DOCUMENT RESUME

ED 296 134 CE 050 420

AUTHOR Levin, Henry M.
TITLE Ability Testing for Job Selection: Are the Economic Claims Justified?
INSTITUTION Stanford Univ., Calif. Center for Educational Research at Stanford.
SPONS AGENCY Spencer Foundation, Chicago, Ill.
REPORT NO 88-CERAS-02
PUB DATE Mar 88
NOTE 41p.; A version of this paper was presented at the Planning Conference of the Commission on Testing and Public Policy (Berkeley, CA, December 11-13, 1986).
AVAILABLE FROM Publications, Center for Educational Research at Stanford, CERAS Bldg., Stanford University, Stanford, CA 94305 ($3.00).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability Identification; *Aptitude Tests; Educational Research; Intelligence Tests; Job Performance; *Personnel Selection; *Predictive Validity; Productivity; *Research Methodology; Research Problems; *Test Validity

ABSTRACT

The use of ability testing for job selection has become widespread in the Federal Government and in the U.S. Employment Service, which assists private sector employers. The justification for the practice is based largely on research findings claiming a high level of validity for such tests in predicting job performance. More recently, such claims have been translated into the dollar increases in productivity that would result if optimal testing strategies were used for selecting employees for jobs. However, a careful review of the claims indicates that they are not supported by research evidence. The utility of any selection procedure depends on (1) its ability to predict worker performance better than alternatives; (2) the selection ratio of employer openings to applicants; and (3) the economic value of the better employee selection relative to the costs of the selection. On the first point, the evidence that general ability tests are superior to other selection criteria in predicting the various indicators of worker performance is not convincing. Furthermore, much of the research on ability testing for job selection ignores the second point, and much contains many unsubstantiated conclusions and overstatements with regard to the third point. (MN)

88-CERAS-02

ABILITY TESTING FOR JOB SELECTION: ARE THE ECONOMIC CLAIMS JUSTIFIED?

Henry M. Levin

March 1988

An earlier version of this paper was presented at the Planning Conference of the Commission on Testing and Public Policy, Graduate School of Education, University of California, Berkeley (December 11-13, 1986). This paper will be published in a volume on testing, edited by Bernard R. Gifford. Support for the research underlying this paper was provided by the Spencer Foundation under the project grant, "Educational Requirements for New Technologies and Work Organization." The author wishes to acknowledge the helpful suggestions of Edwin Bridges, Catherine O'Conner, Edward Haertel, Marshal Smith, and Richard Snow.
The author is Professor of Education and Economics, Stanford University. He may be reached at: CERAS Building. Stanford University, Stanford, CA. 94305. Telepnone (415) 723 0840.

CENTER FOR EDUCATIONAL RESEARCH AT STANFORD
CERAS

The Center for Educational Research at Stanford serves as an overall support organization for the development, implementation, and dissemination of educational research for the School of Education at Stanford University. CERAS is the successor to both the Institute for Research on Educational Finance and Governance (IFG) and the Stanford Education Policy Institute (IFG). CERAS projects are supported by government agencies, private foundations, and the School of Education itself. A Faculty Research Directory that describes research activities is available, free-of-charge.

Publications of CERAS represent a sample of the large number and wide range of publications and subjects associated with faculty and other researchers in the School of Education. Although research activities in the Stanford School of Education are highly diverse and cover virtually all levels of education and educational practice and all of the associated disciplines, a special effort is being made to focus CERAS resources on two major areas: research on teaching and teaching policy and research on the education of children-at-risk.

Abstract

ABILITY TESTING FOR JOB SELECTION: ARE THE ECONOMIC CLAIMS JUSTIFIED?

The use of ability testing for job selection has become widespread in the federal government and in the U.S. Employment Service which assists private sector employers. The justification for the practice is based largely on research findings that claim a high level of validity of such tests in predicting job performance. More recently such claims have been translated into the dollar increases in productivity that would result if optimal testing strategies were used for selecting employees for jobs. This paper assesses the claims and concludes that they are not supported by the research evidence. The underlying research studies overstate their findings and use questionable approaches to make estimates of economic gains.

H. M. Levin
October 1987

ABILITY TESTING FOR JOB SELECTION: ARE THE ECONOMIC CLAIMS JUSTIFIED?

I. INTRODUCTION

The use of tests by employers to evaluate and choose from among prospective employees has a long history. As recently as 1973, the evidence on the ability of personnel tests to predict job performance was considered to be modest, at best (Ghiselli 1966; 1973). Thus, it is rather astounding to find that by the early 1980's published research was arguing that the use of general ability tests to select workers could increase U.S. productivity by almost $90 billion (Hunter and Schmidt 1982: 268). A U.S. Employment Service report estimated that if tests were given optimal use, the Federal Government could save about $16 billion a year and employers who hire through the U.S. Employment Service could save almost $80 billion (Hunter 1983 a).

These claims were quickly picked up by the U.S. Employment Service and by private employers as a basis for using general ability testing for employee placement in jobs. For example, the Job Service of the State of Missouri distributes a pamphlet to employers that states "Recent studies by the U.S. Department of Labor show that test-selected workers produce an average of about $5,500 more per year than those selected using typical hiring procedures (Mueser and Maloney 1987: 32)."

In 1987 the public employment service systems of some 37 states were using ability tests based upon the Validity Generalization (VG) approach of Schmidt and Hunter (1977) to evaluate and refer job applicants to employers. Three more states were planning to use the approach. Nationally, some 70 local public employment service offices have made VG procedures an integral part of their operational procedures for assessing job applicants and referring them to employers' job openings. In many cases the employment services were responding to the requests of private sector employers for using this approach.

It is clear that a major reason for the widespread revival and expansion of ability testing for employee selection is the claim that it has been "scientifically" shown to increase significantly the economic value of work output and productivity. Its leading advocates have asserted:

> In the past, the value of selection procedures had usually been estimated using statistics that did not directly convey economic value. These statistics included the validity coefficient, the increase in the percentage of "successful" workers, expectancy tables, and regression of job performance measures on test scores. In general, organizational decision makers were less able to evaluate these statistics than statements made in terms of dollars. (Schmidt, Hunter, Outerbridge, and Trattner 1986).

