers will also change the sorting of workers across jobs. Its impacts on total output depends on the extent to which the developed abilities measured by employment tests--academic achievement, perceptual speed and psychomotor skills--have larger impacts on worker productivity in dollars in some occupations than in others. This question is examined by analyzing GATB revalidation data on 31,399 workers in 159 occupations and by reviewing the literature on how the standard deviation of worker productivity varies across occupations. The analysis finds that indeed such differentials exist and therefore that reassigning workers who do well on a test to occupations where the payoff to the talent is particularly high will increase aggregate output. The magnitude of the output effect was estimated by reweighting the GATB revalidation data to be representative of the 71 million workers in the non-professional and non-managerial occupations and then simulating various resorting scenarios. Selecting new hires randomly lowered aggregate output by at least $129 billion or 8 percent of the compensation received by these workers. An upper bound estimate of the productivity benefits of reassigning workers on the basis of three GATB composites is that it would raise output by $111 billion or 6.9 percent of compensation. Reassignment based on tests had an adverse impact on Blacks and Hispanics but greatly reduced gender segregation in the work place and substantially improved the average wage of the jobs held by women. These results are based on a maintained assumption--the models of job performance which were estimated in samples of job incumbents are after corrections for measurement error and selection on the dependent variable yield unbiased estimates of true population relationships--that is almost certainly wrong. The biases introduced into the calculation by this assumption lower the estimated costs of introducing random assignment of workers to jobs, exaggerate the benefits of greater test use and exaggerate the changes in demographic composition of occupational work forces.

The paper concludes with a discussion of ways in which employment tests can simultaneously strengthen incentives to learn, improve sorting and minimize adverse impacts on minority groups.

10/20/88

THE ECONOMICS OF EMPLOYMENT TESTING

Employment testing appears to be destined to have a growing role in the allocation of workers to jobs. The skills measured by these tests are becoming increasingly important. Unskilled manufacturing jobs are moving to Asia, Africa and Latin America. The only way American manufacturers can stay in the US and survive is to automate. Automation, however, requires a highly skilled and flexible work force (Adler 1986). Employers are complaining that many new hires and long service employees do not have the reading, math and reasoning skills necessary to learn the demanding jobs being generated by the new information technology. At the same time that the demand for workers competent in basic skills and problem solving is rising, the supply appears to be contracting. The test scores of high school students fell during the 1970s and while they have rebounded somewhat, they have not yet returned to their former level.

These forces are causing American manufacturers to become more selective when they hire new workers. At the same time, the legal impediments to the use of aptitude tests appear to be diminishing (Scarf 1988). Even if the trend of court decisions accepting the claims of validity generalization were to be reversed, employers and society can gain most of the benefits of improved selection by top down hiring from a ranking generated by race normed test scores (Schmidt 1988; Wigdor and Hartigan 1988).[1] As a result, there is no necessary conflict between minority interests and greater use of tests in employment selection. As a result, test use appears to be growing. A 1985 American Society for Personnel Administration survey found that 24 percent of the firms responding had increased testing in the past year and another 44 percent were considering an increase in the amount of testing they do.

Greater use of employment tests will have three kinds of effects on the economy. First, greater use of tests increases the rewards for developing the skills and competencies assessed by the tests and as a consequence the supply of these skills and competencies will likely increase. Part I of the paper examines this effect, the incentive effect.

The second and third effects of greater use of tests by employers arise from their impact on the sorting of workers across jobs and occupations. Employment tests yield information on the probable job performance of job applicants that is not available from other sources (Hunter 1986, Schmidt

1988). If a trait measured by a test has a larger effect on dollars of output in occupation A than in occupation B, recruiting people who do well on the test into occupation A will increase national output. This is the second effect of greater test use. The third effect that greater use of tests for selection is likely to have is on the gender breakdown and ethnic makeup of particular occupations. These two sorting effects of employment testing are examined in the second part of the paper. The paper concludes with a discussion of the incentive and sorting efficiency effects of different methods of selecting workers for jobs and then recommends an approach to employment testing which simultaneously strengthens incentives to learn and improves the sorting of workers across jobs.

PART I. INCENTIVE EFFECTS

The prevalence and nature of employment testing powerfully influences incentives to develop particular skills and competencies. Greater use by employers of tests measuring competence in reading, writing, mathematics and problem solving will increase the supply of these competencies. The effects of testing on the aggregate supply of skilled workers may be considerably more important than its sorting effects. This tentative judgement follows from four propositions which will be defended below:

1. The tests at issue measure malleable developed abilities which are causally related to individual productivity (ie. they are not serving as proxies for inherited learning ability).

2. The labor market under-rewards the developed abilities measured by these tests.

3. Greater use of employment tests measuring a broad range of cognitive achievements such as the ASVAB would increase the economic rewards for learning.

4. People would devote more time and energy to developing these abilities if the rewards were greater.

The first of these propositions is defended in the first two sections of Part I. Section 1.1 presents evidence that gains in the quantity and quality of education and improvements in the cultural environment have caused major increases in average levels of achievement in the skills these tests assess. Because IQ scores of blacks and whites were advancing at roughly

equal rates, the gap between them remained relatively constant for a long period of time. Recently, however, the gap has begun to close apparently in response to the nation's efforts to raise the quality of education received by those from disadvantaged backgrounds. Section 1.2 argues that the distinction between an aptitude test and an achievement test is more in the eye of the beholder than in the psychometrics. The conclusion drawn from the evidence presented in these two sections is that employment tests predict job performance because they measure or are correlated with a large set of developed abilities which are causally related to productivity and not because they are correlated with an inherited ability to learn.

Section 1.3 presents evidence on proposition # 2 and # 3: our economy currently underrewards the academic achievements that are measured by employment tests. The minimal use of tests of academic achievement or credentials based on them as devices for selecting employees is in large part responsible for the failure of the labor market to appropriately reward effort and achievement in high school. Section 1.4 presents evidence for proposition # 4. It examines incentives to study hard in high school, and presents evidence that more powerful labor market rewards for learning are needed to strengthen these incentives.

1.1 The Implications of Secular Trends in IQ for the Nature/Nurture Debate

The cultural, economic and educational environment to which children and adults are exposed has improved dramatically in the last 70 years.[2] If proposition # 1 is true, these improvements in the cultural and educational environment should have resulted in major gains in the mean scores on IQ type tests. Since the gene pool of the nation can not have changed very much during the last 70 years, a genetic explanation of differences in IQ predicts almost no change in the mean IQ of the population. Consequently, the existence and magnitude of secular trends in the mean IQ of national populations is important evidence that the environment has major impacts on the developed abilities measured by employment tests.

This section of the paper, therefore, provides a review of 5 kinds of evidence on trends in the academic achievement and mean IQ of national populations and racial groups:

- o Cross-section and longitudinal data on the relationship between age and scores on individually administered IQ tests.
- o Time series data on the mean IQ of American adults on individually administered IQ tests.
- o Time series data on the scores of American students on group administered paper and pencil IQ tests.
- o Time series data on the mean IQ of students and young adults in other Western nations.
- o Trend data on Black/White differences in academic achievement and IQ.

This review concludes that for every group and in every nation for which there is data, IQ has risen over time. Even tests which were originally designed to measure inherited learning ability--individually administered IQ tests--exhibit this improvement. Rates of gain have not, however, been stable over time or equal across nations or racial groups.

Gains on Individually Administered IQ Tests-- Cross-section and Longitudinal Data

In cross section data the relationship between raw IQ scores and age is curvilinear. The highest scores are typically obtained by people in their twenties. Representative results from the standardization samples of the WAIS and WAIS-R, are presented in Table 1. Looking down the columns one can see that for those over age 30, raw scores for Full Scale IQ appear to decline .027 SDs per year and Verbal IQ scores appear to decline .013 to .014 SDs per year. A cross-section relationship like this is a mixture of aging effects and cohort effects. If the lower scores for those over the age of 30 or 35 reflect deterioration of the intellect over time, the cross-section data are consistent with the mean IQ of the population being stable over time. If, however, scores on these tests do not decline as one grows older, the cross section pattern implies that older cohorts had a lower IQ throughout their lifetime. This, in turn, implies that as new, better educated cohorts replaced older cohorts the mean IQ of the population advanced.

Longitudinal studies which retest individuals over long intervals of time are one way to distinguish between the aging and cohort explanations of the cross-section pattern. These studies have found that Verbal IQ rises with age and that only scores on the timed Performance IQ subtests decline

with age. Full scale IQ which combines the two types of tests does not deteriorate at least until age 60 and above (Bayley 1955; Bradway, et al. 1958; Schaie and Strather 1968; Schaie and Hertzog 1983). This, then, implies that cohort differences must be responsible for the lower raw IQ scores of those over age 35 and that the mean IQ of the population must be rising secularly.

Gains on Individually Administered IQ Tests- U. S. Time Series Data

It has been argued, however, that practice effects may last the full 7+ year interval between test administrations in these studies, so the improvements in IQ with age may be illusory. Therefore, other kinds of data which are not subject to practice effects must be examined to obtain conclusive proof that the mean IQ of the population has risen over time. What is needed is large random samples of the adult population which have been given the same or equated IQ tests many years apart. The standardization studies for the WAIS and WAIS-R adult IQ tests provide the necessary stratified random samples of the population, and five studies have been published which equate the two tests. The equating studies determined the correspondence of scores on the earlier and later versions of the test by administering both tests in counterbalanced order (to neutralize practice effects) to a sample of people. These studies found that it was easier to get a high score on the WAIS which had been standardized on the U.S. population during 1953-54 than on the WAIS-R which was standardized between 1976 and 1980. The implied gain in IQ for the population 16 to 70 years old was 6.37 IQ points (.425 SDs) on full scale IQ and 6.48 IQ points (.432 SDs) on verbal IQ between 1953 and 1978. (Wechsler 1981, Urbina, Golden & Ariel 1982; Smith 1983; Mishra & Brown 1983; Lippold & Claiborn 1983). These equating studies have been used to calculate the relationship between column 1 and 2 and between column 4 and 5 of Table 1. Yearly rates of gain in IQ for each age group are presented in column 3 and 6 of table 1. The IQ gains for cohorts over the age of 25 were generally around .02 SDs per year.[3]

The equating studies comparing other IQ tests obtain similar results: IQ scores on the older tests are almost invariably higher and the magnitude of the difference is linearly related to the time between standardizations of the tests (Thorndike 1975; Flynn 1984). Flynn finds that the rate of

increase in Full Scale IQ is remarkably consistent across age groups and time periods. Analyzing 77 equating studies with a total of 7431 subjects involving 18 different IQ test comparisons, he concludes that between 1932 and 1978 the gains have averaged 3.1 IQ points or .21 standard deviations per decade (Flynn 1987).

Comparisons of white enlisted soldiers serving in World War I and II provide still another measure of secular trends in IQ scores. When a stratified random sample of white WWII recruits took a test very similar to the test that all literate WWI army recruits had taken, they scored .73 standard deviations or 11 IQ point better. (Tuddenham, 1948).[4] Since only the literate 83 percent of white soldiers took the Alpha during WWI, this comparison understates the IQ gain for the population as a whole.

IQ Gains Among American Students

A third source of data on IQ trends is provided by studies of trends in the IQ of students. A search was conducted for studies reporting the results of administering the same IQ test to different cohorts of students at the same school. Only a few such studies were found all of them covering the period between the two world wars. A study of the school children of eastern Tennessee found that, over the decade of the 1930s, 1st graders gained 11 IQ points and 7th and 8th graders gained 10.8 IQ points (more than two-thirds of a standard deviation). (Wheeler, 1942). A study of two high schools in the midwest found no change in the mean IQ of the students at a small rural high school and a 5 point increase between 1923 and 1942 at a large high school serving a small city and the surrounding county (Finch, 1946). Johnson (1935) found a 3 point gain in IQ between 1925 and 1935 at Grover Cleveland High School in St. Louis. Roessel's (1937) comparison of the students in three Minnesota high Schools in 1920 and 1934 and Rundquist's (1936) comparison of Minneapolis high school students in 1929 and 1934 both found increases in IQ.

For the post-WWII era, the best data on trends in the academic achievement of students nearing completion of compulsory schooling comes from the Iowa Test of Educational Development (ITED), a battery of achievement tests very similar to the ACT (Forsyth 1987).[5] Because about 95 percent of the public and private schools in the state of Iowa regularly participated in the testing

program, the analysis of trends in ITED data for Iowa is not plagued by changing selectivity of the population taking the test in the way the SAT and ACT are. The overall trend of ITED scores for Iowa high school seniors has been up, but progress has not been steady. There was an increase of .426 population standard deviations between 1942 and 1967, a decline of .26 population standard deviations between 1967 and 1979, and then a rebound of .185 population standard deviations between 1979 and 1987.

IQ Trends in Europe, Canada and Japan

Improvements in the mean IQ of the population have been even more dramatic in Europe and Japan than in the United States. James R. Flynn (1987) has collected studies of the trend in mean IQ of representative samples of youth and young adults for 14 advanced nations. Table 2 reports the findings of the studies for which there can be no debate about the representativeness of the populations tested. In every country for which data was available (including the countries with weaker studies), there were major gains on IQ tests during the post war period. The findings for France, Netherlands, Norway and Belgium are especially strong. In these countries unchanged tests were given to all male 18 year olds entering their universal military training obligation. In just 25 years, for example, the IQ scores of French 18 year olds rose 25 points on the Ravens, a "culture reduced" test of abstract problem solving ability, and 9.4 points on a more conventional test of verbal and mathematical intelligence. In general, test score gains were smaller on math and verbal tests than on tests of abstract problem solving ability and the performance components of Wechsler IQ tests. The countries such as Japan, Belgium, France and Netherlands that have above average rates of gain in IQ tend also to have above average rates of productivity growth. This correlation probably reflects the combined effect of the IQ gains on national productivity and the effect of improved standards of living on IQ.

Racial Gaps in Academic Achievement and IQ

Gottfredson (1988) argues that racial differentials in IQ are very stubborn and that, consequently, color blind use of employment tests will probably have an adverse impact on blacks for generations. Citing Gordon (in press), she reports that the IQ gap between whites and blacks has been relatively stable ever since 1918. Since mean IQ levels of the entire

population have risen substantially during the 20th century, this implies that the mean IQ of blacks has been rising as well. The failure of the gap to close during the first 50 years of psychometric testing takes on much less significance when one realizes that both groups were rapidly improving.

In more recent data, however, the gap is closing. Blacks born after the civil rights revolution are doing much better in school than those born prior to 1960. The evidence is presented in Table 3. In the first National Assessment of Educational Progress black high school seniors born around 1954 were 5.3 grade level equivalents behind their white counterparts in reading proficiency. In the first assessment of math skills, black high school seniors born around 1957 were 4 grade level equivalents behind in mathematics. The most recent National Assessment data for 1986 reveals that the gap in math proficiency has been cut to 2.9 grade level equivalents in just 12 years and that the reading gap has been cut to 2.6 years in just 15 years. Koretz's (1986 Appendix E) analysis of data from state testing programs supports the NAEP findings. Gains of this magnitude contradict Gottfredson's very pessimistic assessment and suggest that Head Start, Title I and other compensatory interventions are having an impact. The schools attended by most black students are still clearly inferior to those attended by white students so further reductions in the school quality differentials will probably produce further reductions in academic achievement differentials.

