DOCUMENT RESUME

ED 235 221 TM 830 641

TITLE	Overview of Validity Generalization for the U. S. Employment Service.
INSTITUTION	North Carolina Employment Security Commission, Raleigh.
SPONS AGENCY	Employment and Training Administration (DOL), Washington, D.C.
REPORT NO	USES-TRR-43
PUB DATE	83
NOTE	31p.; Report prepared by Psychological Services Incorporated under contract to the Southern Test Development Field Center.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Aptitude Tests; Job Performance; Personnel Evaluation; Personnel Management; *Personnel Selection; Pilot Projects; *Predictive Measurement; *Test Use; *Test Validity; Vocational Aptitude
IDENTIFIERS	*General Aptitude Test Battery; *Validity Generalization; Validity Research

ABSTRACT

The United States Employment Service is now able to expand the General Aptitude Test Battery (GATB) coverage from approximately 400 jobs to all candidates for every job in the Dictionary of Occupational Titles (over 12,000 occupations). In addition, employers can now receive more useful feedback on applicants. Instead of reporting whether a candidate scored high, medium, or low on a test battery, Employment Service offices can refer candidates on a top-down percentile ranking basis, which permits employers to select applicants with greater productive potential. This saves time and money, and makes it possible for the Employment Service to refer those candidates most capable of performing well in the job. It also makes the GATB, which is the most valid predictor of job performance, the primary decision maker rather than other procedures such as the interview, evaluations of training and experience, and the like, which typically have substantially less validity. It is also designed to increase the representation of high ability minority group members faster than alternative methods of selection. A pilot project has demonstrated that the progressive aspects of validity generalization represent exactly the kind of management assistance many employers want from the Employment Service. (PN)

USES TEST RESEARCH REPORT NO. 43

OVERVIEW OF VALIDITY GENERALIZATION FOR THE U. S. EMPLOYMENT SERVICE

DIVISION OF COUNSELING AND TEST DEVELOPMENT
EMPLOYMENT AND TRAINING ADMINISTRATION
U. S. DEPARTMENT OF LABOR
WASHINGTON, D.C. 20213

1983

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENT

The Validity Generalization research summarized in this report was conducted by Dr. John E. Hunter of Michigan State University in cooperation with the test research staff of the U.S. Employment Service. The results presented in this report have been collected from a series of technical documents authored by Dr. Hunter supported in part by contracts from the Test Development Field Centers. This project could not have succeeded without his technical expertise, leadership, and cooperation.

This report was prepared by Psychological Services Incorporated, 1735 Eye Street, N.W., Washington, D.C. under contract to the Southern Test Development Field Center, North Carolina Employment Security Commission, Raleigh, North Carolina. The authors of this report did not participate in the research. The report was prepared for printing by staff of the Western Test Development Field Center, Utah Department of Employment Security.

EXECUTIVE SUMMARY

The U.S. Employment Service can help improve the productivity of American industry on the order of 50 to 100 billion dollars in the upcoming year. Sound ridiculous? Not at all. In fact, the impact of the Employment Service on the U.S. economy can be accomplished by a new use of an already established and widely accepted Employment Service device for matching people and jobs - the General Aptitude Test Battery - in a way that makes the best use of state-of-the-art research evidence.

What we are referring to is local Employment Service use of the General Aptitude Test Battery (GATB) under concepts of validity generalization. For the past three years, leading research industrial psychologists have been collaborating with the research and development arm of the Employment Service. Drawing upon the massive GATB data base collected over the past 35 years, these scientists have created a new technology that truly puts the Department of Labor on the cutting edge of aptitude test research. The results of their efforts can play a major role in the country's economic recovery as well as advancing the professional field of testing. (Well-informed industrial psychologists are aware of this work which has been presented at a number of professional meetings.)

The U.S. Employment Service is now able to expand GATB coverage to all candidates for every job in the Dictionary of Occupational Titles (over 12,000 occupations). Before validity generalization, the GATB covered approximately 400 jobs. In addition, employers can now receive more useful feedback on applicants. Instead of reporting whether a candidate scored high, medium or low on a test battery, Employment Service offices can refer candidates on a top-down percentile ranking basis, which permits employers to select applicants with greater productive potential. This saves time and money, and makes it possible for the Employment Service to refer those candidates most capable of performing well in the job. It also makes the GATB, which is the most valid predictor of job performance, the primary decision maker rather than other procedures such as the interview, evaluations of training and experience, and the like, which typically have substantially less validity.