That is, they have suggested that a major policy breakthrough has been the purported capability of expressing the advantages of ability testing in terms of dollar benefits to employers and the economy. In this way, the value of their selection procedures can be made more persuasive to decision-makers.

The purpose of this article is to examine whether the evidence justifies these economic claims. Placement of a dollar value on

gains in productivity associated with the use of ability tests for personnel selection requires: (1) that general ability test performance of workers is superior to alternative selection procedures in predicting worker output; and (2) the additional work output associated with their use has been properly converted into monetary values. A systematic evaluation of the evidence suggests that neither tenet is supported.

The next section will provide background for the economic claims by presenting a brief description of the use of ability testing for personnel selection and its extension to validity generalization approaches to the topic. Section III will discuss the basis for claims that link ability test scores of prospective workers to their productivity, and Section IV will examine the procedures that have been used to connect alleged increases in worker productivity to economic measures of increased output. The final section of the paper will provide a summary.

II- ABILITY TESTING FOR PERSONNEL SELECTION

The use of tests for personnel selection has a relatively long history (Cronbach and Gleser 1965; Ghiselli 1966). However, the present claims of validity are based upon work that began largely in the latter part of the last decade and was centered at the United States Civil Service Commission. The movement was established to ascertain the validity of general ability tests in predicting work output and the extension of findings to a wide range of jobs in the economy.[1] Subsequent work estimated the

economic value of the gains in productivity associated with more and better use of general ability testing.

There are two major aspects of this movement, validity generalization and the economic valuation of benefits. In general, validity generalization refers to:

> Applying validity evidence obtained in one or more simultaneous estimation, meta-analysis, or synthetic validation arguments (American Educational Research Association, American Psychological Association, National Council on Measurement in Education 1986: 94-95).

In the specific context of employee selection, validity generalization refers to the phenomenon of doing intensive validation on the relation between personnel tests and work performance in a few occupations and generalizing the outcomes to a large number of other occupations. This is accomplished by taking a small set of occupations and analyzing them according to their tasks and duties. Ability tests are given to a group of workers in these occupations to ascertain the relation between the tests and measures of work performance, so-called criterion-related validity.

But, criterion-related validity studies are difficult to carry out for a wide variety of apparently disparate occupations and are very costly. Since all or most occupations share various categories of work duties, it is claimed that the predictive ability of the tests can be extended to other occupations without doing "local" criterion-related validity studies. Rather, the results for the few jobs on which criterion-related studies have been done are generalized to other occupations by "reweighting"

the scores according to the different distributions of duties in the other occupations, so-called validity generalization (VG).

In order to do this a different category of validity is used, construct validity. Construct validity is established through four steps: analysis of occupations to ascertain which duties are performed; analysis of duties to ascertain which abilities are needed for performing those duties; selection of specific sub-tests which measure these abilities; and development of a system of weighting the various sub-tests to match occupational requirements.

Thus a single test, the Professional and Administrative Career Examination (PACE), is used by the U.S. Civil Service Commission to select workers for over 100 occupations. The test attempts to measure: (1) deduction or ability to reason from principles; (2) induction or the ability to examine specific facts to arrive at an understanding of their relations; (3) judgement or the solving of a problem under conditions of imperfect information; (4) memory or the ability to retain a large quantity of information; (5) number or the ability to manipulate numbers; and (6) verbal comprehension or effective command of the English language (McKillip et al. 1977). Although PACE is used to select workers for about 120 different federal jobs, its construct validation is based upon only 27 occupations and its criterion validation is based upon studies of only three occupations.

Statistical support for validity generalization (VG) is found in reviews of research on the General Aptitude Test Battery (GATB)

(Hunter 1983 b) and meta-analyses of validation studies (Hunter, Schmidt, and Jackson 1982). Meta-analysis refers to a family of statistical methods for summarizing the results of many different studies of a specific phenomenon (Glass, McGaw and Smith 1981; Hedges and Olkin 1985). Hunter (1983 b) has claimed that meta-analysis of 515 research studies using the GATB over 45 years has shown the generalized validity of that test battery for selecting employees for 12,000 jobs.

Although the VG approach has had great influence in shaping the personnel selection policies of the federal and state governments, the U.S. Employment Service, and some private employers, it has been far more controversial among researchers. For example, other meta-analyses have not found ability testing to show higher validity than alternative selection devices such as biographical data and peer/supervisor ratings (Schmitt et al. 1984; Reilly and Chao 1982). Muesser and Maloney (1987) demonstrate convincingly that the concurrent validity studies that are used as a basis for validity generalization understate seriously the validity coefficients for education relative to ability tests. Lynn and Dunbar (1986) have raised serious issues regarding predictive biases from validity generalization. Many other questions have also been raised as acknowledged by Schmidt et al. (1985) and commented on by Sackett et al. (1985), with particular concern for the penchant of VG advocates Schmidt and Hunter to exaggerate the magnitude, certitude, and policy consequences of modest findings. In what follows, we will not address the validity generalization

issue directly. However, we will address the three criterion-related validity studies that are used as a basis for validity generalization for the use of PACE by the federal government. Much of the criterion-based evidence for validity generalization and for the economic claims associated with use of testing for worker selection is attributable to these three studies (Schmidt, Hunter, Outerbridge, and Trattner 1986). Accordingly, if the three studies are not supported by the claims, extensions of the results of the studies to other occupations are also suspect.

III. VALIDITY CLAIMS AND PRODUCTIVE WORKERS

The appeal of using general ability test scores for personnel selection is the assumption that such a simple device will lead to selection of more productive workers than alternative selection criteria such that the benefits of additional worker output will exceed the additional cost of testing. Indeed, that is the claim made by advocates of VG. Since the marginal costs of testing job candidates is low, this element is typically discounted. The claim rests primarily upon the assumption that general ability testing will provide workers who are more productive than those selected by alternative devices. In this section, I will examine the way in which worker productivity has been measured for assessing the validity of general ability testing.