The evidence just reviewed implies that (1) the mean IQ of national populations has been growing over time and (2) the relative IQ of different national populations and ethnic groups has been changing over time. Since the genetic makeup of these populations cannot have changed appreciably in just one or two generations, these results clearly imply that changes in the environment--education, nutrition, culture and economic status--can and do produce substantial changes in the developed abilities that are measured by IQ tests and employment aptitude tests.

1.2 Is There a Psychometric Distinction Between Achievement and Aptitude?

Broad spectrum achievement tests correlate almost as highly with verbal and mathematical aptitude tests as alternate forms of the same test correlate with each other. The similarity of subject content is a much more important determinant of correlations between tests than is the distinction between

aptitude and achievement. This is clearly visible when one examines attenuation corrected correlations between College Board achievement and aptitude tests. When subject matter is similar, corrected correlations between aptitude and achievement tests are close to one. The corrected correlation between Math I and the Math SAT is .93 and the correlation between English Literature and the Verbal SAT is .93 (College Board 1984, 1987). When subject matter differs, correlations are much lower. Corrected for attenuation, Math SATs correlate .721 with Verbal SATs, .738-.756 with science achievement and .526-.576 with history achievement. Corrected for attenuation, Verbal SATs correlate .789-.831 with history, .793 with biology and .625-.643 with chemistry and physics.[6]

There are good reasons for high correlations between past achievement and scores on aptitude tests designed to predict future achievement. Past achievement aids learning because the tools (e.g. reading and mathematics) and concepts taught early in the curriculum are often essential for learning the material that comes later. Furthermore, aptitude tests are validated on later achievement levels, not on rates of change of achievement. Consequently, the items that are included in paper and pencil aptitude tests often look a lot like the items that appear on achievement tests. The tests that do not fit this generalization, such as the WAIS-R's digit span, typically have lower validity than other subtests or are measures of something altogether different like short-term memory and psychomotor ability.

Further evidence on this issue is provided by the many studies which have shown that school attendance raises scores on these aptitude tests (Lorge 1945; Husen 1951; Department of Labor 1970), and that taking a rigorous college prep curriculum increases the gains on these tests between sophomore and senior years of high school (Bishop 1985; Hotchkiss 1985). In recognition of the fact that aptitude test scores are significantly influenced by educational background, the College Board now describes the SAT as a measure of "developed verbal and mathematical reasoning abilities (1987, p. 3)"

At this point it is important to address a potential objection to this conclusion. Those who believe that IQ tests truly measure inherited learning ability might argue that productivity is an outcome of on-the-job learning rather than in-school learning, and that employment aptitude tests measure inherited learning ability rather than outcomes of schooling that help one

do a job well. In this view, employment aptitude tests are good measures of inherited learning ability because everyone receives roughly equivalent instruction in the material covered by the test, therefore, differences in knowledge at the end of instruction primarily reflect differences in inherited learning ability. In my judgement, this view does not withstand scrutiny.

Many of its key predictions are contradicted by data. (1) If it were true, we would expect childhood IQ tests to predict adult labor market success just as well as adult GIA tests. In fact, when adult GIA tests compete with childhood IQ tests, it is the adult test not the childhood test which has by far the biggest effect on labor market success (Husen, 1969). (2) In addition, we would expect culture reduced non-verbal IQ tests to be just as good predictors of labor market success as a test of reading and writing skills. In fact, a study of Kenyan workers has found that wages were significantly effected by literacy but not by non-verbal IQ (Brossiere, Knight and Sabot, 1985). (3) Furthermore, we would expect education obtained abroad in non-English speaking countries to be just as good a signal of high inherited learning ability (and therefore just as good a predictor of wage rates in the U.S. economy) as education obtained in the U.S. or English speaking countries. In fact, a year of schooling obtained in a non-English speaking country has a much smaller effect on wage rates than a year of schooling obtained in the U.S. or some other English-speaking country. (Chiswick, 1978). (4) Finally, we would expect that controlling for genotype IQ (e.g. by comparing identical twins) would reduce the effect of test scores on labor market success to zero. Since siblings are genetically similar, we would expect IQ's effect to diminish when siblings are being compared. In fact, the effect of IQ (measured while in school) on labor market success is <u>actually greater when brothers are compared</u> than in standard cross section regressions (Olneck 1977).

These findings suggest that the associations between scores on employment aptitude and IQ tests on the one hand and productivity and labor market success on the other arise because the tests either directly measure developed abilities that contribute to productivity or else correlate highly with such abilities. Therefore, an increase in the incidence of these developed abilities in the working population will increase national output. This raises the question of what determines the incidence of these developed abilities

in the adult population. Propositions # 2, # 3 and # 4 summarize the findings that will be presented in the next two subsections of the paper.

1.3 The Absence of Major Economic Rewards for Effort in High School

The decline in test scores and the poor performance of American students on international mathematics and science tests has stimulated a great deal of concern about the quality of education. An educational reform movement has developed that is attempting to add rigor to the curriculum and improve teaching. These are important objectives, and important progress has been made. If, however, students are not motivated to study harder, the reform initiatives will fail. Too little attention has been given to student motivation. In the area of student motivation, employment testing potentially has an important role to play.

Studies of time use and time on task in high school show that students actively engage in a learning activity for only about half the time they are scheduled to be in school. In 1980, high school students spent an average of 3.5 hours per week on homework. When homework is added to engaged time at school, the total time devoted to study, instruction, and practice is only 20 hours per week. By comparison, the typical senior spent 10 hours per week in a part-time job and nearly 25 hours watching television. Thus, TV occupies more of an adolescents time than learning.

Even more important is the intensity of the student's involvement in the process. Theodore Sizer described American high school students as **"docile, compliant, and without initiative"** (Sizer 1984). Coming to the same conclusion, John Goodlad observed, "the extraordinary degree of student passivity stands out" (Goodlad 1984). When teachers are asked what they feel are the most important problems in education, more than 40% respond, "lack of interest by students". This lack of interest makes it very difficult for teachers to be demanding.

Some teachers are able to overcome the obstacles and induce their students to undertake hard learning tasks. But for most mortals the lassitude of the students is too demoralizing. In too many classrooms an implicit agreement prevails in which the students trade civility for lowered academic demands (Sizer 1984). Most students view the costs of studying hard as much greater

than the benefits, so the peer group pressures the teacher to go easy. All too often teachers are forced to compromise their academic demands.

Students are not, however, the only group that is apathetic. Stevenson, Lee and Stigler's (1986) study of education in Taiwan, Japan and the U.S. found that even though American children were learning the least in school, American parents were the most satisfied with the performance of their local schools. Why do Japanese and Taiwanese parents hold their children and their schools to a higher standard than American parents?

The fundamental cause of the apathy and motivation problem is the way student effort and achievement is recognized and reinforced. The educational decisions of students and their parents are significantly influenced by the costs (in money, time and psychological effort) and benefits (praise, prestige, employment, wage rates, and job satisfaction) that result. The problem is that while there are benefits to staying in high school, most students do not benefit very much from working hard while in high school. This is in large measure a consequence of the failure of the labor market to reward effort and achievement in high school.

Students who plan to look for a job immediately after high school generally see very little connection between their academic studies and their future success in the labor market. Statistical studies of the youth labor market confirm their skepticism about the economic benefits of studying hard:

- For high school students, high school grades and performance on academic achievement/aptitude tests have essentially no impact on labor market success. They have -
  --no effect on the chances of finding work when one is seeking it during high school, and
  --no effect on the wage rate of the jobs obtained while in high school.(Hotchkiss, Bishop and Gardner 1982)

- As one can see in table 5, for those who do not go to college full-time, high school grades and test scores had -
  --no effect on the wage rate of the jobs obtained immediately after high school in Kang and Bishop's (1985) analysis of High School and Beyond seniors and only a 1 to 4.7 percent increase in wages per standard deviation (SD) improvement in test scores and grade point average in Meyer's (1982) analysis of Class of 1972 data.
  --a moderate effect on wage rates and earnings after 4 or 5 years [Gardner (1982) found an effect of 4.8 percent per SD of achievement and Meyer (1983) found an effect of 4.3 to 6.0 percent per SD of achievement],

--a small effect on employment and earnings immediately after high school.

- In almost all entry-level jobs, wage rates reflect the level of the job not the worker's productivity. Thus, the employer immeadiately benefits from a worker's greater productivity. Cognitive abilities and productivity make promotion more likely, but it takes time for the imperfect sorting process to assign a particularly able worker a job that fully uses that greater ability -- and pays accordingly.

The long delay before labor market rewards are received is important because most teenagers are "now" oriented, so benefits promised for 10 years in the future may have little influence on their decisions.

Although the economic benefits of higher achievement are quite modest for young workers and do not appear until long after graduation, the benefits to the employer (and therefore, to national production) are immediately apparent in higher productivity. This is the implication of the finding that tests of reading, mathematics and problem solving ability are valid predictors of job performance in most jobs.

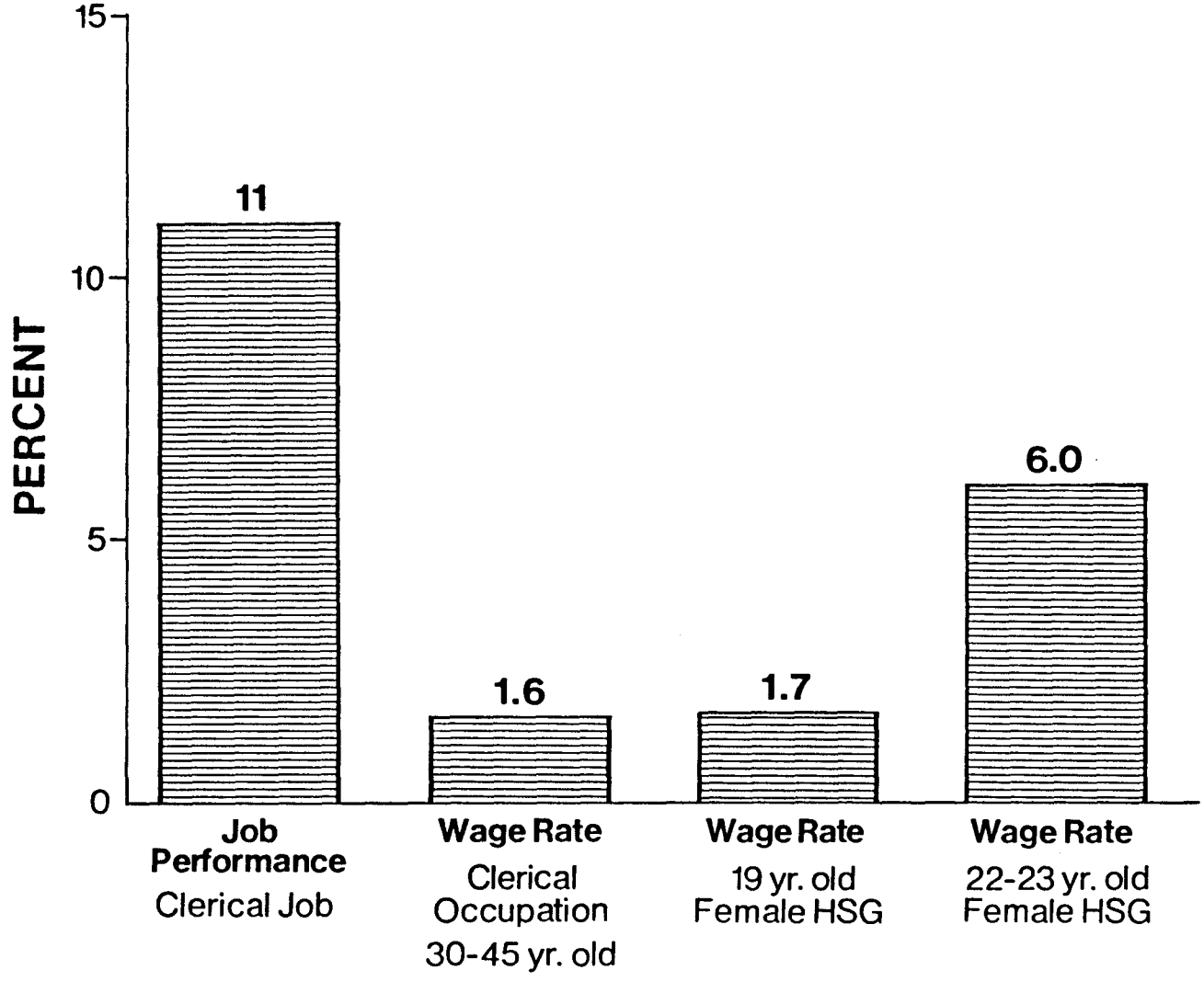
Figure 1 compares the percentage impact of mathematical and verbal achievement (specifically a one standard deviation difference in GPA and test scores) on the productivity of a clerical worker, on wages of clerical workers, and on the wages of all workers who have not gone to college[7,8]. Productivity clearly increases more than wage rates. This implies that when a non-college-bound student works hard in school and improves his or her academic achievements the youth's employer benefits as well as the youth. The youth is more likely to find a job, but not one with an appreciably higher wage. The next sub-section examines reasons for the discrepancy.

Reasons for the Discrepancy between Wage Rates and Productivity on the Job

Employers are presumably competing for better workers. Why doesn't competition result in much higher wages for those who achieve in high school and have strong basic skills? The cause appears to be the lack of objective information available to employers on applicant accomplishments, skills, and productivity.

A 1987 survey of a stratified random sample of small and medium sized employers who were members of the National Federation of Independent Business (NFIB) found that aptitude test scores had been obtained in only 3.15 percent of the hiring decisions studied (Bishop and Griffin 1988). Top down hiring

Figure 1


15
10
5
0
PERCENT
11
1.6
1.7
6.0
Job Performance
Clerical Job
Wage Rate
Clerical Occupation
30-45 yr. old
Wage Rate
19 yr. old Female HSG
Wage Rate
22-23 yr. old Female HSG

on the basis test scores is even more unusual. Prior to 1971, employment testing was much more common. The cause of this change was the fear of costly litigation over the business necessity and validity of aptitude tests. The EEOC's codification of the APA's professional testing standards and its theory of situational and subgroup differences in validity into federal law made the required validation studies so costly it discouraged almost all employers from undertaking the effort (Friedman and Williams 1982).