One reason for the decrease in national productivity over the past decade has been reduced effectiveness in personnel selection. Poor hiring decisions lower workforce productivity. Optimum test use alleviates this problem by assuring that only those with greater ability are recommended for selection. Valid selection bears a direct relationship to workforce productivity, and underlies all the validity generalization research. Validity generalization is designed to increase test utility (that is, the economic benefit to companies using the Employment Service) to a level few previously thought

attainable. It is also designed to increase the representation of high ability minority group members faster than alternative methods of selection.

There is evidence that even the current low level of test use by the Employment Service results in a productivity increase of $1.8 billion dollars annually over and above "random selection," or the use of selection procedures having no validity. Using the same model of test utility and administering the GATB consistent with validity generalization evidence the economic gains have been conservatively estimated to be 50 to 100 billion dollars. Evidence supporting this estimate is available in the validity generalization technical documentation.

The critical question for management at this point is whether it is operationally feasible and cost-effective for local Employment Service offices to make the necessary changes in procedures in order to effect the gains. A pilot project is currently being conducted in the North Carolina State Employment Service and monitored by the Southern Test Development Field Center. The results of this pilot project clearly indicate the potential gains in utility as well as the placement of minorities given the following conditions:

(1) testing all applicants with the GATB;

(2) classifying the employer's job into a special job family structure;

(3) using new composite test scoring procedures recommended by the validity generalization research;

(4) selecting in a manner consistent with the newly developed procedures.

The pilot project has already demonstrated that the progressive aspects of validity generalization represent exactly the kind of assistance many employers want from the Employment Service. Helping America towards economic recovery is a tall challenge, one not normally reserved for an employee selection program. Yet we can conceive of no other more logical or direct method of attack than at the level of workforce productivity, beginning with selection.

INTRODUCTION

Since 1970 it has been no secret that the productivity growth in the U.S. economy has declined at an alarming rate. An article appearing in Time magazine in early 1979 reported the rate dropping from about 3.5 percent a year to about 1 percent. Three years later the Commerce Department reported figures showing the nation's Gross National Product declining at 5.2 percent annual rate during the last quarter of 1981, with experts predicting further short-term declines (Washington Post, February 18, 1982, p.1.). Surely there are many causes for this productivity decline: economic, social, and political.

Recent research findings in industrial psychology point to what may be an overlooked cause of low workforce productivity: reduced effectiveness in selecting people for available positions. Contrary to popular belief, productivity differences between high and low-performing workers are great. In fact, the loss in resulting goods and services to the employer and to the economy as a whole due to less than optimal selection can be staggering. A recent study conducted for the National Science Foundation estimates the productivity loss per year to the economy from poor selection to be on the order of 80-100 billion dollars.[1]

The U.S. Employment Service has embarked upon a selection-oriented test improvement program heretofore unprecedented. Drawing from the results of the study described above for the National Science Foundation and others in that general area, the Employment Service, collaborating with research industrial psychologists, have produced a new, state-of-the-art approach to personnel selection based on ability. This approach utilizes the concept of validity generalization, and combines the results of 35 years of General Aptitude Test Battery (GATB) validity studies for individual jobs in the U.S. economy.

The purpose of this report is to further explain the validity generalization concept and the technical documentation underlying the development of the U.S. Employment Service's validity generalization testing program.

HISTORY OF THE GATB

The U.S. Employment Service developed the General Aptitude Test Battery (GATB) in 1947 for use in State Job Service local offices. The GATB itself is composed of eight paper-and-pencil and four apparatus tests designed to measure nine aptitudes found to be important for successful performance on most jobs. The nine aptitudes are as follows:

G - General Learning Ability
V - Verbal Aptitude

N - Numerical Aptitude
S - Spatial Aptitude
P - Form Perception
Q - Clerical Perception
K - Motor Coordination
F - Finger Dexterity
M - Manual Dexterity

The GATB was developed on the basis of statistical analyses of 59 different kinds of tests used in predicting job performance in a wide variety of occupations. The nine GATB aptitudes were selected because they provide adequate measures of all the major abilities measured by the 59 tests. Since 1947, the GATB has been involved in a continuing research program to validate the tests against successful performance in many different occupations and to insure that the tests meet all the professional standards and legal requirements. The research and development arm of the U.S. Employment Service has produced over 500 studies documenting the extent to which the GATB predicts future job performance, making the GATB the best validated test battery in existence for use in occupational selection.