Economists have devoted considerable thought and empirical work to conceptualizing and measuring worker productivity. In general, it is agreed that worker productivity is not easy to measure (Kendrick 1984). Much work is done in teams where the output is a

result of an interactive process in which it is difficult or impossible to separate out the contributions of individual workers (Alchian and Demsetz 1972). The result is that studies of labor productivity usually use a production function approach in which the output of firms is explained statistically by irputs of different kinds of workers, capital, and other productive resources (Kendrick and Vaccara 1980). The contribution of each input (including different groups of workers) to output is considered to be a measure of the productivity of that input.[2]

The productivity of a worker will depend upon the capital investment of a firm in plant and equipment and the technology or vintage of that investment; the organization of the firm, and the number and characteristics of its workers. In explaining differences in worker productivity in a given job in a given firm (that is with other things held constant), two factors are pertinent: worker capacity and worker effort.

Worker capacity refers to the capability of the worker to be productive with respect to the job requirements. There is a huge literature that explores the various dimensions of worker skill which are considered to be important for worker performance (Dunnette 1983, Fleishman and Quaintance 1984; McCormick 1979; U.S. Employment Service 1965: App. A & B). These include such cognitive dimensions as verbal, mathematical, and thinking ability, categories that are reflected in general ability tests. They also include physical attributes such as perceptual and psychomotor skills, strength, and coordination, characteristics

that are at least partially measured by GATB. Finally, they include social/affective dimensions such as interpersonal skills and ones related to temperament. The social/affective skills are particularly relevant to the four-fifths of the labor force who are found in service occupations. Yet, of the more than 50 specific cognitive, physical, and social dimensions of abilities reviewed in the industrial psychology literature, only a few are covered by the GATB, and none of these are in the social domain.

Even when workers have the capacity to provide a high level of work performance, their actual performance will depend upon the exertion of energy or work effort in applying these skills to the objectives of the workplace. The effort of a worker is thought to be related to his or her personality as well as the supervision, organization, and incentives that are present in the workplace (Stiglitz 1975, Pencavel 1977). In most workplaces it is not uncommon to find diligent workers with modest skills who appear to be more productive than others with superior skills. These differences may be systematically related to the match or mismatch between job requirements and worker characteristics, where those workers who are most closely matched provide greater effort than those who are not (Tsang and Levin 1985).

The literature on worker productivity suggests that workers need both skills or human capital (Becker 1964) and a conscientious application of those skills to be productive. In addition, to the many dimensions of general ability that may be pertinent to a job, there are likely to be specific cognitive abilities, physical

attributes, and socio-affective characteristics that are necessary for particular types of work. Finally, the mere existence of these capacities does not produce work output. Somehow, these skills must be tranformed into work output through the application of worker effort, a fact that is a matter of particular concern to work organizations (Vroom 1964). Given this brief background on the relation between worker characteristics and worker productivity, we can proceed to review the literature that ties general ability testing to worker performance. I will focus on a recent article by Schmidt, Hunter, Outerbridge, and Trattner (1986) which relies on cumulative findings and summarizes the latest thinking on ability testing for job selection. It applies the technique of VG to the hiring of white collar workers in the government:

> This study examines ... productivity gains for most white-collar jobs in the federal government. In the present study, these job performance differences were determined empirically, based on direct multi-method measurement of the job performance of employees who had been selected years earlier, either (a) using cognitive ability tests or (b) using other methods (mostly evaluations of education and experience.)...Results from three different studies show that the job performance of test-selected employees averages approximately one-half a standard deviation higher than of non-test-selected employees. Results also indicate that use of measures of cognitive skills in place of less valid selection methods for selection of a one-year cohort in the federal government would lead to increases in output worth almost $600 million per year for every year the new hires remain on the job (Schmidt et al. 1986: 25-26).

It is important to review the specific way in which employee job performance is measured in the light of the discussion set out above. The authors premise their findings on three studies that were done for the U.S. Civil Service Commission in 1977. Although

Schmidt et al. (1986) refer to these studies as measuring job performance of employees (p. 25) or even increases in output (p. 1), none of the studies measured actual job performance or productivity. Rather, they validated the selection tests on various indicators which are presumed to be related to productivity.

Each of the studies used Test 500 of the federal government's Professional and Administrative Career Examination (PACE) to predict "job performance." The three occupations that were covered by the studies included: Internal Revenue Service revenue officers (O'Leary and Trattner 1977), customs inspectors (Corts, Muldrow, and Outerbridge 1977), and social insurance claims examiners (Trattner, Corts, van Rijn, and Outerbridge, 1977). Let us examine briefly how "job performance" was measured for each occupation.

(1) IRS Revenue Officers

For the 305 IRS revenue officers in the sample, job performance was measured using a job information test, a work sample, and supervisory ratings. The 59 job information items were constructed in a multiple choice format that addressed the 12 major job duties. The work sample asked the respondents to determine what actions should be taken to collect delinquent taxes in five separate cases. Respondents were given the files and asked to select the appropriate actions. The supervisory ratings were based upon behavioral scales for each of the 12 major job duties.

For the sample of 190 customs inspectors the criteria included a job information test, a work sample, and supervisors' ratings and rankings. The job information test was composed of 50 multiple choice questions that were based upon the major job duties. The work sample was not actually a sample of the work of the respondents, but an evaluation by the respondents of a video-taped work sample of another customs inspector. Respondents were given a booklet in which they were asked to record errors in procedures and ways that performance could be improved. The supervisory ratings were based upon using a ten-point graphic scale to rate 33 dimensions of performance over 12 major duties. Supervisory rankings were also based upon rank ordering of the respondents' proficiencies according to each of the duty statements.