Other potential sources of information on effort and achievement in high school are transcripts and referrals from teachers who know the applicant. Both these means are under used. In the NFIB survey, transcripts had been obtained prior to the selection decision for only 13.7 percent of the hiring events in which someone with 12 or fewer years of schooling was hired. If a student or graduate gives written permission for a transcript to be sent to an employer, the Buckley amendment obligates the school to respond. Many high schools are not, however, responding to such requests. The experience of Nationwide Insurance, one of Columbus, Ohio's most respected employers, is probably representative of what happens in most communities. The company obtains permission to get high school records from all young people who interview for a job. It sent over 1,200 such signed requests to high schools in 1982 and received only 93 responses. Employers reported that colleges were much more responsive to transcript requests than high schools. High schools have apparently designed their systems for responding to requests for transcripts around the needs of college bound students not around the needs of the students who seek a job immediately after graduating.

There is an additional barrier to the use of high school transcripts in selecting new employees--when high schools do respond, it takes a great deal of time. For Nationwide Insurance the response almost invariably took more than 2 weeks. Given this time lag, if employers required transcripts prior to making hiring selections, a job offer could not be made until a month or so after an application had been received. Most jobs are filled much more rapidly than that. The 1982 NCRVE employer survey of employers found that 83.5 percent of all jobs were filled in less than a month, and 65 percent were filled in less than 2 weeks.

The only information about school experiences requested by most employers is years of schooling, diplomas and certificates obtained, and area of

specialization. Probably because of unreliable reporting and the threat of EEOC litigation, only 16 percent of the NFIB employers asked the applicants with 12 or fewer years of schooling to report their grade point average. Hiring on the basis of recommendations by high school teachers is also uncommon. In the NFIB survey, when someone with 12 or fewer years of schooling was hired, the new hire had been referred or recommended by vocational teachers only 5.5 percent of the time and referred by someone else in the high school only 3.1 percent of the time.

Consequently, hiring selections and starting wage rates often do not reflect the competencies and abilities students have developed in school. Instead, hiring decisions are based on observable characteristics (such as years of schooling and field of study) that serve as signals for the competencies the employer cannot observe directly. As a result, the worker's wage reflects the average productivity of all workers with the same set of educational credentials rather than that individual's productivity or academic achievement.

This evidence implies that the social benefits of developing one's verbal, mathematical and scientific capabilities are considerably greater than the private rewards. Despite their higher productivity, young workers who have achieved in high school and who have done well on academic achievement tests do not receive higher wage rates immediately after high school. The student who works hard must wait many years to start really benefiting and even then the magnitude of the wage and earnings effect--a 1 to 2 percent increase in earnings per grade level equivalent on achievement tests--is considerably smaller than the actual change in productivity that results.

1.4 Will Larger Economic Rewards for Learning Induce Students to Study Harder ?

Learning that is certified by a credential is rewarded handsomely. The magnitude of the earnings payoff to a credential has been shown to have significant effects on the numbers of students entering college and choosing specific majors (Freeman 1971, 1976). Learning not certified by a credential is either not rewarded or only modestly rewarded. Consequently, there are strong incentives to stay in school; but much weaker incentives to study hard while in school. If students are to be motivated to devote more time and energy to learning, they must believe their effort will be rewarded. If

parents are to be induced to demand better schools and to spend the time supervising homework, they too must believe that better teaching, a more rigorous curriculum and hard study produces learning which will be rewarded in the labor market. When, however, the only signals of learning accomplishment that are available--eg. GPA and rank in class--describe one's performance relative to close friends, the motivation to study and to demand better schools is undermined.

The Zero-Sum Nature of Academic Competition in High School

The second root cause of the lack of real motivation to learn is peer pressure against studying hard. Students report that "in most of the regular classes... If you raise your hand more than twice in a class, you are called a 'teachers pet.'" Its OK to be smart, you cannot help that. It is definitely not OK to study hard to get a good grade. An important reason for this peer pressure is that the academic side of school forces adolescents to compete against close friends. Their achievement is not being measured against an absolute or an external standard. In contrast to scout merit badges where recognition is given for achieving a fixed standard of competence, the only measures of achievement that receive attention in American schools are measures of one's performance relative to one's close friends such as grades and rank in class. When students try hard and excel in school, they are making things worse for friends. Since greater effort by everyone cannot improve everyone's rank in class, the group interest is for everyone to take it easy. At that age peer friendships are all important, so informal pressure from the peer group is able to induce most students to take it easy. All work groups have ways of sanctioning "rate busters." High school students call them "brain geeks", "grade grubbers" and "brown nosers".

Young people are not lazy. In their jobs after school and at football practice they work very hard. In these environments they are not competing against each other. They are working together as part of a team. Their individual efforts are visible to their peers and appreciated by them. On the sports field, there is no greater sin than giving up, even when the score is hopelessly one sided. In too many high schools, when it comes to academics, there is no greater sin than trying hard.

Another reason for peer norms against studying is that most students perceive the chance of receiving recognition for an academic achievement to be so slim they have given up trying. At most high school awards ceremonies the recognition and awards go to only a few--those at the very top of the class. By 9th grade most students are so far behind the leaders, they know they have no realistic chance of being perceived as academically successful. Their reaction is often to denigrate the students who take learning seriously and to honor other forms of achievement--athletics, dating, holding your liquor and being "cool"--which offer them better chances of success.

The lack of standards for judging academic achievement that do not involve comparisons with one's close friends and the resulting zero sum nature of academic competition also influences the school board and the political system. Parents can see that setting higher academic standards or hiring better teachers will not improve their child's grade point average or rank in class. Since the Scholastic Aptitude Test is intended to be curriculum free, adding rigorous science, history and calculus courses to the curriculum is unlikely to change SAT scores. In any case, doing well on the SAT matters only for those who aspire to attend a small number of highly selective colleges. The parents of children not planning to go to college have an even weaker incentive to demand high standards. They believe that what counts in the labor market is getting the diploma not learning algebra (and they happen to be right). Higher standards might put at risk what is really important--the diploma.

The real costs of mediocre schools become apparent only to employers and to officials at higher levels of government. The whole community loses because the work force is less efficient and it becomes difficult to attract new industry. This is precisely the reason why employers, governors and state legislatures have been the energizing force of school reform. State governments, however, are far removed from the classroom and the instruments available to them for imposing reform are limited. If students, parents and school board officials perceive the rewards for learning to be minimal, state efforts to improve the quality of education will not succeed.

Evidence of a Learning Response to Economic Incentives

The tendency to under-reward effort and learning in school appears to be a peculiarly American phenomenon. Grades in school are a crucial determinant of which employer a German youth apprentices with. Top companies in Japan and Europe often hire lifetime employees directly out of secondary school. Teacher recommendations, grades in school, and scores on national and provincial exams have a significant impact on who gets to work at the more prestigious firms (Leestma, et. al., 1987). Japanese parents know that their son or daughter's future economic and social rank in society critically depends on how much he or she learns in secondary school. Furthermore, learning achievement tends to be defined and measured relative to everyone else in the state or nation and not just relative to one's classmates in the school. Entry into the better high schools depends primarily on the child's performance in junior high school, not on where the parents can afford to live as occurs in the US. These are the reasons why Japanese parents demand so much of their children and of their schools. This is why Japanese 5th graders spend 32.6 hours a week in academic activities while American youth devote only 19.6 hours to their studies (Stevenson, Lee and Stigler 1986).

Japanese adolescents work extremely hard in high school, but once they have entered college, they stop working. For most students a country club atmosphere prevails. The reason for the change in behavior is that employers apparently care only about which university the youth attends, not about the individual's academic achievement at the university. Working hard is not a national character trait, it is a response to the way Japanese society rewards academic achievement.

American students, in contrast, take it easy in high school but generally work quite hard in college. This change is due, in part, to the fact that academic achievement in college has important effects on labor market success. When higher level jobs requiring a bachelors or associates degree are being filled, employers pay much more attention to grades and teacher recommendations than when they hire high school graduates. The NFIB survey found that when someone with 16 or more years of schooling was hired, 26 percent of the employers had reviewed the college transcript before making the selection, 7.8 percent had obtained a recommendation from a major professor and 6.3 percent had obtained a recommendation from a professor outside of the graduates major or from the colleges's placement office.

PART II. SORTING EFFECTS

Hunter and Schmidt (1982) employ Brogden's formula to calculate the effect of test use on the efficiency of the economy's matching of workers to jobs. In this context Brogden's formula can be viewed as a way of representing the derivative of a worker's true productivity measured in dollars with respect to a test score:

$$(1)\quad \frac{\partial \underline{P}^{t}_{ij}}{\partial T_{ij}} = \frac{\mathrm{Cov}(\underline{P}^{t}_{ij}, \underline{T}_{ij})}{\mathrm{Var}(T_{ij})} = r_{TP}\,\frac{\mathrm{SD}(\underline{P}^{t})}{\mathrm{SD}(T)}$$

They point out that tests are more valid predictors of job performance (eg. have higher r_{TP}) in the more complex jobs that are traditionally better paid and therefore probably also have larger standard deviations of productivity in a dollar metric, $SD(P^t)$. When this is the case, output will increase if high scoring individuals are recruited into the most complex jobs and low scoring individuals are recruited into the less complex jobs. They make a simplifying assumption that the ratio of the standard deviation of output in dollars to the wage is the same in all jobs but argue it is quite large, about 40 percent of salary. Under this assumption, they calculate that distributing all workers across four major occupational categories on the basis of a single measure of academic ability will raise productivity 4 percent above the level resulting from random assignment of workers to major occupational category. They also report that assigning workers on the basis of a simple multi-variate selection model involving tests of perceptual speed and spatial ability as well as academic ability would increase productivity by 8 percent relative to random assignment.

However, since people are already recruited into high status jobs on the basis of years of schooling, SAT scores, college major, grades, previous work experience and performance in past jobs (which have independent associations with job performance and together explain much of the variance of test scores), greater use of tests by employers would probably have much smaller effects on national output than those calculated by Hunter and Schmidt. Hunter and Schmidt acknowledge this when they say, "Employers do not select randomly from among applicant pools.... many of these [selection] procedures have low validity, but average productivity levels associated with current methods are certainly above those that would result from random selection

from applicant pools, though less effective than our univariate selection strategy (p. 270)". Michael Rothschild (1979) has proposed two other sources of upward bias in their estimate. He argues that the assumption of optimal placement is unreasonable. Tests would never be used by all firms, for all jobs and optimally in every case, so the full benefits calculated would never be realized. A second source of bias, in Rothschild's view, is the possibility that errors in measuring productivity may be positively correlated with test score, and that consequently the estimates of true validity and the standard deviation of true output used in the analysis may be biased. Hunter and Schmidt argue to the contrary that their estimates are conservative because they assume that (1) coefficients of variation of productivity are the same for all occupations, (2) at most three test scores are used to reassign workers and (3) only 4 categories of occupations are analyzed. They point out that these features of the calculation cause it to understate the effects of greater test use on national productivity.

The only way to analyze whether the H/S estimates are too high or too low is to change as many of the problematic assumptions as possible and then redo the calculation. That is what will be attempted in this part of the paper. The objective is an improved estimate of the magnitude of the efficiency gains that may result from greater test use, not a definitive estimate. In the current state of knowledge, a definitive estimate is infeasible for some important sources of bias cannot be eliminated. There is no way of knowing, for example, how effectively tests will be incorporated into selection decisions and whether the measurement errors of job performance are correlated with test scores or not, so it will not be possible to formally address two of Rothschild's objections to H/S's estimates. Most of the factors that Hunter and Schmidt argue cause their estimates to be conservative are dealt with , however, so the resulting estimates are probably upper bounds on the likely impact of greater test use on the productivity of the economy.

Greater use of tests will increase aggregate output either if tests are more valid predictors of job performance in some jobs than others or if improvements in job performance have larger effects on output valued in dollars in some occupations than others. I begin, therefore, by examining how test validity varies across occupations. This is accomplished by estimating "structural" models of relative productivity as a function of three tests

scores (general academic achievement, perceptual speed and psychomotor skills), years of schooling, age, total occupational experience, tenure, gender, race and Hispanic background for 8 different occupational categories in the United States Employment Service's General Aptitude Test Battery Revalidation Individual Data File.

The next step is a review of the literature on how variable output is across workers doing the same job and how this variability is affected by the nature of the job. The major finding here is that the standard deviation of output is substantially higher in the more cognitively complex and better paid jobs.

The effect of alternative ways of assigning workers to jobs is calculated by simulating such changes in the USES Individual Data File after reweighting it to be representative of all workers outside of professional, managerial and sales representative occupations. The parameters of the "structural" models are used to predict the productivity (in standard deviation units) during the first ten years on the job of all 31,399 workers in the data set in each of the 8 occupational categories analyzed. The mean predicted productivity of workers who currently occupy each job is then compared to the productivity that would result from (1) a random assignment of new hires to jobs and (2) a resorting of new hires across jobs based on the productivity predictions generated by regression equations similar to the structural models but absent data on gender, race and Hispanic background. These results are then translated into a dollar metric by multplying changes in mean productivity in standard deviation units by estimates of the standard deviation of productivity in dollars obtained from the literature review. The impact of reassignment based on test scores on the gender, racial, and Hispanic composition of each occupation is also simulated and discussed. Part 2 then concludes with a critique of the estimated "structural" models of job performance and the resulting estimates of productivity gains from resorting the workforce on the basis of employment tests.

2.1 ANALYSIS OF GATB VALIDATION STUDIES

Data on the relative productivity of a large and reasonably representative sample of workers is available from the US Employment Service's program for revalidating the General Aptitude Test Battery (GATB). This data set contains

data on job performance, the 9 GATB "aptitudes" and background data on 36,614 individuals in 159 different occupations. Professional, managerial and high level sales occupations were not studied but the sample is quite representative of the rest of the occupational distribution. It ranges from drafters and laboratory testers to hotel clerks and knitting-machine operators. The simulations of the effect of changes in selection policies are also conducted in this data set after it has been reweighted to be representative of the 71,132,000 workers who are employed in these occupations.

Since a major purpose of these validation studies was to examine the effects of race and ethnicity on the validity of the GATB, the firms that were selected tended to have an integrated workforce in that occupation. Firms that used aptitude tests similar to the GATB for selecting new hires for the job being studied were excluded. The employment service officials who conducted these studies report that this last requirement did not result in the exclusion of many firms. A total of 3052 employers participated.

Each worker took the GATB test battery and supplied information on their age, education, plant experience and total experience. Plant experience was defined as years working in that occupation for the current employer. Total experience was defined as years working in the occupation for all employers. The dependent variable for this study is a sum of two separate administrations (generally two weeks apart) of the Standard Descriptive Rating Scale. This rating scale (available from the author), obtains supervisory ratings of 5 aspects of job performance (quantity, quality, accuracy, job knowledge and job versatility) as well as an "all around" performance rating. Some studies employed rating scales specifically designed for that occupation and in one case a work sample was one of the job performance measures. None of the studies used ticket earnings from a piece rate pay system as the criterion. Studies which used course grades or tests of job knowledge as a criterion were excluded. Firms with only one employee in the job classification were excluded, as were individuals whose reported work experience was inconsistent with their age.