Although the validity evidence for the GATB is most impressive, the number of jobs in the economy total approximately 12,000. New jobs are being created faster than individual validation studies can be conducted. Thus the Job Service could never hope to complete individual validation studies for all jobs in the economy. It appeared to be an insurmountable problem until the scientific breakthrough of validity generalization. GATB validity generalization research briefly summarized in the following pages has overcome that problem. The research shows that the GATB is in fact a valid predictor of successful performance for all 12,000 jobs.

UNDERLYING DIMENSIONALITY OF THE GATB

Before the validity generalization solution could be applied to the GATB, it was necessary to satisfy two statistical assumptions: (1) that the nine individual GATB aptitudes represent some orderly, underlying factor structure; and (2) that the validity evidence reported in over 500 studies can be attributed to these general underlying factors. Several multivariate statistical procedures were used to confirm these assumptions. The results of these analyses (based on large-scale sample sizes totaling over 23,000) demonstrated that the nine GATB aptitudes break into three general clusters or factors: a cognitive cluster containing the GVN components, a perceptual cluster defined by the SPQ components, and a psychomotor cluster made up of the KFM components (see Tables 1 and 2).

Table 1

Intercorrelations of Aptitudes and Their Reliabilities
(N=23,428)*

Aptitude	G	V	N	S	P	Q	K	F	M
G - General Learning Ability	1.00	.84	.86						
V - Verbal Aptitude	.84	1.00	.67						
N - Numerical Aptitude	.86	.67	1.00						
S - Spacial Aptitude	.74	.46	.51	1.00	.59	.39			
P - Form Perception	.61	.47	.58	.59	1.00	.65			
Q - Clerical Perception	.64	.62	.66	.39	.65	1.00			
K - Motor Coordination	.36	.37	.41	.20	.45	.51	1.00	.37	.46
F - Finger Dexterity	.25	.17	.24	.29	.42	.32	.37	1.00	.52
M - Manual Dexterity	.19	.10	.21	.21	.37	.26	.46	.52	1.00
Reliability	.88	.85	.83	.81	.79	.75	.86	.76	.77

*Source: USES, 1970, pp. 34, 269

Table 2

Confirmatory Factor Analysis of the Aptitude
Intercorrelation Matrix; Correlations Between Aptitudes and Factors

Aptitudes	Cognitive VN	Perceptual PQ	Psychomotor KFM
G - General Learning Ability*	--	--	--
V - Verbal Aptitude	.82	.68	.32
N - Numerical Aptitude	.82	.77	.42
S - Spatial Aptitude	.59	.61	.35
P - Form Perception	.64	.81	.66
Q - Clerical Perception	.78	.81	.54
K - Motor Coordination	.48	.60	.64
F - Finger Dexterity	.25	.46	.67
M - Manual Dexterity	.19	.45	.72

* G was left out of the analysis since it is not defined independently of V, N, or S.

** S was left out of the perceptual factor because it is closer to the cognitive factor than are P and Q.

The cognitive and psychomotor clusters are relatively independent, but both are highly related to the perceptual factor (Table 3). The multiple regression analysis in Table 4 confirms that performance on the perceptual factor is almost perfectly predicted by the cognitive and psychomotor components. This pattern or results justified combining the individual aptitudes into these three composites with no appreciable loss of predictive power, given the second assumption; that is, that the overall prediction of job performance was due to the general underlying factors rather than the individual specific components.

Table 3

Correlations Between the Factors
(N=23,428)

Factors		VN	PQ	KFM
Cognitive	VN	1.00	.88	.46
Perceptual	PQ	.88	1.00	.75
Psychomotor	KFM	.46	.75	1.00

Table 4

Multiple Regression of the Perceptual Factor
onto the Cognitive and Psychomotor Factors
(N=23,428)

Factors	Beta Weight
Cognitive	.68
Psychomotor	.44
Multiple Correlation: .96	

The evidence presented in Table 5 confirms that most of the GATB variance is attributable to the general factors. It follows that generalized composites are good predictors of job performance, as they account for most of the variance in test scores.