For the sample of 252 social security claims authorizers, the criteria included a job information test, work sample, and supervisory ratings. The job information test comprised 42 multiple choice questions. The work sample consisted of a standardized claim that had to be adjudicated by the respondent. First-line supervisors were asked to rate respondents according to their performance on eight job duties as well as to rank-order the respondents. Table One shows the validity coefficients for the indicators of job performance for each of the three occupations.[3]

Table One--Validity Coefficients of Test 500 Total Scores for Three Occupations With Three Indicators of Job Performance [1]

	Job Information Test	Work Sample	Supervisory Rating
Customs Inspector[2]	.56	.52	not sig.[3]
Internal Revenue Officers[4]	.55	.51	.25
Social Insurance Claims Examiner[5]	.59	.39	.28

[1] All validity coefficients are based on method of obtaining multiple correlation with optimally weighted raw subscores. The patterns are similar when other methods are used. The coefficients are also corrected for bias according to the Burket (1964) procedure.

[2] Corts et al. (1977): 49.

[3] statistically insignificant

[4] O'Leary and Trattner (1977): 22.

[5] Trattner et al. (1977): 29.

Job Information Test

Coefficients for the job information test range from about .56 to .59. However, in interpreting these relative high coefficients, we must keep in mind that: (1) a test of job information is not a direct assessment of job performance, but

only a measure of job knowledge; and (2) that the specific method of measuring job information is a multiple choice test with a format similar to the predictor, Test 500. The first stipulation means that we must not equate a multiple choice test of job information with an actual assessment of job performance, regardless of the casual interchangeability among those terms in Schmidt et al. (1986). A test of job information is not a direct measure of job performance, but only one of many potential indicators or determinants of job performance. It tells us nothing about worker effort or a plethora of interpersonal skills that are important in production and organizational life. The second means that the validity coefficient is likely to be overstated to the extent that it reflects overlapping method variance, that is the degree to which respondents who do well on multiple choice tests will have higher scores on both test 500 and the job information test, exclusive of their true ability and knowledge levels. Persons with good test-taking skills on multiple choice items will tend to do better on both types of tests than equally able persons with poorer test-taking skills.

The advocates of VG do not attempt to correct validity coefficients for methods variance. Rather they argue that the existence of such a problem is negated because the general ability tests correlate equally highly with job sample measures:

> Job sample measures are not written tests and would not be expected to share methods variance with ability tests. The fact that ability tests correlate about equally with job sample measures and with training performance measures indicates that what is important are the ability, knowledge, and skills

measured, not the methods used to measure them (Schmidt, Hunter, Pearlman, and Hirsh 1985: 733).

And, as Table One shows, this assertion appears to be supported by the relatively high and comparable validity coefficients for the work samples. But, a closer inspection of the work samples shows that far from being samples of job performance in an actual workplace, they are--at best--simulations of work tests that depend heavily on test taking skills.

Work Samples

As the name implies, work samples refer to the use of a representative sample of work activity that is used as a basis for analysis (e.g. to validate employee selection criteria). But in the case of customs inspectors, no work sample was administered to the respondents on whom the ability tests were being evaluated. Rather, they were asked to view a video-taped sample of the work of some other customs inspector (selected especially for the video-taping) to identify errors in procedures and indicate ways in which procedures could be improved. A special booklet was provided to write down answers in a test format. This criterion is certainly not an evaluation of a work sample of the customs inspectors who were the basis of the validity study, even though it is referred to as a work sample. Rather, it appears to be relevant only as a different form of a job information test.

In the cases of the internal revenue officers and the social insurance claims examiners, the work samples were "simulated" rather than actual samples of work that were evaluated in real work settings. The internal revenue officers were given five

Taxpayer Delinquent Accounts for which they had to make collection decisions based upon the information contained therein. The goal was to determine the course of action to take to resolve the case in the best interest of the government. The work sample for the social security claims examiners was a single case that had to be adjudicated on the basis of the information submitted in support of the claim. Each was scored on the appropriateness of the actions taken.

But, even in these cases the simulated samples were reduced to paper and pencil testing situations that were abstracted from the real work setting. For example, in setting out the work duties of the internal revenue officer, the duty on which the officer spends the largest amount of time was ignored in both the job information and work sample tests.

> This duty, Locating and Contacting Taxpayers, mainly involves social contact with taxpayers and did not lend itself to measurement in either of these two criteria. Performance on this duty was, however measured on the supervisory rating form (O'Leary and Trattner 1977: 12).

An inquiry to the Internal Revenue Service indicated that the evaluation of worker performance in delinquent tax cases cannot be done in the absence of seeing how the officer uses information and discussions in these contacts to resolve issues. For example, there is the problem of finding the taxpayer. Some officers are better at this than others. Second, there is the issue of negotiating a settlement that maximizes the government interest, taking into account feasibility of the agreement from the taxpayer's perspective and avoiding expensive collection

procedures and legal action on the part of the government. Third, there is the search for leverage on the situation such as getting information on the employer to pose the threat of wage garnishment or locating other financial interests and assets of the taxpayer that could be attached to pay the debt. All of these acts require detective work, intuition, and important social skills in determining how to proceed, and none of them can be carried out without contacting the delinquent taxpayer and other persons with whom he is linked.

In the case of the simulated work sample for the social insurance claims examiner, there are also serious flaws relative to the real work setting. Only a single case was used as a basis for evaluation. No provision was made for investigation or communication with others to obtain more information, even though many claims are incomplete and require more documentation or assistance from specialists. No study was made of actual productivity in terms of the number of cases that were handled by examiners within a given work period.

In summary, one of the work samples was not a sample of work of the subjects who were being evaluated, and the other two work samples were far removed from the real work setting and were constrained to reflect only limited dimensions of the jobs. These evaluations could be better described as assessments of simulated task performances using a pencil and paper format under testing conditions rather than evaluations of actual work performance. They were limited to exercises that did not allow for the wider

range of behavior that is necessary to performing competently in the workplace, and they were carried out within testing time constraints. The result is that they too are likely to share methods variance in the calculation of validities.