Academic achievement is the sum of two GATB composites, G and N, that have been put into a population SD metric by dividing by 38.8. The G composite is an average of normalized scores on a vocabulary test, an arithmetic reasoning test and a 3-dimensional spatial relations test. The mathematical

achievement index (N) is an average of normalized scores on the same arithmetic reasoning test and on a numerical computations test. These two GATB composites were aggregated together because previous analyses had found that when both were entered simultaneously into models predicting relative job performance, the coefficients on both composites were very similar (Bishop 1987). Perceptual Speed is the sum of the P and Q aptitudes of the GATB divided by 36.72 to put it in a population SD metric. Psychomotor Ability is the sum of the K, F and M aptitudes of the GATB divided by 51.54 to put it in a population SD metric.

Because wage rates, average productivity levels and the standards used to rate employees vary from plant to plant, mean differences in ratings across establishments have no real meaning. Only deviations in rated performance from the mean for the establishment (R^m_j) were analyzed. The variance of the job performance distribution was also standardized across establishments by dividing (R^m_{ij}-R^m_j) by the standard deviation of performance calculated for that firm (or 3 if the sample SD is less than 3).[9] The model that was fitted to the data was the following:

$$(2)\quad \frac{R^m_{ij}-R^m_j}{SD_j(R^m_{ij})} = R_{ij}-R_j = \beta_0 + \beta_1(T_{ij}-T_j) + \beta_2(S_{ij}-S_j) + \beta_3(X_{ij}-X_j) + v_1$$

where T_{ij} is a vector of the three GATB composites and S_{ij} is the schooling of the i^{th} individual. T_j and S_j are the means of test scores and schooling for the j^{th} establishment. Two models were estimated for each major occupation. In the first model, the vector of individual characteristics, $X_{ij}-X_j$, included the deviations from the firm's mean of gender, Black, Hispanic, age, age squared, plant experience, plant experience squared, total occupational experience and total occupational experience squared. The calculation of the effects on aggregate output of reassigning workers to jobs that is carried out in section 2.3 is based on the predictions of this model. It should be recognized that because of the selectivity of the application and hiring process and of turnover and promotions, the results obtained from fitting this model are not estimates of the true structural relationships prevailing in the full population (Brown 1978; Mueser and Maloney 1987). Since no data sets exist which would enable analysts to model these selection processes, estimates of the true population relationships do not appear to be feasible. An effort will be made in section 2.4 to discuss how the

simulation results would probably change if better estimates of true population relationships were available.

In the second model, gender, race and Hispanic were not included. Because it is illegal for firms to select workers on the basis of gender, race and ethnicity, the selection process must be assumed to ignore this information so the simulation exercise conducted in section 2.3 assumes that workers are assigned to jobs on the basis of performance predictions generated by model two.

The results of estimating Model 2 are presented in Table 5. When test scores are controlled, years of schooling appear to have very small and sometimes negative effects on job performance.[10] The effects of the three test score composites are reported in columns 2-4 of the table. When the metric of job performance is within-job standard deviations, academic achievement has roughly comparable effects on job performance in all occupations except operatives and sales clerks. The effect of academic achievement on the performance of operatives is highly significant but only about two-thirds of the size of the other occupations. Perceptual speed has smaller effects on job performance, but the coefficients are nevertheless significant in all but technical and sales clerk (where the sample is quite small) occupations. Psychomotor skills are significantly related to performance in all occupations but in the better paid and more complex jobs the magnitude of the effect is only about one-third of that of academic achievement. The effect of psychomotor skills is larger in the three least skilled occupations--operatives, sales clerks and service except police and fire. For operatives and sales clerks the impact of psychomotor skills is roughly comparable to the impacts of academic achievement. These results are consistent with previous literature (Hunter 1983). Models were estimated containing squared terms for academic achievement and psychomotor skills but these additions did not produce significant reductions in the residual variance. Estimating model 1 by adding dummy variables for gender, race and Hispanic to model 2, tends to reduce test score coefficients a little but the pattern remains the same.

The effects of occupational experience and tenure are also quite substantial for all occupations except for sales clerks. The negative coefficients on the square terms for occupational experience and tenure imply

they are subject to diminishing returns. For workers who have no previous experience in the field, the expected gain in job performance is about 12-13 percent of a standard deviation in the first year and about 8-9 percent of an SD in the fifth year. The effect of tenure on job performance stops rising and starts to decline somewhere between 16 and 24 years of tenure. Increases in occupational experience lose their positive effect on performance even later--at 37 years for operatives, at over 55 years for craft workers and high skill clerical workers and at 19-31 years for other occupations. Except for technicians, age has large curvilinear effects on job performance as well.

The substantial effects of age and previous occupational experience on job performance are consistent with current hiring practices which give great weight to these job qualifications. These results suggest that a job applicant who has age and relevant work experience in his favor but low test scores may nevertheless be preferable to a young applicant who has high test scores but no relevant work experience. This is particularly likely to be the case if turnover rates are high for the productivity benefits of age and previous relevant work experience are large initially but diminish with time on the job. These results point to the desireability of studying the effects of test scores on job performance in the context of a multivariate model which includes controls for as many other factors as possible. They also remind us that employment tests should never be the sole criterion by which workers are selected. Tests would supplement not displace other criteria for selecting the best job candidate.

2.2 A REVIEW OF STUDIES OF OUTPUT VARIABILITY

The second determinant of the payoff to using tests to select workers is the extent of the variability across workers in their productivity on the job. A search for studies of output variability yielded 49 published and 8 unpublished papers covering 94 distinct jobs. The results are summarized in column 1 of table 7 and column 2 of table 6. (The detailed results are reported in Appendix tables 1 through 4). Most of the studies reviewed measured physical amounts of output produced over periods generally lasting one to four weeks and report a ratio of the standard deviation of output to mean output, coefficient of variation or CV. Relative output levels vary

over time, so coefficients of variation for a one or five year period are inevitably smaller than the coefficients of variation for a one or two week period. Hunter, Schmidt and Judiesch (1988) review a number of studies which provide evidence on the correlation between output levels over time and how these correlations vary with the length of the time interval studied. This information was then used to construct estimates of the output CVs for periods of a year or more. It is these corrected estimates of the CV which are reported. For semi-skilled factory jobs paid on an hourly basis the coefficient of variation averaged about 14 percent. Output variability is greater in the higher paid technical and precision production jobs. The coefficient of variation averages 27.6 percent in craft jobs and 33.8 percent in technical jobs.

Clerical jobs were divided into a high skill and low skill categories. The description of the job in the Dictionary of Occupational Titles was reviewed and jobs which appeared to require greater skill or involve discretion and decision making were classified as "high skill clerical." Jobs which were included in this category were stenographer, computer operator, administrative clerk, supply specialist, claims processor, head teller, ticket agent, customer service representative and teacher aide. Jobs categorized as "routine" were key punch operator, hotel clerk, cashier-checker, telephone operator, mail carriers, file clerks, stock clerk, typists, and toll ticket sorters. This distinction appears to be a real one for the high skill clerical jobs were generally better paid than the routine clerical jobs and the workers in these jobs scored one third of a standard deviation higher on the GATB academic achievement composite than those who occupied the more routine clerical jobs. Furthermore, the variability of job performance appears to be substantially greater in the jobs that require decision making. The coefficient of variation was 25.5 in the high skill clerical jobs and 16.7 percent in the routine jobs.

Data was available for only three service occupations. These three jobs represent too small a sample to produce reliable estimates of the CV for all service jobs except police and fire fighting so the estimate of the service CV employed in the paper is an unweighted average of the CVs for operatives, low skill clerical workers and 20.6, the average for the three service jobs for which there is data on the variability of output. For sale clerks records

of sales transactions were employed to calculate the CV and the result was an estimate of 29.8 percent.

When a firm expands by hiring extra workers, it incurs significant fixed costs. It must rent space, buy equipment, hire supervisors and recruit, hire, train, and payroll the additional production workers. If output can be increased by hiring more competent workers, all of these costs can be avoided and the firm's capital becomes more productive. These factors tend to magnify the effects of work force quality on productivity. They imply that the ratio of the standard deviation of worker productivity in dollars (SD$) to average worker compensation is much larger than the productivity CV for that job (Klein, Spady and Weiss 1983; Frank 1984).

Estimates of productivity standard deviations (SD$) in 1985 dollars are reported in column 2 of the table 6. In most cases the author of the study made no attempt to estimate SD$'s, so estimates of SD$ were derived as a product of the CV, the mean compensation for that job and 1.52, the ratio of value added to compensation for private non-farm business excluding mining, trade, finance and real estate. The value added to compensation ratio in retailing and in real estate was much too high to be used as an adjustment factor. So for all sales occupations it was assumed that SD$ = CV times average compensation. The SD$ that result are $13,668 for technicians, $12,399 for craft workers, $5062 for semiskilled factory jobs, $8925 for high-skill clerical jobs, $4934 for routine clerical jobs, $4068 for service workers other than police and fire fighters and $5228 for sales clerks. While it is possible to debate the accuracy of specific estimates and the reliability of the 15th, 50th, and 85th percentile method of measuring SD$, the basic pattern of rapidly increasing standard deviations of output as one move up in the occupational distribution is unlikely to be disturbed by new data or a revised methodology.

What about jobs where capital equipment controls the pace of work? It has been argued that in automated continuous process industries the amount and quality of output is determined by technology and computer programs not by the skills and talents of the workers. In fact, however, programs cannot be written to handle all contingencies and machines are never completely reliable so human operators have an important role to play (Hirschorn 1984; Adler 1986). In capital intensive industries with high rates of energy and

materials consumption, small errors can cause substantial losses. Small adjustments which increase fuel efficiency can save a utility or refinery millions of dollars a week. This has been demonstrated by a very careful study of the variability of the job performance of the operators of electric utility plants(see Table A2). In the study of the operators of electric generating plants commissioned by the Edison Electric Institute, committees of technical experts were organized and asked to make consensus estimates of the frequency and costs of the most common types of operator errors. Once the relationship between specific operator errors and the purchase costs of replacement power was established, the experts estimated what would be expected (in dollar terms) from an operator at the 15th, 50th and 85th percentile of job performance. The study concluded that the standard deviation for the productivity of control room operators is about $278,000 in 1985 dollars at nuclear plants and $115,000 at fossil fuel plants (Dunnette et al 1982).[1] When the results of Wroten's study of output variability among refinery operators is combined with the results of the Dunnette et al study, the estimated SD$ for this small but very important set of jobs is $91,020. The SD$ of plant operators is more than 6 times larger than any of the other occupations in the USES Individual Data File. As a result, resorting to maximize total output implies that workers who would be above average producers in all occupations should be assigned to this occupation.

2.3 SIMULATION RESULTS

The question posed in this section is "What will happen to aggregate output and to the gender and ethnic composition of various occupations, if firms are allowed and/or encouraged to use employment tests to select new hires?" To simulate the effect of changes in the allocation of workers across jobs on aggregate output, one needs estimates of how the effects of test scores and other worker characteristics on productivity vary across jobs. If the data were available, we would want to estimate, for random samples of the population, linear regressions in which the true relative productivity in dollars, $P^t_{ij}-P^t_j$, of the i^{th} worker in the j^{th} job is a function of the worker's characteristics. Unfortunately, in most studies the only indicators of productivity are supervisory ratings which are not defined on a ratio scale and have only limited reliability. If, however, outside estimates of the standard deviation of true productivity, $SD_j(P^t_{ij})$, are available and

assumptions are made about the measurement error in these ratings, estimates of the effect of test scores on true productivity in that occupation can be derived from regression models in which ratings are predicted by test scores and other worker characteristics. The measurement assumptions made by Hunter and Schmidt and most other contributors to the literature are:

$$(3)\quad \frac{R^m_{ij}-R^m_j}{SD_j(R^m_{ij})} = r_{pp}[(P^t_{ij}-P^t_j)/SD_j(P^t_{ij})] + v$$

where r_{pp} is the reliability of supervisory ratings (eg. the correlation between independent ratings by two different supervisors) and v is uncorrelated with true productivity. The upper bound on the reliability of job performance measures like the Standard Descriptive Rating Scale has been found to be .6 (King, Hunter and Schmidt, 1980). In other words, the ratings of relative job performance are assumed to be cardinal measures of productivity that are linearly related to true productivity and that errors in assessing productivity are negatively associated with true productivity. This assumption implies that measurement error in the dependent variable attenuates the true relationship and that the impact of a right hand side variable on true productivity in standard deviation units can be calculated by multiplying the coefficients reported in Table 5 by 1.29, the inverse of the square root of criterion reliability. It is further assumed that $SD_j(P^t_{ij})$ is equal to the $SD\$_j$, the standard deviation of productivity in dollars discussed in section 2.2. While these assumptions may seem reasonable, there do not appear to be any studies which have demonstrated even, in one particular case, that errors in assessing job performance are negatively correlated with true productivity and that $SD_j(P^t_{ij}) = SD\$_j$. On the other hand, there also appears to be no evidence that these assumptions are wrong. To facilitate comparisons with previous literature, the calculations of output effects presented below are based on the assumptions detailed above.

The second problem that must be dealt with is the fact that job performance outcomes have been used to select the sample used in the analyses. Since incompetent workers are fired or induced to quit and high performing workers are promoted to jobs of a higher classification, job incumbants are a restricted sample of the people originally hired for a job (Bishop 1988). The systematic nature of attrition from the job substantially reduces the variance of job performance and biases coefficients of estimated job

performance models toward zero. When all variables are multivariate normal, the ratio of the coefficients estimated in the selected sample to the true coefficient estimated in an unselected population is equal to:

$$(4) \quad \beta^*/\beta = VR/(1-R^2(1-VR)) = VR + R^{*2}(1-VR)$$

where VR is the ratio of the variance of y in the selected sample to its variance in the full population, R^2 is the multiple coefficient of determination of y on $\underline{x}$ in the full population and R^{*2} is the multiple coefficient of determination of y on $\underline{x}$ in the selected population (Goldberger 1981). Estimates of VR, the ratio of incumbant job performance variance to new hire job performance variance can be derived from the NCRVE employer survey analyzed in Bishop (1987, 1988). Using reported productivity in the 3rd through 13th week after being hired for two different workers as the data, a variance ratio was calculated by dividing job performance variance of incumbants (pairs of workers both of whom were still at the firm at the time of the interview a year or so after being hired) by the job performance variance of a group of very recent hires (pairs of workers both of whom stayed at least 13 weeks but who may or may not have remained at the firm through the interview). The resulting estimate of VR was .486. Assuming multi-variate normality and noting that the R^2 of the models in table 5 averages about .16, our estimate of β/β^*, the multiplier for transforming the coefficients estimated in the selected sample into estimates of population parameters, is 1.76.