Table 5

Percentage of General, Specific, and Error Factor Variance
for the Nine Aptitudes and for the Three Composite Scores

		General Factor Variance			Specific Factor Variance	Error Factor Variance
		Cognitive	Perceptual	Psychomotor		
G	General Learning Ability	79			13	8
V	Verbal Aptitude	67			18	15
N	Numerical Aptitude	67			16	17
S	Spacial Aptitude		37		44	19
P	Form Perception		65		14	21
Q	Clerical Perception		65		10	25
K	Motor Coordination			41	45	14
F	Finger Dexterity			45	31	24
M	Manual Dexterity			52	25	23
GVN	Cognitive Composite	80			12	8
SPQ	Perceptual Composite		79		11	10
KFM	Psychomotor Composite			75	16	9

VALIDITY GENERALIZATION FOR 12,000 JOBS

In the very early days of occupational test validity research, psychologists implicitly assumed an orderly world where similar jobs have similar aptitudes and ability requirements. William H. Stead (1940), writing of the occupational research conducted by USES in the 1930s stated:

> It is believed that a proper appraisal of certain characteristics of job seekers, concerning which human judgment is more unreliable, requires the use of certain techniques or aids. Moreover, adequate information must be available concerning the relationship of worker traits to various occupational requirements.
>
> It is believed, furthermore, that occupations should be thought of in terms of their worker requirements and perhaps in terms of families of occupations for which similar worker characteristics are necessary. The grouping of occupations is particularly significant when one considers that there are probably 20,000 separate occupations. Until these occupations are arranged and understood according to common denominators or workers' skills, aptitudes, and other characteristics, the full range of employment opportunities cannot be made available to job seekers.

However, as data were collected on the relationship between occupational requirements and worker traits such as abilities and aptitudes, a wide variation was observed in the validity coefficients.

For years industrial psychologists and personnel selection specialists interpreted the wide variation in validity coefficients between studies where jobs and tests were identical as due to subtle job or job-context differences that the human job analyst or observer was simply unable to detect. As a result, it was impossible to accurately predict test validity in a particular setting. Empirical validation was recommended for every job-test combination. It was hypothesized that, although a test might be valid for determining the job aptitude of potential candidates for one organization, that same test might be invalid in another organization for the same job. These interpretations underlied the belief that test validity was situation-specific.

In the mid 1970's a new analytic procedure was developed which demonstrated most of the observed variability in validity coefficients for similar job-test type combinations was <u>not</u> real. Rather, this variance was due to a number of statistical artifacts; primarily sampling error inherent in small sample size studies, criterion unreliability, test unreliability, and

restriction in the range of test scores. By correcting for these sources of error using conventional statistical and measurement principles, it was shown that these statistical artifacts accounted for nearly all of the differences in validities across studies.[2]

Other sources of error are known to exist but cannot be corrected for using known statistical techniques. These are criterion contamination and deficiency, computational or typographical errors in published and/or unpublished studies, and slight differences in underlying factor structure between similar tests used to select for identical jobs. Because ways could be found to correct for only some of the statistical artifacts above, the true extent of validity generalization is underestimated.

Another belief that guided testing over the years is that tests are job specific; that is, a test might be valid for selecting machinists but not for cooks. This belief has also fallen under the weight of cumulative evidence using large samples of data -- cases numbering in the hundreds of thousands. The results now show that test validity is in fact stable across jobs and settings. Differences in the task structure of jobs do not cause aptitude tests to be valid for some jobs and not for others.[3] Therefore relatively small differences between jobs belonging to the same job type or job family will not produce large variations in test validity. As a result, the true test validity obtained after correcting for the various statistical artifacts can technically be "generalized" or "transported" over space and time.

The present research analyzed the cumulative results of over 500 individual validity studies taken from 35 years of U.S. Employment Service GATB research in order to determine the best job family structure for validity generalization purposes over all 12,000 jobs in the Dictionary of Occupational Titles (DOT).

Table 6

The Distribution of Observed and True Validity for Aptitude Composites Across the Entire Job Spectrum*

	Composites		
	GVN	SPQ	KFM
Mean observed validity	.25	.25	.25
Uncorrected standard deviation	.15	.15	.17
Corrected standard deviation (sampling error only)	.08	.07	.11
Observed 90 percent confidence interval	.05, .45	.05, .45	.03, .47
Corrected 90 percent confidence interval (sampling error only)	.15, .35	.16, .34	.11, .39
Mean true validity	.47	.38	.35
Standard deviation	.12	.09	.14
90 percent confidence interval of true validity	.31, .63	.26, .50	.17, .53

* True validities were obtained by correcting for sampling error, criterion unreliability, and restriction in range on the test.

Table 6 shows that all three GATB ability composites are valid predictors of performance in all jobs across the entire spectrum. However, the amount of validity varies substantially within each composite. A few validities are very low and others very high for each of the composites in some jobs (see 90 percent range). A uniformly higher level of prediction can be obtained if job families or groups of similar jobs are found for which there is a uniformly high level of ability substitution in prediction. This would entail finding those jobs or groups of jobs where, for example, one ability is more important relative to all others and can be used to account for more of the variation in performance.