The one criterion that is likely to take all of the characteristics of the job into account is that of supervisory ratings. First-line supervisors are able to assess actual productivity of workers or at least to observe the proficiencies of workers in performing work tasks as well as productive work effort. Both the job information tests and the simulated work samples tend to focus on much narrower dimensions of the job as well as to ignore such matters as effort or cooperation and communication with colleagues, behaviors which are important to organizational productivity. Indeed, in the case of the internal revenue examiners, it was claimed that the most time-consuming aspect of the job could only be validated by the supervisory ratings.

Supervisory Ratings

Supervisory ratings for the three occupations are shown in the third column of Table One. The most noticeable pattern is that the validity coefficients for the supervisor ratings are considerably lower than those for the testing situations reflected in the job information tests and work samples. In the case of the customs inspectors, there is no statistically significant relation between supervisory ratings and Test 500 scores. In the other cases, the validity coefficients are in the range reported for

many selection criteria and are hardly impressive. Indeed, they are less than half of those calculated from the job information test.

In terms of specific validity coefficients, those for customs inspectors and for internal revenue officers are worthy of further comment. The researchers explained the insignificant validity coefficient for the customs inspectors by asserting that the job environment and nature of supervisory duties do not permit adequate direct observation of inspectors performing individual duties (Corts, Muldrow, and Outerbridge 1977: 43).

> It is also recognized that the ratings and rankings obtained may be based upon general impressions of the inspectors' work, and therefore could contain a large component of cooperativeness, speediness, "knowing the ropes," acceptability within the group, maturity, and other similar characteristics... the conclusion is that these data contain components of error and other variance with which Test 500 could not be expected to correlate. (Corts, Muldrow, and Outerbridge 1977: 43).

There are two important features of this explanation. The first is that the explanation is ex post. That is, after ascertaining that the validity coefficient was insignificant, there is a concerted effort to show that the supervisory ratings are not appropriate measures of validity for this occupation. It is instructive to note that this concern did not emerge in the very extensive design phase of the study with its close attention to detailed analysis of the occupation and its supervision.

The second aspect is the focus on individual performance as the exclusive focus of productivity differences. Productivity analyses in economics (Spence 1974; Williamson 1975) and industrial organization (Pasmore and Sherwood 1978) have

emphasized an organizational approach in which activities among workers are considered to be interdependent. As such, they require analysis of productivity by work group rather than limiting assessments to individual performance for narrowly specified duties. That is, interpersonal skills, communication with clients and co-workers, and group problem-solving skills are often as important as the work skills for individual work performance. Although supervisors can evaluate this entire range of skills and work performance, the authors of this study view such components of work performance as "error" rather than as central to an understanding of work output and productivity.

Given that supervisors can observe work performance directly in both its individual dimensions and those that affect organizational productivity, supervisory ratings are likely to be more valid than measures derived from paper and pencil tests in non-work settings. This conclusion is reinforced by the fact that the most important duty in terms of time allocation for internal revenue officers is "social contact with taxpayers" which could only be evaluated by supervisors.

On the basis of this information, it is reasonable to hypothesize that the supervisory ratings are more nearly valid indicators of work performance than the test data for job knowledge and work samples. But, these coefficients are well within the boundaries of validity coefficients associated with a wide variety of selection devices (Reilly and Chao 1982; Schmitt, Gooding, Noe, and Kirsch 1984).

Although the validity coefficients for the two other validity criteria have higher coefficients, they are only indirect indicators of work performance. If we assume a relatively high correlation of .5 between these indicators and actual work performance, there would still be an overall validity coefficient of less than .30 between Test 500 and the true score for work performance. Using either this coefficient or those associated with supervisory ratings, less than 10 percent of the variance in work output is explained by Test 500. This result is about the same as for other selection criteria, and it is hardly a basis for arguing that general ability tests are a powerful predictor of worker productivity.

IV. ECONOMIC VALUE OF ESTIMATED GAINS IN WORKER PERFORMANCE

Schmidt et al. (1986) stress the usefulness of converting the value of selection methods into dollar terms. Such terms suggest to decision-makers the concrete and calculable economic gains to be made from using various methods of worker selection. They argue that on the basis of the validity coefficients set out in Table One, they can estimate the economic gains from using general ability tests to select white collar workers in the government. This requires the conversion of the putative gains in worker performance into dollar values. Specifically, they find that on the basis of these validity coefficients, the use of Test 500 for selecting a one year cohort of such workers would increase government output by up to $600 million a year or increase output by almost 10 percent.

Although economists have had considerable experience in estimating the value of worker productivity (Chinloy 1981; Denison 1985; Kendrick 1984), the authors do not refer to any economic literature. Moreover, they reject a cost-accounting approach without reference to that literature. Rather, they rely on disparaging comments made by Cronbach (Cronbach and Gleser 1965) about a doctoral dissertation in psychology done in 1961 by Roche at Southern Illinois University that used a cost-accounting approach (Hunter and Schmidt 1982: 248). Even Cronbach stated that:

> This study relies heavily on the discipline--or art--of accounting, and Roche, a psychologist was necessarily dependent on the advice of the accountants. It is not entirely certain that the accountants perceived the program clearly, and it may well be that in future studies a more thoroughly interdisciplinary attack will produce better solutions to the acccunting problems (Cronbach and Gleser 1965: 266).

Nevertheless, a summary of a psychology dissertation (not even the original document) from almost three decades ago is used as the basis for rejecting a cost-accounting approach.

They then use their own approach to estimating the economic value of the putative increases in worker productivity by asking supervisors to estimate the dollar value to the organization of the products and services produced by the average employee, by one at the 85th percentile, and by one at the 15th percentile (Hunter and Schmidt 1982: 248-251). Since, the 15th and 85th percentiles are one standard deviation above and below the mean respectively, they estimate the standard deviation of worker performance in terms of dollars. Estimated increases in worker performance from

using general ability tests for selection are converted into standard deviation units and translated into dollar values. Such values are set out into ratios of the standard deviation of productivity in dollars relative to the arithmetic mean of productivity in dollars.

Schmidt and Hunter (1983) argue that the standard deviation of worker output is at least 20 percent and as high as 40 percent of the mean output of workers. These assumptions are used to estimate gains in national productivity from worker selection based upon general ability testing (Hunter and Schmidt 1982) as well as to estimate such gains for all white collar workers in government (Schmidt et al. 1986) and for specific occupations such as computer programmers (Schmidt et al. 1979).