The Productivity Loss from Random Assignment of Workers to Jobs

The first simulation exercise is a comparison of the mean predicted productivity of workers who currently occupy each job to the productivity that would result from a random assignment of new hires to jobs. The parameters of the first model were used to predict the productivity (in standard deviation units) during each of the first ten years on the job of all 31,399 workers in the data set in each of the 8 occupational categories analyzed. The effects of age and previous occupational experience at the time of hire were included along with test scores, schooling, gender and ethnicity. A present discounted value of each worker's predicted productivity during the first ten years was then calculated under the assumption of a 6 percent real interest rate and a monthly turnover rate of 1 percent. Based on occupation, race and Hispanic status, each worker was assigned a weight

so that the USES Individual Data File would become representative of all 71,132,000 workers in these 8 occupations (see Appendix Table B1 for a description of how these weights were derived). The weighted mean present value of predicted productivity resulting from random assignment of new hires to occupations was then subtracted from the weighted mean present value of predicted productivity during the first ten years on the job for the current set of individuals in that occupation. This was then translated into dollars by multiplying first by 1.29, second by 1.76 and then by the $SD\$_j$ for that occupation.

The results of this simulation exercise are presented in Table 6. The loss in productivity that would result from random assignment of workers to jobs is estimated to be about $1800 dollars per worker or 8 percent of mean compensation. The aggregate loss is $129 billion in 1985 dollars. The reductions in productivity primarily occur because: (1) workers who had higher than average productivity during their early years at the firm due to previous experience in the occupation are often randomly assigned to an occupation where this previous experience is of no value and (2) workers with high test scores are much less likely to be assigned to high skill jobs which use their talents than is the case currently. These results are clearly an extreme lower bound estimate of the benefits (relative to random assignment) of the current process of matching workers to jobs. If other worker characteristics such as occupationally specific education, tastes and talents for particular occupations and performance in previous similar jobs had been included in the model, estimates of productivity loss resulting from random assignment of workers to occupations would have been substantially greater.

Re-Sorting Workers on the Basis of Test Scores

The effect of greater use of employment tests to select workers on productivity was explored by simulating the effects of reassigning new hires on the basis of the productivity predictions derived from model 2. A present discounted productivity (for the first ten years after being hired) was calculated for each worker in each occupation. The 8 occupations were arrayed in a hierarchy according to the magnitude of the dollar change in productivity that results from a unit change in academic achievement. Plant operators were at the top of the hierarchy. The computer program sorted all workers

by the present discounted value of their predicted productivity as plant operators (based on model 2) and then assigned just enough people from the top of that ranking to fill all 228,000 of the nation's plant operator jobs. The remaining workers were then sorted by their productivity in technical occupations and those found at the top of the ranking were assigned to these occupations until all 5,261,000 technical jobs were filled. This procedure was repeated next for craft jobs, then for high skill clerical jobs, for low skill clerical jobs, for service jobs, and for operative jobs. Those left over after operatives were selected became sales clerks.[11] The simulated effects of this reassignment scheme on productivity are presented in Table 7. Output rises by $1561 per worker or by 6.9 percent of mean compensation. The total gain from applying this plan to the 71 million workers represented in the data base is $111 billion. There are major improvements in the productivity of plant operators, technicians and craft workers which more than offset large declines in the productivity of operatives and sales clerks.

The simulated effect of the reassignment scheme on the mean test scores, schooling and demographic character of each occupation is presented in the even numbered columns of Table 8. The characteristics of those who are crrently in each occupation are presented in the odd numbered columns. Currently workers in technical and high skill clerical occupations have the highest academic achievement and operatives and service workers have the lowest. The simulation results in the workers with the stongest academic achievement being reassigned to plant operator, technical and craft occupations and the workers with the weakest academic achievement being reassigned to operative and sales clerk occupations. Some of the changes are truly dramatic--the mean test score of plant operators rises by 2 population standard deviations and the mean score of sales clerks falls by 1.6 population standard deviations. This outcome is a result of placing the plant operator occupation at the top of the hierarchy and the sales clerk occupation at the bottom. The simulation also produces an increase in the schooling of plant operators and a decline in the mean schooling of sales clerks.

Reassigning workers on the basis of test scores, age and previous work experience but not gender or ethnicity produces large changes in the demographic composition of some occupations. Women end up with most (77 percent) of the plant operator jobs and roughly half of the craft jobs.

Occupations which have historically been predominantly female become more evenly split between men and women. As anticipated, black representation decreases in plant operator, technical, craft, clerical and service occupations and increases in operative and sales clerk occupations. Similar but more modest changes occur for Hispanics. Since, however, employers know the minority status of job applicants, the adverse impact on minorities of using tests to select employees can be eliminated by within-group scoring of the tests or by other affirmative action efforts.

How do these results compare to those of Hunter and Schmidt (1982)? The estimated total effect of going from random selection of new hires to optimal use of tests, age and previous work experience is 15 percent of the compensation of workers subject to reassignment. This is much larger than the 8 percent figure H/S obtain in their three test score selection model when SD$ is 40 percent of each occupation's mean compensation. The reasons for the discrepancy are: (a) the estimates of differences in SDY across occupations are much larger than the one's assumed in their simulation, (b) the restriction of range correction (which was based on actual data on the reductions in job performance variance resulting from the selective nature of turnover) is larger than the one they assumed, (c) job assignment is based on a composite of test scores, schooling, age and previous occupational work experience that has greater validity than test scores alone and (d) 8 rather than 4 occupational categories are analyzed.

2.4 A CRITIQUE OF THE SIMULATIONS

The simulation results just presented are based on a maintained assumption that the models of relative job performance described in section 2.1 (which were estimated in samples of job incumbents) are, after the correction for errors in measurement and restriction of range, unbiased estimates of true population relationships. This assumption is almost certainly incorrect and this inevitably results in the findings of the simulation exercise being biased as well. The underlying performance model is biased for two reasons: omitted variables and the selection process that determines which members of the population are hired for the job.

While model 1 is a more complete specification than is typically found in the literature, it lacks controls for important characteristics of the worker which are often known by hiring decision makers and which are

associated with worker productivity. Examples of things left out of the model are occupationally specific schooling, grades in relevant subjects in school, reputation of the school, the amount and quality of previous on-the-job training, performance in previous jobs, interview performance, physical strength and a desire to work in the occupation. Quite clearly, if random assignment of new hires to jobs involves ignoring all of this additional information as well as information on schooling and years of experience in the occupation, the loss in productivity would be substantially larger than the numbers reported in table 6.

The omission of so many important determinants of job performance also biases the simulations of the impact of greater test use. If these variables had been included in the job performance models, the coefficients on test scores would probably have been smaller and adding test scores to the factors considered in hiring selections would have resulted in fewer workers being reassigned. This in turn reduces the output gain that results from greater use of employment tests for selection and exaggerates the predicted changes in demographic composition of occupational work forces.

The other source of problems is selection effects. The selectivity bias caused by turnover and promotion decisions that depend on realized levels of job performance has already been discussed and corrected for. Another form of selectivity bias is introduced by the selection that precedes the hiring decision. If hiring selections were based entirely on X variables included in the model, unstandardized coefficients such as ß would be unbiased and correction formulas would be available for calculating standardized coefficients and validities. Unfortunately, however, incidental selection based on unobservables such as interview performance and recommendations is very probable (Thorndike 1949; Olson and Becker 1983; Mueser and Maloney 1987). In a selected sample like accepted job applicants, one cannot argue that these omitted unobservable variables are uncorrelated with the included variables that were used to make initial hiring decisions and, therefore, that coefficients on included variables are unbiased. When someone with 10 years of formal schooling is hired for a job that normally requires an associates degree, there is probably a reason for that decision. The employer saw something positive in that job applicant (maybe the applicant received a particularly strong recommendation from previous employers) that led to the

decision to make an exception to the rule that new hires should have an associates degree. The analyst is unaware of the positive recommendations, does not include them in the job performance model and, as a result, the coefficient on schooling is biased toward zero. This phenomenon also causes the estimated effects of other worker traits used to select workers for the job such as previous relevant work experience to be biased toward zero. Variables which were not used to select new hires such as the GATB test scores will probably have a positive correlation with the unobservable. Since the unobservable probably has its own independent effect on job performance (ie it is not serving solely as a proxy for test scores), test score coefficients are likely to be positively biased. Mueser and Maloney (1987) experimented with some plausible assumptions regarding this selection process and concluded that coefficients on education were severely biased but that test validities were not substantially changed when these incidental selection effects are taken into account.

Consequently, the estimates of the effects of greater test use presented in Table 7 almost certainly exaggerate its true effect. If the simulations had been conducted using the true structural model of job performance rather than the biased one that was available, many fewer people would have been reassigned and productivity gains would have been smaller. Still another problem with the simulations is that they took no account of turnover risks. The large effects of tenure on the productivity of plant operators, technicians and craft workers implies that specific training is particulary important in these occupations and that minimizing turnover should be an important goal of a firm's hiring selections. Some of the workers assigned to plant operator jobs in the simulation might have been college students working part time who would have been unlikely to remain long in the job. As Mike Rothschild has argued there are countless barriers to the complete reshuffling of the work force that would be necessary for employment testing to have its maximum effect (the effect that is simulated in Table 7). Employers would have to become much better informed about employment testing. If they all sought advice from industrial psychologists, long queues would result and consulting fees would skyrocket. Some test batteries are expensive to administer. If a number of worker aptitudes are to be reliably measured, a couple of hours must be devoted to the testing. This would impose a very serious burden on

job seekers in some labor markets and many low wage industries would, consequently, eschew testing altogether. The simulation model did not ask the workers who were being transferred whether they wanted the higher paying jobs. Some would have refused. The simulation ends gender segregation of occupations and makes wholesale transfers of clerical workers to plant operator and craft jobs. Improved structural models would probably reduce the size of these shifts, but even more modest shifts would be difficult to pull off. Affirmative action goals and/or the use of race normed test scores in selection would also reduce the sorting impacts of greater test use. Clearly, the EEOC regulation of employment testing is not the only barrier to a more efficient allocation of workers across jobs and many of these other barriers would have to fall before testing could have its full effect. Consequently, the likely productivity benefits and resorting effects of allowing employers a free hand with regard to employment testing are much smaller than those presented in Table 7.

The simulated effects of substituting random selection of new hires for the current job-worker matching system reported in Table 6 are, by contrast, gross underestimates of the true costs. The selected nature of the sample and the many variables omitted from the "structural" models of job performance, cause very large biases in these simulations. Depending on how far one goes down the road toward random selection, the loss in sorting efficiency might be 2 or even 4 times those estimated.[12] Rates of involuntary seperation would increase and this would increase unemployment. In addition, economic incentives to go to school and study hard would be greatly reduced and this would cause further reductions in total output and standards of living. These results suggest that the current system of matching workers to jobs which makes almost no use of tests (tests were given prior to hiring in only 3.2 percent of hiring events sampled in the NFIB study) is not doing all that bad a job. This conclusion would appear to contrast somewhat with Hunter and Schmidt's (1983) characterization of current selection processes quoted at the beginning of part 2.

On the more important issue of how increased employment testing will effect national output, there is no real disagreement with Hunter and Schmidt. The simulations imply that the improvements in the matching of workers to jobs resulting from increased employment testing will significantly increase

output. The 6.9 percent figure might fall to 2 or 3 percent of employee compensation once one takes the biases and the barriers to optimal use of tests into account. On the other hand, taking constraints off the use of tests will also reduce tryout hiring and turnover, increase investment in specific human capital and reduce aggregate unemployment (ie. the rate of unemployment at which inflation begins to accelerate will fall if the minimum wage constraint is not binding). These effects were not part of the simulations. Since, however, total compensation of labor will exceed $3 trillion in 1988, extrapolation of these simulations to all workers implies that the productivity gain from unconstrained employment testing would eventually increase national income by 60 to 90 billion dollars and maybe much more when the unemployment response is factored in. These effects would not arrive suddenly for the tests only influence hiring decisions. Current employees would not be fired and replaced by new hires selected on the basis of tests because the gains from better selection will seldom be sufficient to justify firing employees who have developed extensive firm specific knowledge. It would, therefore, be a generation before the full effect of testing on the allocation of workers to jobs would be realized.

The $60 to $90 billion estimate is clearly a guess. A better estimate of the effect of greater test use on sorting efficiency requires better estimates of SDY, a better understanding of the magnitude and nature of the biases in job performance models, a model of employment testing's effects on turnover and unemployment and above all an understanding of how employers would use tests if they were given the opportunity. Clearly, much more research is needed on these topics.

PART III. POLICY RECOMMENDATIONS

The findings presented in the first two parts of the paper imply that improved signaling of worker skills and competencies to employers will probably have significant positive effects on productivity and standard's of living. Productivity gains occur both because more valid selection procedures improves the match between workers and jobs and because the supply of workers with the talents measured by the tests rises in response to the increase in labor market rewards for the talents. The distributional consequences of greater

test use are that it benefits women but tends to lower the representation of Blacks and Hispanics in occupations where the payoff to cognitive skills is particularly high such as plant operator, craft worker and technician.[13]

This adverse impact can be avoided, however, by race norming the test scores (as the GATB currently does) and affirmative action. Consequently, impacts on minority groups should not be the basis for deciding whether to use an employment test or which test to use. Other instruments are available for achieving employer and societal goals regarding integration on the job and the representativeness of a firm's workforce. When, however, it comes to generating incentives to develop the skills needed on the job and efficient matching of workers with talents to jobs, there appears to be no other selection instrument that will sort efficiently while generating the correct incentives. These are the two criteria by which alternative employee selection policies should be evaluated. That is the task undertaken in the remainder of the paper.

Sorting efficiency will tend to be maximized when employment tests are part of the selection process for jobs in which the particular competencies measured by the tests have a particulary high productivity payoff. In other words, effort should be made to maximize differential validity. Tests should be used but they should supplement not displace consideration of occupationally relevant training and experience. If most of the people hired into an entry job move up to other more responsible positions, the criteria applied at the port of entry needs to take the higher level jobs into account.

The analysis presented in the first part of the paper implies that student incentives to learn and parental incentives to demand a quality education are maximized when the following is true: (1) <u>significant</u> economic rewards depend directly and <u>visibly</u> on academic accomplishments, (2) the accomplishment is defined relative to an <u>externally imposed standard of achievement</u> and not relative to one's classmates, (3) the reward is received immediately, (4) everyone, including those who begin high school with serious academic deficiencies, has an <u>achievable goal</u> which will generate a significant reward and (5) progress toward the goal can be monitored by the student, parents and teacher.

We will see shortly that it is not easy to design a system of signaling and certifying academic achievement which satisfies all of these requirements.

Consequently, it will generally be desireable to use more than one signal of academic achievement and to use different signals when selecting for different jobs. Let us examine the alternatives.

Diplomas:

From the point of view of incentives, the standard high school diploma satisfies requirement 2, 3, 4 and 5 but it fails to satisfy requirement # 1, the most critical requirement of all. Minimum competency tests for receiving a high school diploma are an improvement, for they are an example of an externally imposed standard of achievement. They are a step in the right direction, especially when they are taken early in high school, and remedial classes are offered after school and during the summer for those who fail on the first try. However, some students arrive in high school so far behind that setting a high minimum would cause many to give up trying. Consequently, the minimum standard is not set very high and fails to challenge most students.