Five strategies for forming job families were considered in order to determine the job grouping that most powerfully predicted composite aptitude validity. These were as follows:

(1) job analyst judgments of which aptitudes are important for job success;

(2) estimated mean aptitude ratings taken from the Dictionary of Occupational Titles);

(3) Department of Labor's functional job analysis (the "data-people-things" hierarchy used to determine the complexity of a job);

(4) Department of Labor's Occupational Aptitude Patterns (OAP); and

(5) the Position Analysis Questionnaire.

All five job analysis techniques were shown to predict observed composite aptitude validity along the order of r = .30. If this figure is corrected for sampling error, the correlation rises to about r = .45, representing a substantial improvement over considering all jobs together. The results of these analyses can be found in Hunter[4], who concluded that the functional job analysis approach best characterized the data when carried through to a full validity generalization solution. This job analysis method is described below.

A Brief Introduction to Functional Job Analysis

Functional job analysis (FJA) was used to develop and classify all jobs in the Dictionary of Occupational Titles. This job analysis technique identifies what a person does on the job and the results of one's behavior - that is, what gets accomplished. Worker task activities are defined in relation to "data" (information, facts, ideas and statistics), "people" (clients or co-workers), and "things" (machines or equipment). Every job in the DOT is classified by these three functions. Levels were developed for these functions representing varying degrees of complexity. Thus successive levels within the data function include comparing, copying, computing, compiling, analyzing, coordinating, and synthesizing. The people function includes signaling, persuading, instructing, negotiating, and mentoring. The things function is defined by the hierarchical arrangement of handling, feeding-offbearing, tending, manipulating, driving-operating, operating-controlling, precision working, and setting up. See Appendix.

The Relationship between the DOT Code and the Job Families Produced by FJA

The middle three digits of the nine digit DOT occupational code correspond to the "data-people-things" functions performed in any occupation. Since every job requires a worker to function to some extent in relation to data, people, and things, a separate digit (4th, 5th, or 6th) expresses the worker's relationship to each of these groups. The lower numbers in these three lists correspond to worker functions that involve more complex responsibility and judgment. Functions which are less complicated are assigned higher numbers.

Several FJA levels were merged and tested to produce a final optimum set of five job families:

Job Family Number	Name	DOT Code
1	Setting Up	Things = 0
2	Feeding, Offbearing	Things = 6
3	Synthesizing, Coordinating	Data = 0, 1
4	Analyzing, Compiling Computing	Data = 2, 3, 4
5	Copying, Comparing	Data = 5, 6

The "people" function dropped out of the final set, as it added little information beyond that contained in the "data" and "things" dimensions.

Table 7

Percentage Composite Contributions to Job Families

Job Family	GVN	SPQ	KFM
1	59	30	11
2	13	0	87
3	100	0	0
4	73	0	27
5	44	0	56

Table 7 shows the relative contribution made by each composite to the job families; that is, the importance of each composite score within a job family for validity. The results show that as the level of job complexity increases (job families 1, 3 and 4) the contribution made by the cognitive factor GVN increases relative to the psychomotor factor KFM. As job complexity decreases (job families 2 and 5) the contribution made by psychomotor is higher than cognitive. This job analysis solution locates those jobs for which there is ability substitution in predicting future job performance. Industrial "set up" work (job family 1) is the only job category under this system that is influenced by perceptual ability (SPQ).

Validity Generalization Results for Training and Job Proficiency Criteria

Drawing from the results of the "data-people-things" category analysis, the five job family set into which all jobs can be classified was carried forward to a full validity generalization study. This analysis was performed for both training and job proficiency criteria.

Table 8

Average True Validity Across All Jobs
(Corrected for Restriction in Range and Criterion Unreliability)
for Each Composite Ability by Job Performance and Training Criteria

Study Type	Number of Jobs	GVN	SPQ	KFM	Average
Training Success	90	.54	.41	.26	.40
Job Proficiency	425	.45	.37	.37	.40
Average	515	.47	.38	.35	.40

Table 8 presents the analysis of true validity across all jobs for both training and job proficiency criteria after correcting for sampling error, restriction in range, and unreliability in the criteria measures. Average true validity across all jobs for both training and job proficiency studies is .40. Training success shows a higher validity than does job proficiency for cognitive ability, while psychomotor ability is higher for job performance criteria. True criterion reliability has generally been found to be .80 for training success and .60 for job performance ratings.