But, this procedure is fundamentally flawed for at least two reasons. First, the method for obtaining the economic value of additional productivity is highly dubious. Basically, the procedure entails a survey of supervisors that asks their opinions about the value of output of workers on different parts of the productivity distribution. There are internal contradictions in this procedure that emanate from the very studies on which Hunter and Schmidt build their argument.

When supervisors are asked to rate worker performance, something directly observable by them and within their domain of experience, the validity coefficients tend to be low or even insignificant as in the case of customs inspectors (See Table One). One explanation for these weak results by authors of the VG literature

is that supervisors' ratings of workers are highly errorful, despite the fact that they reflect a central duty to which supervisors are regularly assigned (Corts, Muldrow, and Outerbridge 1977). But, if supervisors do such a poor job of performing a function at which they are presumably skilled and knowledgeable and which is directly observable, how can we assume that they can estimate the economic value of a standard deviation of worker performance--something for which they lack experience, information, and a basis for calculation. If cost accounting approaches as interpreted by a doctoral student in psychology are considered to be problematic, opinion sampling approaches without cost-accounting information are likely to be even more unreliable.

Second, even if the economic values were appropriate, the straightforward application of the validity coefficients in Table One will overstate vastly differences in productivity due to differences in worker selection criteria. This procedure equates a z-score or standard deviation increase in the indicators of work performance as measured by tests of job knowledge, "simulated" work samples, and supervisory ratings with a similar increase in worker productivity. As I argued in the previous section, the validity of general ability tests to predict actual worker productivity is likely to be less than .3, explaining less than 10 percent of the variance.

Schmidt et al. (1986) use the much higher validity coefficients generated by the indicators of worker performance, not the actual worker performance. This means that any estimated improvement in

worker performance on the validity criteria associated with general ability tests will represent a much smaller increase in actual worker productivity or work output. Such unmeasured dimensions as worker effort, interpersonal abilities, and a variety of other determinants of worker performance that are not reflected in the criteria validation studies will explain the rest of the variance in worker performance.

But, the technique of placing dollar values on standard deviations of worker performance attributes all of the difference in worker performance to differences in estimated productivity created by ability selection. This is far from the true case, since the validity criteria are never based upon actual work performance but only potential indicators of that performance. That is, a one standard deviation improvement in the indicators of worker performance as measured in the validity studies is likely to yield a much smaller increase in actual worker productivity than one standard deviation. The result is a substantial overstatement of the dollar value of probable productivity gains attributible to ability testing.

V. SUMMARY

The work of the validity generalization theorists is rich in heuristic value and its attempt to extend the value of employee selection criteria to a large set of occupations and decision criteria. But, this should not detract from the fact that it is a literature of vast overstatement that often appears to be drafted more for its persuasive power than its scientific validity. A

number of rigorous studies have illustrated the magnitude of such biases.

The utility of any selection procedure depends upon (1) its ability to predict worker performance better than alternatives; (2) the selection ratio of employee openings to applicants; and (3) the economic value of the better employee selection relative to the costs of the selection.

On the first of these, the evidence is not convincing that general ability tests are superior to other selection criteria in predicting the various indicators of worker performance. Mueser and Maloney (1987) have shown that validities of general ability tests will be systematically overstated relative to education in concurrent validity studies where the subjects of the study have already been selected on education (Mueser and Maloney 1987). Meta-analyses of selection criteria by other authors find that biographical data and peer/supervisor evaluations show equal or even higher validities than general ability tests (Schmitt et al. 1984; Reilly and Chao 1982).

Second, they assume that selection ratios are low, where the selection ratio is defined as the number of persons who are accepted for employment relative to the number of applicants. This means that it is possible to choose from among a large number of job candidates. The larger the choice, the greater the potential benefits of the most preferred selection criterion. In contrast, if the applicant pool does not exceed the number of persons hired, no selection is possible and no selection berefits are

forthcoming. Their research suggests that professional and technical jobs represent the occupations for which general ability testing is likely to yield the largest selection benefits (Hunter 1983 b). But these are the occupations in which there are rarely a surplus of candidates relative to positions. In fact, there are often shortages of candidates. By assuming low selection ratios, they overstate substantially the benefits of any improvement in selection.

For example, in their study of computer programmers, Schmidt et al. assume a selection ratio of .2 (only 20 percent of the applicants will be hired). Using this assumption, they calculate the benefits of using a programmer aptitude test over previous selection procedures. They conclude that the test would produce a benefit to the employer of almost $65,000 for each programmer hired or a total gain in productivity for the American economy of $11 billion. Such a claim is not grounded in reality. As Cronbach (1984) summarizes:

> This projection is a fairy tale. The economy utilizes most of the persons who are trained as programmers, and only the most prestigious firm can reject 80 percent of those who apply. If 90 percent of the programmers are hired somewhere, the tests merely give a competitive advantage to those firms that test (when other firms do not test).

Third, we have pointed to other sources of overstatement of the economic consequences of general ability testing. For example, the literature makes claims about how the use of general ability tests can increase worker productivity, worker output, and the output of industries and the economy, but the actual measures of worker performance are highly incomplete and artificial measures of

workplace behavior. Further, the setting of economic values on putative gains in worker performance are vastly overstated by the rather simplistic estimation technique that is used. Finally, in some attempts to extend their findings from a few occupations to entire industries, there is a tendency to ignore the compositional fallacy in which gains in worker abilities for some employers will mean losses to other employers (Schmidt et al. 1986). Even when this is recognized (Hunter and Schmidt 1982), it is not clear how a highly decentralized economy in which employment decisions are made at micro-levels would result in an optimal redistribution of talent along the lines that are recommended (Rothschild 1979).