Schooling is a valid predictor of job performance but to a great degree its validity derives from its correlation with test scores. The evidence on its incremental contribution to validity once test scores are controlled is more mixed. The estimations reported in Table 5 found very weak effects of schooling but this is probably an artifact of the selection biases discussed in section 2.4 (Mueser and Maloney 1988). Selection into the military is based explicitly on the test scores and high school graduation, not on unobservables as in the civilian sector. Since selection is based on X variables, selection effects can be corrected for (Dunbar and Linn 1986). Analysis of military data finds that high school graduation has its own unique impacts when test scores are controlled.

Weiss's (1985) study of Western Electric employees found that completing high school is a valid predictor of low absenteeism and low turnover but not job performance. Thus even when studies find that graduating from high school has little effect on job performance, it appears to effect retention. Consequently, from a sorting efficiency point of view, the high school diploma probably belongs on the list of credentials considered by employers even when test scores are available.

Competency Profiles:

Competency profiles are check lists of competencies that a student has developed through study and practice. The ratings of competence that appear on a competency profile are relative to an absolute standard, not relative to other students in the class. By evaluating students against an absolute standard, the competency profile prevents one student's effort from negatively affecting the grades received by other students. It encourages students to share their knowledge and teach each other.

A second advantage of the competency profile approach to evaluation is that students can see their progress as new skills are learned and checked off. The skills not yet checked off are the learning goals for the future. Seeing such a check list getting filled up is inherently reinforcing.

With a competency profile system, goals can be tailored to the student's interests and capabilities, and progress toward these goals can be monitored and rewarded. Students who have difficulty in their required academic subjects can, nevertheless, take pride in the occupational competencies that they are developing and which are now recognized just as prominently as course grades in academic subjects. Upon graduation, the competency profile would be encased in plastic and serve as a credential certifying occupational competencies. If the ratings by teachers (and the sponsoring employers of cooperative education students) are reliable indicators of competence, employers will find this information very valuable, and the students who build a good record will be handsomely rewarded. I am not aware of any studies of the validity of school developed competency profiles.

Hiring Based on Grades in High School:

Using grades to select new hires results in a very visible dependence of labor market outcomes on an indicator of academic accomplishment. There are, however, two disadvantages. It results in zero-sum competition between classmates and consequently contributes to peer pressure against studying and parental apathy about the quality of teaching and the rigor of the curriculum. The second problem is that it induces students to select easy courses which tend to cause grade inflation. These problems can be mitigated somewhat if employers take the rigor of courses into account when evaluating grades, give preference to schools with tough grading standards, and vary the number hired from particular schools in response to the actual job

performance of past hires from that school.

From the employer's point of view, the disadvantage of high school GPA is that it is difficult to adjust these grades for the grading standards of the school but without such adjustment grades may have rather low validity. Most of the published studies of the validity of grades probably used information that had been collected by the firm when hiring decisions were being made. As a result, most of the validity coefficients reported for grades are probably negatively biased by the selection effects discussed in section 2.4.

Job Tryout and Promotions Based on Performance:

From the point of view of motivating students to study, the problem with job tryout and performance reward systems is that the dependence of labor market outcomes on academic achievements is both invisible and considerably delayed.

From the employer's point of view, the disadvantages of job tryout are the costs of training workers who end up being fired, its unpopularity with workers who will spend months unemployed if they are fired, and its potential for generating grievances.[14] Performance evaluations are known to be unreliable, and this makes workers reluctant to take jobs in which next year's pay is highly contingent on one supervisor's opinion. Pay that is highly contingent on performance can also weaken cooperation and generate incentives to sabotage others. The benefits of performance reward systems are that they motivate better performance, they tend to attract high performers to the firm, and they tend to induce the high performers to stay at the firm. When these factors are balanced, it appears that most workers and employers choose compensation schemes in which differentials in relative productivity result in relatively small wage differentials (Bishop 1987).

IQ Tests:

Students, parents and teachers view IQ tests as measuring something that schools do not teach. Even though this public perception is not entirely correct, the perception is not likely to change in the near future, so hiring on the basis of IQ tests fails requirement # 1. Students will not see the connection between how hard they study and higher IQ scores. Other problems with this approach are that IQ tests are less valid than broad spectrum

achievement tests like the ASVAB and individually administered IQ tests are too expensive and cannot be made secure.

Job Knowledge Tests:

From the point of view of learning incentives, the disadvantage of job knowledge tests is that they do not generate incentives to study math, science, history and literature and may induce students to over-specialize in school. If at some point in their career a job in the field for which they prepared is not available, they are left high and dry.

From the point of view of sorting efficiency, job knowledge tests have much to recommend them for they maximize differential validity. They are particularly appropriate if the applicants vary in their knowledge and background in the occupation and training costs are substantial. If new hires are likely to be quickly promoted into higher level jobs, the job knowledge test should also cover the skills required in these jobs. Job knowledge tests are less useful when none of the applicants has experience in the field and training costs are low. The possibility of court challenges complicates matters here for validity generalization may not apply, and each job knowledge test may have to stand on its own merits. Development costs are high, so small occupations may never have job knowledge tests developed for them.

Broad Spectrum Achievement Tests:

From the point of view of incentives to study a broad range of academic subjects, broad spectrum achievement tests such as the ASVAB are the best of the alternatives reviewed so far. If some of the subtests in the battery include material covered in the standard college prep high school curriculum such as algebra, statistics, chemistry, physics and computers, the use of such tests for selection would generate parental pressure for an upgraded curriculum and encourage high school students to take more rigorous courses. When many employers use achievement tests to select new employees, everyone who wants a good job faces a strong incentive to study, and those not planning to go to college will find the incentive especially strong. The best paying firms will find they can set higher test score cutoffs than low paying firms, so the reward for learning will become continuous. Whether one begins 9th grade way behind or way ahead, there will be a benefit on the margin to studying hard for it will improve one's job prospects.

Broad spectrum achievement tests covering science, computers, mechanical principles, and technology as well as mathematics, reading and vocabulary are the preferred method of assessing general cognitive skills from the employer's point of view as well. Test batteries which cover the full spectrum of knowledge and skills taught in high school are more valid predictors of job performance than tests which assess math and verbal skills only. Evidence for this statement comes from examining the relative contributions of various subtests to the total validity of the ASVAB battery. Sims and Hiatt's (1981) analysis of the job performance of 23,061 Marine recruits found, for example, that validity (corrected for restriction of range) was .38 for auto shop information, .43 for mechanical comprehension, .42 for electronics information, .46 for general science, .42 for word knowledge, .50 for mathematics knowledge, and .48 for arithmetic reasoning. Tests measuring electronics, mechanical, automotive and shop knowledge--material that is generally studied only in vocational courses--have high validity. Analyzing this and other military data sets, Hunter, Crosson and Friedman (1985) concluded that the "general cognitive ability" construct that best predicted performance in all military jobs included subtests in general science, electronics information, mechanical comprehension and mathematics knowledge as well as conventional word knowledge and arithmetic reasoning subtests. The addition of these four subtests to the construct increased validity by 9 percent and the proportion of true job performance variance explained from .306 to .364. They also found that auto and shop information significantly improved the prediction of performance in jobs in the mechanical job family. I suspect that tests measuring understanding of statistics, business, economics, marketing and psychology would similarly improve the validity of batteries used to select workers for most white collar jobs in the private sector.

Will the courts allow firms to use broad spectrum achievement tests covering subjects not offered until the final years of high school? My fear is that, since the research on test validity in the civilian sector has used the GATB almost exclusively, everyone may be forced to use reading, vocabulary, and arithmetic reasoning tests that are demonstrably similar to their GATB counterparts. If the studies of the ASVAB's validity in predicting performance in military jobs were not accepted as evidence for similar jobs in the civilian sector, it might be a decade before tests measuring general science and

electronics knowledge could be used as a general selection device for blue collar jobs. Courts might require that employers demonstrate that each item on a science test have a specific application in each job for which it is a proposed selection device. To avoid having to redesign the test for each job, test developers would dumb the test down and include only simple questions covering scientific principles that are learned in grade school. Costly validity studies covering tens of thousands of workers might be necessary before broad spectrum achievement tests covering the material included in rigorous high school courses have their validity generalized and become available as selection tools.

To maximize the incentive effects, it is essential that students, parents and teachers be aware that local employers are using tests for selection and what kind of material is included on these tests. Employers should seek out ways of publicizing their use of broad spectrum achievement tests. Unfortunately, the fear of litigation may cause employers to give only limited publicity to their use of tests and so constrain the type of tests that are used that many of the potential beneficial incentive effects of employment testing may never be realized.

Performance on Achievement Exams Taken at the End of Secondary School

In Japan and most European countries, the educational system administers achievement test batteries (eg. the 'O' and 'A' Levels in the UK, the Baccalaureate in France) which are closely tied to the curriculum. These are not minimum competency exams. Excellence is recognized as well as competence. In France, for example, students who pass the Baccalaureate may receive a "Tres Bien", a "Bien", an "Assez Bien" or just a plain pass. Most job applications request information on which exams were taken and what scores were obtained on each exam. These exams generate credentials which signal academic achievement to all employers and not just the employers who choose to give employment tests. The connection between one's effort in school and performance on these exams is clearly visible to all. Consequently, school sponsored achievement exams like those used in Europe would have much stronger incentive effects than employer administered broad spectrum achievement tests.

This approach to signaling academic achievement has a number of advantages. Because it is centralized and students take the exam only once,

job applicants do not have to take a different exam at each firm they apply to and the quality and comprehensiveness of the test can be much greater. There is no need for multiple versions of the same test and it is much easier to keep the test secure. By retaining control of exam content, educators and the public influence the kinds of academic achievement that are rewarded by the labor market. Societal decisions regarding the curriculum (eg. all students should read Shakespeare's plays and understand the Magna Charta) tend to be reinforced by employer hiring decisions. Tests developed solely for employee selection purposes would probably place less emphasis on Shakespeare and the Magna Charta.

The disadvantages of schools administering the achievement exams is that students have only one chance to demonstrate their competence. If one has an off day, a year typically must pass before the exam can be retaken. With employer administered exams, having an off day is less damaging for one will shortly have a chance to do better at another employer. Employers may also find it is easier to compare job applicants who have all taken the same employer administered exam.

With regard to validity, there is probably little to choose between the two systems. Separate scores are reported for each subject so employers may focus on the tests which have special relevance to their jobs. School administered tests are more reliable measures of achievement because they sample a much larger portion of the student's knowledge of the field (the ASVAB General Science subtest, by contrast, allows the student 11 minutes to do 24 items). They may also be more valid because they are not limited to the multiple choice format. Thus, even though the topics covered in the school exam are probably less relevant to the firm's jobs, the school exam is probably just as valid a predictor of job performance as a specially designed employment test.

FOOTNOTES

1. Gottfredson (1988) argues against race norming on the grounds that it will produce a white backlash and is counter to democratic principles. I did not, however, find her argument persuasive. In the first place, there are equities on both sides of the issue. Many whites support special preferences in employment for blacks to try to undo the effects of centuries of slavery and discrimination. There may be instances in which affirmative action has produced a backlash but for most whites it is a low priority concern. In any case, a person's views on this issue are determined largely by values, not scientifically verifiable facts.

2. If Americans born between 1897 and 1901 are compared to those born 50 years later, the proportion born on a farm fell from 42.4 percent to 10.6 percent, the proportion growing up in a broken family fell from 17 to 13 percent, the average number of siblings fell from 4.8 to 3.3, and father's average years of schooling rose from 6.9 years to 10.7 years. (Hauser and Featherman, 1976). Time spent in school has increased dramatically. Between 1890 and 1960, the average length of the school term increased 19 percent, average daily attendance rates rose 40 percent and mean years of schooling completed increased more than 50 percent.

3. Estimates of how IQ changed as a cohort aged can be obtained by comparing young adults in the WAIS standardization to the people 25 years older in the WAIS-R standardization. This comparison suggests that for Verbal IQ the 20-29 year olds in 1953/54 gained 4.7 points (.315 SDs) in 25 years and the 30-39 year olds gained 2.85 points (.19 SDs). For Full Scale IQ, which contains the nonverbal performance tests all of which are timed, small declines occurred: -0.2 IQ points (-.013 SDs) for the 20-29 year olds and -2.3 IQ points (-.16 SDs) for the 30-39 year olds.

4. The Wells Alpha scores were translated into Army Alpha scores using a table of percentile equivalents developed by Lorge (1936). Means were then calculated and compared to the mean of the WWI army recruits using the SD of the WWI sample for a metric. The resulting estimate of the gain between WWI and WWII is smaller than that reported by Tuddenham. This estimate of the GIA gain for army recruits may well underestimate the GIA gain for all youth. The WWI army sample appears to have been more selected. Veterans of WWII accounted for 77 percent of all males 20-29 in 1947, while only 38.5 percent of the men 20-29 in 1920 were veterans of WWI. The sample used by Tuddenham was selected on the basis of AGCT scores to yield a distribution like that of inductees entering during 1943. During much of 1943 there were no limitations on recruitment of illiterates. (Department of the Army, 1965) Illiterate recruits included in the representative sample of WWII recruits took the Wells Alpha along with everyone else. During WWI 25 percent of all recruits and 17 percent of white recruits either could not read or write or had fewer than four or six (depending on the army processing center) years of schooling and took the Beta exam instead of the Alpha. They were, therefore, not part of the WWI sample.

5. The ITED was revised six times between 1942 and 1985 to incorporate

changes in curriculum. Scores on the new versions of the test were made equivalent to the old by administering the old and new test to large samples of students (in a way that insured random assignment of test version to individual students) and then equating them using the equal percentile method. Trends were translated into population standard deviation units by multiplying scores standardized using the sample standard deviation for Iowa seniors by .74.

6. Similar results are obtained for other test combinations. For example, alternative form reliabilities average .75 for 7 SRA subtests and .87 for the G aptitude of the GATB. The G aptitude of the GATB has an average correlation of .7 with the 7 SRA subtests, .75 with the WAIS verbal IQ, .78 with the ITED composite and .81 with the ACT composite (Hunter, Crosson & Friedman 1985; Department of Labor 1970). Jencks and Crouse (1982) make this same point in an article that proposes using achievement tests rather than aptitude tests for college admissions decisions.

7. For 12th graders such an improvement is approximately equal to 3.5 grade equivalents. By reporting the percentage changes in labor market outcomes that result from a one standard deviation change in GPA or performance on a test, we make the results of studies done on very different cohorts of workers comparable over time and understandable to the layman.

8. Studies that measure output for different workers in the same job at the same firm, using physical output as a criterion, have found that the standard deviation of output varies with job complexity and averages about .164 in routine clerical jobs and .278 in clerical jobs with decision making responsibilities(Hunter, Schmidt and Judiesch 1988). Since there are fixed costs to employing an individual (facilities, equipment, light, heat and overhead functions such as hiring and payrolling), the coefficient of variation of marginal products of individuals will be considerably greater (Klein, Spady, and Weiss 1983). On the assumption that the coefficient of variation of marginal productivity for clerical jobs is 30 percent[1.5*(.33*.278+.67*.164)], a .5 validity for general mental ability implies that an academic achievement differential between two individuals of one standard deviation (in a distribution of high school graduates) is associated with a productivity differential in the job of about 11 percent (.5*.74*30%). The ratio of the high school graduate test score standard deviation to the population standard deviation is assumed to be .74. For a more thorough discussion of the evidence on this issue see Bishop 1987b.