Table 9

Average True Validity for Each Job Family by Criterion Type

Job Proficiency:

Job Family	Number of Jobs	True Validity GVN	SPQ	KFM	Best Single	Beta Weights GVN	SPQ	KFM	R
1	17	.56	.52	.30	.56	.40	.19	.07	.59
2	20	.23	.24	.48	.48	.16	-.13	.49	.49
3	36	.58	.35	.21	.58	.75	-.26	.08	.60
4	151	.51	.40	.32	.51	.50	-.08	.18	.53
5	201	.40	.35	.43	.43	.35	-.10	.36	.51
Average	425	.45	.37	.37	.48	.42	-.09	.27	.51

Training Success:

Job Family	Number of Jobs	True Validity GVN	SPQ	KFM	Best Single	Beta Weights GVN	SPQ	KFM	R
1	4	.65	.53	.09	.65	.57	.21	-.21	.68
2	0	--	--	--	--	--	--	--	--
3	24	.50	.26	.13	.50	.72	-.30	.03	.53
4	54	.57	.44	.31	.57	.57	-.07	.15	.58
5	8	.54	.53	.40	.54	.34	.17	.20	.59
Average	90	.55	.41	.26	.55	.59	-.10	.11	.57

Table 9 presents the average true validities and multiple regression analyses for the five job families, compiled separately for job proficiency and training criteria. The data show the effects of job complexity: as job complexity increases, the cognitive ability validity is high while psychomotor ability validity is low. The use of the best predictor composite for each job family would raise validity from .40 to .48 for job proficiency and from .40 to .55 for training success. Therefore there is a substantial increase in validity due to ability substitution.

Table 10

Best and Worst Case Analysis for Validity
in Predicting Job Proficiency and Training Success

Table 10a

The Best Case and Worst Case Analysis for Validity
in Predicting Job Proficiency

	Worst Case			Best Case		
Complexity Level	GVN	SPQ	KFM	GVN	SPQ	KFM
1	.52	.52	.25	.60	.52	.35
2	.15	.08	.21	.31	.40	.75
3	.38	.35	.21	.78	.35	.21
4	.31	.26	.12	.69	.54	.52
5	.36	.20	.24	.44	.50	.61
Average	.34	.24	.19	.56	.50	.54

Table 10b

The Best Case and Worst Case Analysis for Validity
in Predicting Training Success

	Worst Case			Best Case		
Complexity Level	GVN	SPQ	KFM	GVN	SPQ	KFM
1	.65	.32	.09	.65	.74	.09
2	--	--	--	--	--	--
3	.29	.26	.01	.71	.26	.25
4	.36	.34	.16	.78	.54	.46
5	.49	.53	.40	.59	.53	.40
Average	.37	.33	.14	.74	.47	.38

Table 10 shows a "best case" and "worst case" analysis of validities for all complexity levels for both job proficiency and training criteria. The "worst case" is the 10th percentile point of the distribution, that is, a value so low that only one in 10 validity values would be lower. The "best case" is the 90th percentile point of that distribution, that is, a value so high that only one in ten validity values would lie above that value. For example, the average validity of cognitive ability in predicting proficiency of setup work is .56 and the standard deviation is .03. For a specific setup job, the validity might be as low as .52 or it might be as high as .60.

Table 10a shows that even in the worst case, the validity of cognitive ability falls below .31 only for feeding and offbearing jobs. Even for these jobs, the worst case value is .15 which is considerably greater than zero. Thus, Table 10a shows that cognitive ability is a valid predictor of job performance for all jobs. Similar but less striking results are found for perceptual and psychomotor abilities. Table 10b shows similar results for training criteria. However, for psychomotor ability, the worst case drops to .01. Thus, there are high-complexity jobs where psychomotor may have no validity for predicting training success.

The focus on the worst case for validity is important both for theoretical reasons and practical since it bears on the issue of invalid prediction. The worst case analysis shows that well-constructed general cognitive, perceptual, and psychomotor tests are never invalid except in the one case noted above. However, the worst case analysis is a very slanted analysis from an applied point of view. The worst case value is deliberately chosen to be an unlikely value. The best case value is just as likely as the worst case value. The most likely validity values are the mean values shown in Table 9.

These validities can be shown to have very large workforce productivity implications, productivity that can be assessed in dollar terms.