The effects of all of these biases and overstatements is likely to be substantial. Rothschild (1979) has tried to analyze some of them in a formal model of the worker selection and production process and suggests that they are multiplicative rather than additive. He concludes that:

> Hunter and Schmidt's estimates should be scaled down by a factor of 8. Thus, the range of possible improvements in productivity due to a more systematic use of ability tests is .4% to 1.75% instead of 3.2% to 14%. Similar gains in productivity would be observed if everyone worked from 9.6 to 42 minutes longer in a forty-hour week (Rothschild 1979: 25).

Even this assessment does not take full account of the full range of sources of overstatement. In short, the economic claims are vastly exaggerated, and the research and findings are not adequate to support such claims.

Future Research

Although there are many problems with the approach taken by the VG advocates, a substantial number of them seem to be attributible to an inadequate understanding of labor markets and the measurement of worker productivity. These are domains in which economists have worked for over a century. It would seem that a major endeavor to improve the estimates of the effects of different worker selection criteria on productivity must be multi-disciplinary in which economists and industrial psychologists work together. Such a collaboration should also take account of the incentives to employers and potential employees of selection criteria (Maloney and Mueser 1987) as well as the relative costs and benefits of different selection approaches. The fact that so much of the validity generalization literature on the economic gains from general ability testing has made virtually no reference to the pertinent economic literature is a very telling sign.

This concern is especially sharpened by the potential in measuring worker productivity directly for the three occupations that were discussed in this paper and that have represented the base of so much of the validity generalization work. The productivity of internal revenue officers could be measured by randomly assigning delinquent taxpayer cases to a sample of such officers over a two or three year period. Productivity would be measured by the yields in additional payments that they were able to derive in behalf of the U.S. government, taking account of any

differential costs imposed on the government (e.g. through collection procedures or litigation).

Customs inspectors could be evaluated by a initiating a secondary search of randomly selected persons who had been initially screened by the customs inspectors in the study. Estimates could be made of their productivity by valuing their services in terms of the average number of persons whom they serve in a given period and the accuracy of their assessments. Such assessments could converted into monetary terms by evaluating the recovery of customs duties and the avoidance of social costs associated with illegal contraband such as drugs or banned agricultural products as well as savings. Benefits would also include the resource savings when additional persons are served by an inspector in a given period. A similar approach could be used to evaluate the performance of social insurance examiners by randomly assigning cases and assessing the number of cases that are processed as well as the costs to the agency and taxpayer of errors (e.g. the cost of appeals and re-evaluations of cases).

These measures of output would take account of the ability of workers to use their interpersonal skills and to obtain information from others in a collaborative setting.
In addition, they would permit a better benefit-cost analysis of alternatives than the VG method allows by taking account of both the costs of selection and workplace costs associated with productivity for each worker. For example, internal revenue officers who are able to obtain collections from delinquent

taxpayers with minimal dependence on the courts, collection agencies, and other personnel in the bureaucracy impose a lower cost on their employer than ones who obtain settlements that rely on heavy use of these other resources. Differences in institutional costs associated with performance will not be picked up in job information tests or the synthetic work samples that depend on individual behavior under test conditions and that do not consider differences in organizational consequences among workers.

In the long term it is best to view the choice of employee selection methods in the context of benefit-cost decisions (Levin 1983, 1987; Mishan 1976). An attempt should be made to consider all of the benefits and costs of the alternatives. Benefits and costs for the employer should be calculated for the organization as a whole rather than for individual workers in the absence of organizational consequences. And, estimates of impacts for the economy as a whole must be far more sophisticated than ones that assume that a result obtained for a few workers or firms can be generalized to the entire economy without taking account of compositional fallacies and interdependence among decentralized decisions.

Footnotes

[1] A good elementary review of validation approaches for personnel selection in the present context is General Accounting Office (1979), Chap. 3. Also see Cascio (1982), Chap. 7 and Cronbach and Gleser (1965).

[2] See Tsang (1987) for an example of an empirical study.

[3] The validity coefficient is generally defined as the correlation of test score with the outcome or criterion score. For classic discussions of validity coefficients and employee selection, see Brogden (1949) and Cronbach and Gleser (1965). For the derivation of multiple correlations with optimally weighted raw subscores, see the details in the three studies cited in Table One.

Bibliography

Alchian, A., and H. Demsetz. (1972, December). Production, information costs, and economic organization. American Economic Review, 62(5), pp. 777-795.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1986). Standard for educational and psychological testing. Washington, DC: American Psychological Association.

Becker, Gary S. (1964). Human capital. New York: Columbia University Press.

Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective efficiency. Psychometrika, Vol. 14, pp. 169-182.

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. Psychometric Monographs, Vol. 12, pp. 1-66.

Cascio, Wayne F. (1982). Costing human resources: The financial impact of behavior in organizations. Boston: Kent Publishing Co.

Chinloy, Peter. (1981). Labor Productivity. Cambridge, MA: Abt Books.

Corts, Daniel B., Tressie W. Muldrow, and Alice M. Outerbridge. (1977, December). Research base for the written test portion of the professional and administrative career examination (PACE): Prediction of job performance for customs inspectors, PS-77-4 (Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Cronbach, Lee J. (1984). Essentials of psychological testing, Fourth Edition. New York: Harper & Row, Publishers.

Cronbach, Lee J., and Goldine C. Gleser. (1965). Psychological tests and personnel decisions. Chicago: Illini Books.

Deneson, Edward F. (1985). Trends in American economic growth, 1929-1982. Washington, DC: The Brookings Institution.

Dunnette, Marvin D. (1983). Aptitudes, abilities, and skills. In Marvin D. Dunnette (Ed.), Handbook of industrial and organizational psychology. New York: John Wiley and Sons, pp. 473-520.

Fleishman, Edwin, and Marilyn Quaintance. (1984). Taxonomies of human performance. New York: Academic Press.

Glass, Gene V., Barry McGaw, and Mary Lee Smith. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage Publications.

Hedges, L. V., and I. Olkin. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Hunter, John E. (1983a). The economic benefits of personnel selection using ability tests: A state of the art review including a detailed analysis of the dollar benefit of U.S. employment service placements and a critique of the low-cutoff method of test use. UESE Test Research Report No. 47. Washington, DC: Employment and Training Administration, U.S. Department of Labor, 1983.