9. The formula was $SD(R^m_j) = (R^m_{ij}-R^m_j)^2/N-1$. Occasionally employers who had only 2 or 3 employees gave them all the same rating. Consequently, a lower bound of 40 percent of the mean $SD(R^m_j)$ was placed on the value the SD could take. Models were also estimated which did not standardize job performance variance across firms and which instead standardized the variances only across the occupation. None of the substantive findings were changed by this alternative methodology.

10. Mueser and Maloney (1988) argue persuasively that since schooling is a very important factor in the selection process, the coefficients on

schooling in estimations like these are negatively biased estimates of true population relationships. This argument probably applies as well to the coefficients on work experience in the occupation but not at the firm.

11. This hierarchical process for allocating new hires to jobs is not fully optimal. Some workers will not be assigned to the occupation in which they have the greatest comparative advantage. A computer program that assigns all new hires optimally would be much more complex. Given the biased nature of the underlying models of job performance it is not clear that the extra investment in programming time would be worthwhile.

12. The legal theories that have been used to attack employment tests on EEO grounds are equally applicable to other selection criteria. If the theory of differential validity by subgroup and employer were applied to selection criteria like years of schooling, school reputation, GPA and recommendations from previous employers (all of which have adverse impact), these criteria would probably fail court tests for jobs like those in the GATB Revalidation data. If the 1970s trend of court decisions restricting employer prerogatives to select the "best" job applicant had continued rather than being reversed, we might have moved a considerable distance down the road toward random selection of new hires for these jobs. Sandra Day OConnor's concurring opinion in Watson (1988) signals a major shift in the application of the Griggs adverse impact test, so the trend now seems to be in the direction of greater freedom for the employer.

13. This adverse impact results not because tests are unfair but because academic achievement contributes to worker productivity and because there are, unfortunately, real differences in mean levels of academic achievement between groups (Jones 1988). The tests are giving us the unhappy news that educational opportunities and achievement have not been equalized. The cause of the situation is the low quality of the education that so many Blacks and Hispanics received in segregated schools. Progress has been made in reducing these quality differentials and achievement gaps are diminishing. This means the problem will diminish over time, but racial differences in the average academic achievement of adults will not disappear quickly.

14. Mueser and Maloney (1987) develop a model of job tryout hiring which they claim implies that it may be efficient to ignore available information on stable worker competencies signaled by high test scores. They apparently do not recognize that the model also implies that information on education and previous work experience should also be ignored. They acknowledge that "Although employing applicants for long enough to observe performance entails costs of training and lost productivity, it may increase the incentives workers have to apply effort to learning their jobs by enough to compensate for such costs." In fact, however, turnover costs are so large--training costs are generally about one month's wages and fired workers suffer a couple of months of unemployment--, that a sequential decision strategy will always dominate the strategy they consider. It will hardly ever be optimal to hire ten people for one position and then fire 9 of them after a tryout. In any job requiring even a modest amount

of specific training or transitional unemployment, the optimal strategy is to use all the inexpensive information available to make an initial selection and then to give those selected a tryout but to plan on seldom having to fire the new employee. It is true, however, that the option of firing the worst performers results in Brogden's formula overstating the private benefits of a selection method.

BIBLIOGRAPHY

Adler, Paul. "New Technologies, New Skills, California Management Review, Volume XXIX, Number 1, Fall 1986.

Bayley, Nancy. "On the Growth of Intelligence." The American Psychologist, 10 (1955) pp. 805-818.

Bishop, John H. Preparing Youth for Employment. Columbus, Ohio: National Center for Research in Vocational Education, 1985.

Bishop, John H. "The Recognition and Reward of Employee Performance." Journal of Labor Economics. October, 1987a.

Bishop, John H. "Match Quality, Turnover and Wage Growth". Center for Advanced Human Resources Studies Working Paper # 88-03, Cornell University, Ithaca, New York, 1988.

Bishop, John and Griffin, Kelly. Recruitment, Training and Skills of Small Business Employees, (National Federation of Independent Business Foundation, Washington, DC, forthcoming).

Boudreau, J. W. "Utility Analysis Applied to Human Resource Productivity Improvement Programs." 1986, To appear in M.D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology, (2nd edition).

Brogden, H. E. "When Testing Pays Off." Personnel Psychology, 1949, Vol. 2, pp. 133-154.

Brown, Charles. "Estimating the Determinants of Employee Performance." Journal of Human Resources, Spring 1982, Vol. XVII, No. 2, pp. 177-194.

Brossiere, M.; Knight, J. B. and Sabot, R. H. "Earnings, Schooling, Ability, and Cognitive Skills." American Economic Review, December 1985, Vol.75, Vol. 5, pp. 1016-1030.

Chiswick, Barry R. "The Effect of Americanization on the Earnings of Foreign-born Men." Journal of Political Economy, 1978, Vol. 86, No. 5. pp. 897-921.

College Board. The College Board Technical Handbook for the Scholastic Aptitude and Achievement Tests. College Entrance Examination Board, Princeton New Jersey, 1984.

College Board. ATP Guide for High Schools and Colleges. College Entrance Examination Board, Princeton New Jersey, 1987.

Department of The Army. Marginal Man and Military Service. December, 1965. pp. 1-270.

Department of Labor. General Aptitude Test Battery Manual 1970, United State Department of Labor, Manpower Administration.

Dunbar, Stephen B. and Linn, Robert L. "Range Restriction Adjustments in the Prediction of Military Job Performance". Prepared for Committee on the Performance of Military Personnel, National Research Council / National Academy of Sciences. September 1986.

Dunnette, Marvin D.; Rosse, Rodney L.; Houston, Janis S.; Hough, Leaetta M.; Toquam, Jody; Lammlein, Steven; King, Kraig W.; Bosshardt, Michael J. and Keyes, Margaret. "Development and Validation of an Industry-Wide Electric Power Plant Operator Selection System," Personnel Decisions Research Institute, Minneapolis, Minnesota, Report submitted to Edison Electric Institute, 1982.

Finch, F. H. "Enrollment Increases and Changes in the Mental Level of the High-School Population." American Psychological Association, June 1946.

Flynn, James R. "The Mean IQ of Americans: Massive Gains 1932 to 1978." Psychological Bulletin, 1984, Vol. 95, No. 1, 29-51

Flynn, James R. "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure." Psychological Bulletin, 1987. Vol. 101, No. 2, 171-191

Forsyth, Robert A. (personnel communication) "Achievement Trends for the Iowa Tests of Educational Development in Iowa: 1942-1985," Iowa City:Iowa Testing Programs, 1987.

Freeman, Richard B. The Market for College-Trained Manpower: A Study in the Economics of Career Choice. Cambridge, MA: Harvard University Press, 1971.

Freeman, Richard. The Overeducated American. New York: Academic Press, 1976.

Friedman, Toby and Williams, E. Belvin. "Current Use of Tests for Employment." Ability Testing: Uses, Consequences, and Controversies, Part II: Documentation Section, edited by Alexandra K. Wigdor and Wendell R. Gardner. Washington, DC: National Academy Press, 1982, pg. 999-169.

Gardner, John A. Influence of High School Curriculum on Determinants of Labor Market Experience. Columbus: The National Center for Research in Vocational Education, The Ohio State University, 1982.

Ghiselli, Edwin E. "The Validity of Aptitude Tests in Personnel Selection." Personnel Psychology. 1973: 26, 461-477.

Goldberger, Arthur S. "Linear Regression after Selection." Journal of Econometrics. Vol.15, 1981, pp. 357-366.

Goodlad, J. A Place Called School. New York: McGraw-Hill, 1984.

Gordon, R. A. (In Press). IQ Commensurability of Black-White Differences in Crime and Delinquency. Personality and Individual Behavior.

Gottfredson, Linda. "Reconsidering Fairness: A Matter of Social and Ethical Priorities." Journal of Vocational Behavior, Vol. 31, No. 3, December 1988.

Griggs vs. Duke Power Company, 3 FEP 175 (1971).

Hause, J. C. "Ability and Schooling as Determinants of Lifetime Earnings, or If You're So Smart, Why Aren't You Rich." In Education, Income, and Human Behavior, edited by F. T. Juster. New York: McGraw-Hill, 1975.

Hauser, Robert M. and Featherman, David L. "Equality of Schooling: Trends and Prospects." Sociology of Education, 1976, Vol. 49, April, 99-120.

Hauser, Robert M. and Daymont, Thomas M. "Schooling, Ability, and Earnings: Cross-Sectional Evidence 8-14 years after High School Graduation." Sociology of Education, July 1977, 50, 182-206.

Hotchkiss, Lawrence. Effects of Schooling on Cognitive, Attitudinal and Behavioral Outcomes. Columbus: The National Center for Research in Vocational Education, The Ohio State University, 1984.

Hunter, John Test Validation for 12,000 Jobs: An Application of Job Classification and Validity Generalization Analysis to the General Aptitude Test Battery. Washington, DC: US Employment Service, Department of Labor, 1983.

Hunter, John. "Cognitive Ability, Cognitive Aptitudes, job Knowledge and Job Performance." Journal of Vocational Behavior, Vol. 29, No. 3, December 1986. pp. 340-362.

Hunter, John and Schmidt, Frank. "Fitting People to Jobs", Human Performance and Productivity, edited by Marvin Dunnette and Edwin Fleishman, (1982), pp. 258-271.

Hunter, John E.; Crosson, James J. and Friedman, David H. "The Validity of the Armed Services Vocational Aptitude Battery (ASVAB) For Civilian and Military Job Performance," Department of Defense, Washington, D.C., August, 1985.

Hunter, John E.; Schmidt, Frank L. and Judiesch, Michael K. "Individual Differences in Output as a Function of Job Complexity." Department of Industrial Relations and Human Resources, University of Iowa, June 1988.

Husen, Torsten. "The Influence of Schooling Upon IQ." Theoria, 1951.

Husen, Torsten. Talent Opportunity and Career, (Almquist and Wiksell, Uppsala, 1969).

Jencks, Christopher and Crouse, James. "Aptitude vs. Achievement: Should We Replace the SAT?" The Public Interest, 1982.

Jencks, Christopher et.al. Who Gets Ahead? The Determinants of Economic

Success in America. Basic Books, Inc., Publishers, New York, 1979.

Johnson, George R. "High School Survey," Public School Messenger, 1935, 33: 3-37.

Kang, Suk. "Time Profile of Youths' Labor Market Outcomes: An Analysis of High School and Beyond Data." In High School Preparation for Employment, pp. 95-135. Columbus: The National Center for Research in Vocational Education, The Ohio State University, 1984.

King, L. M.; Hunter, J. W., and Schmidt, F. L. "Halo in a Multidimensional Forced Choice Performance Evaluation Scale." Journal of Applied Psychology, 65 1980, 507-516.

Klein, Roger; Spady, Richard; and Weiss, Andrew. Factors Affecting the Output and Quit Propensities of Production Workers. New York: Bell Laboratories and Columbia University, 1983.

Koretz, Daniel, et. al. Trends in Educational Achievement. Washington: Congressional Budget Office, 1986.

Leestma, Robert, et. al. "Japanese Education Today." A report from the U.S. Study of Education in Japan prepared by a special task force of the OERI Japan Study Team.

Lippold, Stephen; and Claiborn, James M. "Comparison of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale -Revised." Journal of Counseling and Clinical Psychology. 1983, Vol. 51, No. 2, 315.

Lorge, Irving. "Schooling Makes a Difference." Teachers College Record, 1945 Vol. 46 p. 483-492.

Lorge, Irving. "A Table of Percentile Equivalents for Eight Intelligence Tests Frequently Used With Adults." Journal of Applied Psychology, 1936, 19 392-395.

Matarazzo, Joseph. Wechsler's Measurement and Appraisal of Adult Intelligence, The Williams & Wilkins Company, Baltimore, MD 1972.

McClelland, David C. "Testing for Competence rather than Intelligence," American Psychologist, January 1973, pp. 1-14.

Meyer, R. "Job Training in the Schools." In Job Training for Youth, edited by R. Taylor, H. Rosen, and F. Pratzner. Columbus: The National Center for Research in Vocational Education, The Ohio State University, 1982.

Mishra, Shitala; and Brown, Kenneth H. "The Comparability of WAIS and WAIS-R IQs and Subtest Scores." Journal of Clinical Psychology, September, 1983. Vol. 39. No. 5 754-757.

Mueser, Peter and Maloney, Tim. "Cognitive Ability, Human Capital and Employer Screening: Reconciling Labor Market Behavior with Studies of Employee

Productivity." Department of Economics, U. of Missouri-Columbia, June 1987.

National Assessment of Educational Progress. The Reading Report Card. Princeton, New Jersey: Educational Testing Service, 1985.

National Assessment of Educational Progress. Who Reads Best?. Princeton, New Jersey: Educational Testing Service, 1988.

National Assessment of Educational Progress. The Mathematics Report Card. Princeton, New Jersey: Educational Testing Service, 1988.

Olneck, M. "On the Use of Sibling Data to Estimate the Effects of Family Background, Cognitive Skills, and Schooling: Results from the Kalamazoo Brothers Study." In Kinometrics: The Determinants of Socio-economic Success Within and Between Families, edited by P. Taubman. Amsterdam: North Holland, 1977.

Olson, Craig A and Becker, Brian E. "A Proposed Technique for the Treatment of Restriction of Range in Selection Validation." Psychological Bulletin. Vol. 93, No 1, 1983, pp. 137-148.

Reubens, Beatrice. From Learning to Earning: A Transnational Comparison of Transition Services. R&D Monograph 63, 1969, U.S. Department of Labor.

Roessel, F. P. "Comparative Mental Ability of High School Pupils in Three Minnesota Towns in 1920 and 1934." Minnesota Studies in Articulation. University of Minnesota Press, 1937, 122-28.

Rothschild, Michael "Social Effects of Employment Testing." Madison, Wisconsin: Department of Economics, University of Wisconsin, 1979.

Rundquist, E. A. "Intelligence Test Scores and School Marks of High School Seniors in 1929 and 1934." School and Society. Vol. 43, 1936, 301-04.

Schaie, K. Warner and Strother, Charles. "A Cross-Sequential Study of Age Changes in Cognitive Behavior." Psychological Bulletin, 1968, Vol. 70, No. 6, 671-680.

Schaie, K. Warner and Hertzog, Christopher. "Fourteen-year Cohort Sequential Analysis of Adult Intellectual Development." Developmental Psychology. 1983, Vol 19, No. 4, 531-543.

Schmidt, Frank. "The Problem of Group Differences in Ability Test Scores for Employment Selection." Journal of Vocational Behavior, Vol. 29, No. 3, December 1986. pp. 340-362.