GATB TEST UTILITY

The economic impact or benefit of valid selection procedures on workforce productivity is commonly referred to as test utility. For years it was assumed that personnel selection methods had little impact upon resultant employee performance and productivity. In other words, it made little difference in terms of employee performance how an organization selected its workforce so long as it hired from among the "qualified" applicants. This assumption, like those enumerated in the previous section, has proven erroneous when analyzed empirically. Standard regression-based methods can be used to estimate the gain resulting from different ways of using tests for selection; gains that can be measured in terms of dollars and cents.[5]

The classic formulas for deriving test utility have been available for years. The only condition for using these equations is that there must be a direct, linear relationship between GATB aptitude test scores and future job performance. Intensive investigation of over 3300 GATB test and job performance relationships involving 23,428 cases found evidence for non-linearity in only 5 percent of the cases; that is, at exactly the chance level. The overwhelming evidence on the GATB therefore supports the assumptions of linearity.

The average gain from the use of the GATB can be defined as the difference between average performance for those selected using the test and average

performance for those selected using alternative procedures. This average gain can also be defined in terms of the difference between optimal use of tests (ranking candidates on the basis of test scores and selecting top-down), and any other selection method (random selection of individuals with test scores above low minimum cutoffs, interviews, etc.). Each selection method has its own workforce productivity implications. The basic formula for the dollar benefit ($\bar{u}$ for utility) using valid selection tests is:

$$\bar{u}/\text{selectee} = r_{xy} s_y \bar{x}$$

where

r_{xy} = true validity of the test for predicting job performance,

s_y = standard deviation of true job performance in dollar terms, and

$\bar{x}$= average applicant test scores of those selected from the applicant pool.

The number r_{xy} is the correlation between test and job performance corrected for statistical artifacts, and is based on the full range of ability for applicants rather than the restricted range of those selected onto the job.

The number s_y is the standard deviation of yearly job performance measured in dollars. Based upon the results of previous empirical studies, this figure can be estimated conservatively at .40 of average annual salary or wage for the job.[6] This estimate is based upon test utility research for a wide variety of jobs.

The average test score $\bar{x}$ varies according to the proportion of candidates selected. The smaller the proportion selected, the higher the average test performance, given optimal (top-down) selection. The average test scores are computed in standard score form, with mean equal to 0 and standard deviation equal to 1. The total utility for a given year is the utility per person multiplied by the number of people selected over the course of that year and their average tenure.

An Example of GATB Test Utility: XYZ Corporation

Suppose the XYZ Corporation were to use the services of the U.S. Employment Service under the new validity generalization program as an aid in the selection of new workers. Further assume that the situation at XYZ Corporation requires that they select 20 entry-level machine operators from an applicant pool of 100. These machine operators will be paid an average annual wage of $14,500. Typically, the standard deviation of workers of this salary level is $5,800. Optimal test use under validity generalization

requires hiring the top 20 percent on the basis of GATB composite test score rankings, which corresponds to a mean standard test score of 1.39*. This particular machine operator falls in job family 1. Using the prediction equations shown in Table 9 based on validity generalization gives a multiple correlation of .59 for selecting machine operators with the GATB.
The gain in productivity that would accrue to XYZ Corporation under these conditions is:

$$\bar{U} \text{ /selectee} = r_{xy} s_y \bar{x}$$

$$\bar{U} \text{ /selectee} = (.59)(\$5,800)(1.39)$$

$$\bar{U} \text{ /selectee} = \$4,756.58$$

The number represents the gain per year per person selected. In this case XYZ hires 20 operators. The total increase in productivity represented by optimal GATB test use at XYZ Corporation for one year for 20 hirees is

$$U = N\, r_{xy} s_y \bar{x}$$

$$U = (20)(.59)(\$5800)(1.39)$$

$$U = \$95,131.60$$

The savings over a period of time can be figured by multiplying $95,131.60 by a tenure factor. If the average operator stays at XYZ Corporation for 3.6 years (the current average tenure), the total savings resulting from optimal test use for this one year is $95,131.60 times 3.6, or $342,473.76.

The U.S. Employment Service placed 4,022,019 applicants in jobs during 1980. If these candidates had been selected on the basis of validity generalization, productivity gains realized by the business community have been conservatively estimated at 98 percent over and above those selection procedures in place at the time. Thus the movement by the U.S. Employment Service towards optimal test use under validity generalization can increase potential workforce productivity by 50 to 100 billion dollars per year.