Hunter, John E. (1983b). Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the general aptitude test battery. USES Test Research Report No. 45. Washington, DC: Employment and Training Administration, U.S. Department of Labor, 1983.

Hunter, John E., and Frank L. Schmidt. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E. A. Fleishman and M. D. Dunnett (Eds.), Human performance and productivity: Vol. I: Human capability assessment. Hillsdale, NJ: Erlbaum, pp. 233-284.

Hunter, John E., and Frank L. Schmidt. (1983, April). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. American Psychologist, pp. 473-478.

Hunter, John E., Frank L. Schmidt, and Gregg Jackson. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage Publications.

Kendrick, John W. (1984). Improving company productivity. Baltimore: Johns Hopkins Press.

Kendrick, John W., and Beatrice N. Vaccara (Eds.). (1980). New developments in productivity measurement and analysis. Studies in Income and Wealth, Vol. 44. Chicago: University of Chicago Press.

Levin, Henry M. (1987), Summer). Cost-benefit and cost-effectiveness analyses. In David S. Cordray, Howard S. Bloom, and Richard J. Light (Eds.), Evaluation practice in review--New directions for program evaluation, No. 34. San Francisco: Jossey-Bass, pp. 83-99.

Levin, Henry M. (1983). Cost-effectiveness: A primer. Beverly Hills, CA: Sage Publications.

Linn, Robert L., & Stephen B. Dunbar. (1986). Validity generalization and predictive bias. In Ronald A. Berk (Ed.), Performance Assessment: Methods and application. Baltimore: Johns Hopkins Press.

McCormick, Ernest. (1979). Job analysis: Methods and applications. New York: AMA-COM.

McKillup, Richard H., Marvin H. Trattner, Daniel B. Corts, and Hilda Wing. (1977, April). The professional and administrative career examination: Research and development, PRR-77-1. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Mishan, Ezra J. (1976). Cost-benefit analysis. New York: Praeger Publishers.

Mueser, Peter, and Tim Maloney. (1987, June). Cognitive ability, human capital and employer screening: Reconciling labor market behavior with studies of employee productivity. Unpublished paper available from Peter Mueser, Department of Economics, University of Missouri, Columbia, MO.

O'Leary, Brian, and Marvin H. Trattner. (1977, August). Research base for the written test portion of the professional and administrative career examination (PACE): Prediction of job performance for internal revenue officers, TS 77-6. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

O'Leary, Brian S. (1977, August). Research base for the written test portion of the professional and administrative career examination (PACE): Prediction of training success for social insurance claims examiner, TS 77-5. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Pasmore, William A., and John J. Sherwood. (1978). Sociotechnical systems: A sourcebook. San Diego: University Associates.

Pencavel, John H. (1977b). Work effort, on-the-job screening, and alternative methods of remuneration. In R. G. Ehrenberg (Ed.), Research in Labor Economics, Vol. I. Greenwich, CT: JAI Press, pp. 225-259.

Reilly, Richard R., and Georgia T. Chao. (1982). Validity and fairness of some alternative employment selection procedures. Personnel Psychology, Vol. 35, pp. 1-62.

Rothschild, Michael. (1979, June). Social effects of ability testing. Unpublished paper available from author at Department of Economics, University of California, La Jolla.

Sackett, Paul R., Neal Schmitt, Mary L. Tenopyr, Jerard Kehoe, and Sheldon Zedeck. (1985). Commentary on forty questions about validity generalization and meta-analysis. Personnel Psychology, Vol. 38, pp. 697-798.

Schmidt, Frank L., and John E. Hunter. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, Vol. 62, pp. 529-540.

Schmidt, Frank L., John E. Hunter, Kenneth Pearlman, and Hannah Rothstein Hush. (1985). Forty questions about validity generalization and meta-analysis. Personnel Psychology, Vol. 38, pp. 697-798.

Schmidt, Frank L., John E. Hunter, Robert C. McKenzie, and Tressie W. Muldrow. (1979) Impact of valid selection procedures on work-force productivity. Journal of Applied Psychology, 64(6), pp. 609-626.

Schmidt, Frank L., John E. Hunter, Alice N. Outerbridge, and Marvin H. Trattner. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. Personnel Psychology, Vol. 39, pp. 1-29.

Schmitt, Neal, Richard Z. Gooding, Raymond A. Noe, and Michael Kirsch. (1984). Meta-analysis of validity studies, published between 1964 and 1982 and the Investigation of study characteristics, Vol. 37, pp. 407-422.

Spence, A. Michael. (1974). Market signaling: Informational transfer in hiring and related screening processes. Cambridge, MA: Harvard University Press.

Stiglitz, J. (1975, Autumn). Incentives, risk and information: Notes toward a theory of hierarchy. Bell Journal of Economics, 6(2), pp. 552-579.

Trattner, Marvin H., Daniel B. Corts, Paul P. van Ryn, and Alice M. Outerbridge. (1977, September). Research base for the written test portion of the professional and administrative career examination (PACE): Prediction of job performance for claims authorizers in the social insurance claims examining operation, TS 77-3. Washington, DC: Personnel Research and Development Center, U.S. Civil Service Commission.

Tsang, Mun Chiu. (1987). The impact of underutilization of education on productivity: A case study of the U.S. Bell Companies. Economics of Education Review, 6(3), pp. 239-252.

Tsang, Mun C., and Henry M. Levin. (1985). The economics of overeducation. Economics of Education Review, 4(2), pp. 93-104.

U.S. Employment Service. (1965). Dictionary of Occupational Titles, 2 vols., 3rd ed. Washington, DC: U.S. Government Printing Office.

U.S. General Accounting Office. (1979, May 15). Federal employment examinations: Do they achieve equal opportunity and merit principle goals? Washington, DC: Comptroller General of the United States.

Vroom, V. H. (1964). Work and motivation. New York: J. Wiley and Sons, Inc.

Williamson, Oliver E. (1975). Markets and hierarchies. New York: The Free Press.