Sizer, Theodore R. Horace's Compromise: The Dilemma of the American High School. The First Report from A Study of High Schools, co-sponsored by the National Association of Secondary School Principals and the Commission on Educational Issues of the National Association of Independent Schools, Boston: Houghton Mifflin Co., 1984.

Smith, R. Spencer. "A Comparison Study of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale-Revised in a College Population." Journal of Consulting and Clinical Psychology. Vol 51. 414-419.

Stevenson, Stevenson; Lee, Shin-Ying and Stigler, James W. "Mathematics Achievement of Chinese, Japanese & American Children." Science, February 1986, pp. 693-699.

Taubman, P. and Wales, T. "Education as an Investment and a Screening Device." In Education, Income, and Human Behavior, edited by F. T. Juster. New York: McGraw-Hill, 1975.

Thorndike, R. L. Personnel Selection: Test and Measurement Techniques. Ney York: Wiley, 1949.

Thorndike, R. L. "Mr. Binet's Test 70 Years Later," Educational Researcher, Vol. 4, 1975, 3-7.

Tuddenham, R. D. "Soldier Intelligence in World Wars I and II. American Psychologist, 3, 1948, 54-56.

Urbina, Susana P.; Charles Golden, and Rona Notestine Ariel. "WAIS/WAIS-R: Initial Comparisons." Clinical Neuropsychology. Vol 4. 145-146.

Wechsler, David. WAIS Manual. The Psychological Corporation, New York, New York 1955.

Wechsler, D. WAIS-R Manual. The Psychological Corporation, New York, New York. 1981.

Weiss, Andrew. "High School Graduates: Performance and Wages" Bell Communications Research, Inc. Economics Discussion Paper # 3, 1985.

Wheeler, Lester, R. "A Comparative Study of the Intelligence of East Tennessee Mountain Children." The Journal of Educational Psychology, Vol. 33, 1942, 321-334.

Wigdor, Alexandra K. and Hartigan, John A. editors. Interim Report-Within Group Scoring of the General Aptitude Test Battery. Washington, DC: National Academy Press, 1988.

Table 1

AGE & COHORT EFFECTS ON IQ

Age Group	Full Scale IQ WAIS 1953-54	Full Scale IQ WAIS-R 1976-80	Full Scale IQ Growth Rate per yr	Verbal IQ WAIS 1953-54	Verbal IQ WAIS-R 1976-80	Verbal IQ Growth Rate per yr
16-17	.007	.229	.009	-.253	-.028	.009
18-19	.142	.254	.004	-.075	.052	.005
20-24	.277	.759	.019	.065	.507	.018
25-29	.414	.829	.022	.251	.697	.022
30-34	.156			.055		
35-39	.223	.439	.013	.169	.429	.013
40-44	.000			.051		
45-49	-.067	.325	.020	.054	.473	.020
50-54	-.288			-.131		
55-59	-.385	.032	.017	-.169	.302	.019
60-64	-.408			-.182		
65-69	--	-.260	--	--	.163	--
70-74	--	-.498	--	--	.024	--
Yrly Diff by Age 25-35 to 55-64	.027	.027		.013	.014	
Sample Size	1880	1100		1880	1100	

The table reports the mean score for specific age groups relative to grand mean in the 1953-54 WAIS standardization divided by an average of age specific standard deviations. It was derived from table 10 in Wechsler (1955) and table 7 of Wechsler (1981). The WAIS and WAIS-R were normed on representative samples of the U.S. population. The five equivalence studies used to determine the difference between the mean IQ of the two standardization samples (6.37 pts for Full Scale IQ and 6.48 pts for Verbal IQ) are referenced in the text.

Table 2

Increases in IQ Test Scores Over Time

Country	IQ Point Gain	Period	Test	Age Group	Status
United States	11.0	1918-1943	Army--Wells Alpha	18-33	(4)
	6.0	1932-1953	SB--WAIS	16-48	3
	9.9	1932-1971	SB-LM--SB-72	2-18	2
	6.4	1954-1978	WAIS--WAIS:R	16-70	(3)
	5.3	1942-1987	ITED-Iowa Seniors	17	(3)
United Kingdom	7.4	1939-1979	Ravens	8-30	3
France	25.1	1949-1974	Ravens	18	3
	9.4	1949-1974	Verbal & Math	18	3
Japan	20.0	1951-1975	Wechsler	6-15	3/4
Netherlands	20.0	1952-1982	Ravens	18	1
Norway	8.8	1954-1968	Ravens	19	1
	8.2	1954-1968	Verbal & Math	19	1
Edmonton, Canada	11.0	1956-1977	CTMM	9	1
Belgium	6.8	1958-1967	Ravens/Shapes	18	1
	3.7	1958-1967	Verbal/Math	18	1

Note: WAIS--WAIS:R, ITED and Army Alpha results are discussed in the text. For all other comparisons the source is Flynn 1987. SB stands for Stanford Binet,CTMM stands for California Test of Mental Maturity, ITED stands for Iowa Test of Educational Development, and Ravens stands for the Ravens Progressive Matrices test of Abstract Reasoning. All tests have been adjusted to give them a standard deviation of 15. Flynn's classification of the reliability of the estimate is given in the column headed by status. It has the following key 1 = verified, 2 = probable, 3 = tentative, and 4 = speculative. The status classifications in parenthesis were assigned by the author.

Table 3

Racial Gap in Reading and Math Proficiency

[In Grade Equivalent Units]

	Test Date							
	1971	1973	1975	1978	1980	1982	1984	1986
Reading								
At Age 17	5.3	--	5.0	--	4.8	--	3.3	2.6
At Age 13	4.2	--	3.9	--	3.3	--	2.8	2.3
Math								
At Age 17	--	4.0	--	3.8	--	3.2	--	2.9
At Age 13	--	4.6	--	4.2	--	3.4	--	2.4

Source: National Assessment of Educational Progress, The Reading Report Card. 1985, Data Appendix and Who Reads Best?, February 1988, Table 1.1. The difference between the scores of 17 year olds and 9 year olds was 75 points on the NAEP scale used in the report covering 1971 through 1984 and 18 on the scale used in the report on the 1986 assessment. Consequently, a grade equivalent unit was defined as 9.375 points on the NAEP scale used in the 1971-84 report and 2.25 points on the scale used in the report on the 1986 assessment. The Mathematics Report Card. June 1988, Figure 1.2. The difference between the scores of 17 year olds and 9 year olds was 80.3 points on the NAEP scale. Consequently, a grade equivalent unit was defined as 10 points on the NAEP scale.

Table 4

Effect of Academic Achievement
on the Wage Rates of High School Graduates

Study and Data Set	Date of Graduation	Age	Achievement Measures	Percent Change in Wage Rate Male	Female
Wage Rates					
Kang & Bishop (1985)	1980	19	Test-Math,Voc,Read	-1.9	-.5
High School & Beyond			GPA in Grade 12	.6	2.2
Gardner (1983)	1976-1982	19-24	AFQT	4.8	4.8
NLS Youth					
Daymont & Rumberger	1976-1979	19-21	GPA in Grade 9	.3	2.7
NLS Youth (1982)					
Meyer (1982)	1972	19	Class Rank Grade 12	0.0	2.5
(Weekly earnings)			Test Composite	1.2	2.2
Class of 1972					
Earnings					
Hause (1975)	1961	19	IQ,Test-Math	-3.7	--
Project Talent (white)		23	IQ,Test-Math	6.1	--

The table reports the percentage response of the wage rate or earnings to a one standard deviation improvement in a measure of academic achievement. For high school seniors a one standard deviation differential on an achievement test is about equal to 3.5 grade level equivalents or 110 points on the Verbal SAT. For GPA, one standard deviation is about .7 when C's = 2.0, B's = 3.0 and A's = 4.0.

TABLE 5
DETERMINANTS OF RELATIVE JOB PERFORMANCE

	Yrs of Schooling	Academic Achievement	Perceptual Speed	Psychomotor Skills	Age	Age Square	Occ Exp	Occ Exp Square	Tenure	Tenure Square	R^2	N
Plant Operators	-.013	.244***	.112*	.117**	.048*	-.00053	.024	-.00039	.096*	-.002	.181	651
	(.43)	(3.89)	(1.68)	(2.30)	(1.69)	(1.45)	(.51)	(.28)	(1.93)	(1.36)		
Technician	.028*	.277***	.024	.117***	-.005	-.00008	.041***	-.00097**	.084**	-.0023***	.115	2384
	(1.75)	(8.25)	(.72)	(4.35)	(.33)	(.36)	(2.93)	(2.11)	(5.47)	(3.66)		
Craft Workers	-.017**	.249***	.060**	.079***	.046***	-.00065***	.046***	-.00034***	.064***	-.0016***	.141	10061
	(2.48)	(15.00)	(3.36)	(5.96)	(5.86)	(6.51)	(8.43)	(2.27)	(11.37)	(8.60)		
High Skill Clerical	.013	.272***	.085***	.094***	.035**	-.00051**	.020	-.00017	.117***	-.00316***	.145	2570
	(.82)	(8.75)	(3.17)	(3.63)	(2.31)	(2.55)	(1.35)	(.36)	(7.35)	(5.07)		
Low Skill Clerical	-.015	.296***	.107***	.092***	.035***	-.00057**	.042***	-.00090**	.095***	-.0027***	.135	4124
	(1.28)	(11.91)	(4.43)	(4.48)	(3.46)	(4.29)	(3.36)	(2.15)	(6.73)	(4.94)		
Service	-.024	.298***	.072**	.138***	.045***	-.00056***	.084***	-.0022***	.052***	-.0012	.152	1928
	(1.45)	(8.14)	(1.96)	(4.65)	(3.43)	(3.28)	(5.16)	(4.16)	(2.70)	(1.61)		
Operatives & Laborers	-.049**	.189***	.079***	.140***	.047***	-.00064***	.038***	-.00052	.078***	-.00166***	.137	8167
	(6.59)	(10.65)	(4.37)	(9.53)	(6.62)	(6.79)	(3.77)	(1.58)	(7.38)	(4.65)		
Sales Clerks	-.024	.119	.118	.167**	.071***	-.00084**	-.009	.0012	.026	-.0008	.087	417
	(.70)	(1.34)	(1.41)	(2.38)	(2.63)	(2.45)	(.26)	(1.08)	(.62)	(.50)		

Table 6

LOSS IN PRODUCTIVITY IF RANDOM ASSIGNMENT WERE SUBSTITUTED FOR THE CURRENT ALLOCATION OF WORKERS [LOWER BOUND ESTIMATE]

	Average Compensation per FTE	Standard Deviation of Output	Loss Per Worker	Number of Workers (1000's)	Aggregate Loss (billions)
Plant Operators	$33,808	$91,020	-$9,652	228	-$ 2.3
Technicians	$26,649	$13,668	-$8,672	5261	-$45.6
Craft Workers	$29,655	$12,399	-$3,700	13073	-$48.4
High Skill Clerical	$23,065	$ 8,925	-$4,914	5227	-$25.7
Routine Clerical	$19,472	$ 4,934	-$1,512	12082	-$18.3
Service Exc. Police & Fire	$15,496	$ 4,068	+$ 889	12724	$11.3
Operatives & Laborers	$23,828	$ 5,062	+$ 250	16816	$ 4.2
Sales Clerks	$17,542	$ 5,228	-$ 723	5682	-$ 4.0
All Workers	$22,566	$ 6,708	-$1,815	71,132	-$128.7

Estimates compare the predicted productivity of current members of each occupation with the mean predicted productivity in that occupation of everyone in the USES data set. Predicted job performance was calculated using Model 1, the best fitting model of job performance which included individual variables for gender, race and Hispanic. Dollar impacts were then calculated by first adjusting for the unreliability of the criterion in the standard manner (i.e. dividing by .6), then correcting for restriction of range by multiplying by 1.76 and then multiplying by the standard deviation of output in dollars (column 2 of Table 7).]

Table 7

THE EFFECT OF RE-SORTING
ON AGGREGATE OUTPUT
[UPPER BOUND ESTIMATE]

	Coefficient of Variation	Impact of Resorting on Average Output		Aggregate Gain (billions $)
		Percent	Dollars	
Plant Operators	---	---	$159,282	36.3
Technicians	33.8	17.8	$ 12,667	66.7
Craft Workers	27.6	7.1	$ 5,623	73.6
High Skill Clerical	25.5	.9	$ 579	3.0
Routine Clerical	16.7	.6	$ 190	2.3
Service Exc. Police & Fire	17.3	1.3	$ 537	6.9
Operatives & Laborers	14.0	-3.4	-$ 2,152	-36.3
Sales Clerks	29.8	-23.8	-$ 7,322	-41.5
All Workers			$ 1,558	111.0

Estimates compare the predicted productivity of current members of each occupation with the predicted productivity of those assigned on the basis of model 2 (which ignores race and ethnicity). Model 2 performance prediction were made for each occupation and each worker. Because the standard deviation of output measured in dollars of plant operators was so high, this occupation got first pick. Then came technicians, craft occupations etc. Those not selected for one of the top 7 occupations became sales clerks. Once workers were assigned to occupations on the basis of Model 2, predicted job performance was then calculated using Model 1, the best fitting model of job performance which included individual variables for gender, race and Hispanic. Dollar impacts were then calculated by first adjusting for the unreliability of the criterion in the standard manner (i.e., dividing by .6), multiplying by 1.76 to correct for range restriction and then multiplying by the standard deviation of output in dollars (column 2 of Table 7).]

Table 8

THE EFFECT OF THE RE-SORTING
ON THE ABILITY, GENDER AND ETHNICITY
OF OCCUPATIONS

	General Ability (Pop SD's)		Education		Percent Female		Percent Black		Percent Hispanic	
	Current Level	Change	Current Level	Change	Current Level	Change	Current Level	Change	Current Level	Change
Plant Operator	.09	+2.03	12.1	2.09	2	+75	11.1	-11	5.1	-5
Technician	.28	1.03	13.7	.22	55	+1	8.1	- 6	3.4	-1
Craft	-.09	.65	11.9	.53	4	+43	7.1	- 4	7.4	-3
High Skill Clerical	.32	.02	12.9	.67	83	-16	10.1	- 4	5.4	-1
Low Skill Clerical	.00	.03	12.6	-.48	82	-18	10.6	- 4	5.7	0
Service Exc. Police & Fire	-.52	.12	11.8	.01	82	-14	18.0	- 6	8.3	0
Operative	-.59	-.46	11.3	-.01	66	0	14.7	+ 7	10.0	+2
Sales Clerk	-.02	-1.59	12.3	-1.00	86	-20	8.2	+24	5.6	+3
All Occupations	-.21		12.1		62		11.9		7.3	

This table reports the gender, ethnicity, schooling and test scores of current members of each occupation and the changes in each of these variables that would result if new hires had been selected on the basis of the Model 2 predicted productivity regressions (which ignore race and ethnicity). The simulation was conducted by first calculating the Model 2 performance predictions for each worker in each occupation. Because the standard deviation of output measured in dollars of plant operators was so high, this occupation got first pick. Then came technicians, craft occupations etc. Those not selected for one of the top 7 occupations became sales clerks.