---

*The average test score in standard score form can be computed in two ways. First, it can be derived as the arithmetic average of the test score distribution. Second, it can be computed by an equation that depends on the normal curve. For example, if the situation requires hiring the top 20 percent of applicants, look up the normal curve table in a statistics book and find the height (ordinate) of the normal curve that corresponds to the proportion. This value is symbolized $\emptyset$. For this example $\bar{x} = \emptyset/p = .278/.20 = 1.39$.

Table 11

Percentage of Very Poor Workers Selected Given Optimal Test Use

	Selection Ratio				
Validity	.80	.50	.20	.10	.05
.30	8.0	5.8	3.9	3.0	2.4
.40	7.2	4.6	2.4	1.7	.12
.50	6.3	3.4	1.4	.7	.4
.60	5.3	2.3	.7	.2	.1
.70	4.4	1.3	.2	0	0

Table 12

Percentage of Workers with Promotion Potential Given Optimal Test Use

	Selection Ratio				
Validity	.80	.50	.20	.10	.05
.30	11.5	14.3	18.7	21.8	24.5
.40	12.1	15.7	22.1	26.8	31.6
.50	12.3	16.9	26.1	32.6	39.4
.60	12.3	18.1	30.2	39.4	49.4
.70	12.3	19.2	35.2	47.2	59.5

Table 11 depicts the percentage of very poor workers selected given optimal GATB test use as a function of selection ratios (the number of hirees to applicants) and validity coefficients. One notes that the extent of reduction in poor workers selected depends on the validity coefficient and the selection ratio. The higher the validity coefficient and the lower the selection ratio, the greater the reduction in the number of very poor workers selected. Under validity generalization, XYZ Corporation could reduce the number of poor selectees to 0.7 percent. By the same token, optimal test use can increase the percentage of workers selected who lie in the top talent category; that is, those with promotion potential. Table 12 presents these percentages in the same manner as shown in Table 11. Under the same conditions above, the increase in top talent would be on the order of 30.2 percent over and above random selection.

GATB TEST IMPACT

Clearly optimal use of the GATB under validity generalization maximizes potential productivity among those hired. A legitimate consideration for any organization is the impact of tests and other selection procedures on the racial, sex, and ethnic composition of the workforce. Companies have an obligation to maximize potential gain, but many are also very concerned about affirmative action and equal employment opportunity. This issue is a complex one. Opposing forces must be carefully weighed; that is, the inevitable trade-off between maximal workforce productivity and societal goals such as proportional minority representation throughout occupational structures. Various schemes exist for achieving racial balance; however, they all reduce the average productivity of those selected to varying degrees. One such example is a preferential selection system based on top-down hiring within each subgroup using valid tests. This particular scheme has been shown to increase minority employment faster than selecting at random among all "qualified" candidates and, in addition, incurs much less economic cost to the organization.

Racial and ethnic groups do not have identical scores on ability tests. The GATB is no exception. Table 13 shows the relationship in standard score form (mean = 0, SD =1) of racial and ethnic group means on GATB general abilities relative to majority group standard scores. Although average GATB test score differences exist between racial and ethnic subgroups, this does not imply that people in the high average group all score higher than those in low mean groups. There are people in all groups found at every level of ability. However, groups with lower mean ability will have proportionately fewer candidates selected under optimal test use procedures.

Table 13

Racial and Ethnic Group Means Expressed in Majority Group Standard Scores

	Cognitive GVN	Perceptual SPQ	Psychomotor KFM
Majority	.00	.00	.00
Asian/Pacific Islander	-.13	-.02	.34
Hispanic	-.51	-.29	.18
Black	-.75	-.67	-.23
American Indian	-.85	-.16	.27

Table 13 indicates that group differences are more pronounced on cognitive ability than either the perceptual or psychomotor abilities. Note for

psychomotor ability, three of four minority subgroups have a higher average standard score than the majority. This is important because the validity generalization research has shown psychomotor ability to be a better predictor than cognitive ability for many jobs. If psychomotor ability is used to select for these jobs then adverse impact is greatly reduced. If the cutoff score were set to select the top 50 percent of the majority candidates (white applicants), then 57 percent of the Hispanic applicants would also be included, as opposed to only 20 percent if cognitive ability is used. In this sense the GATB is unique: it finds jobs where minority groups are hired at faster rates at no expense of validity or economic benefit to the organization.

Although adverse impact exists, subgroup differences are less extensive as one moves from cognitive to psychomotor abilities. However, use of the GATB test will not produce a racially or ethnically balanced workforce, unless there is proportional hiring within subgroups. Does this mean that the GATB is unfair to minorities in the sense that they underestimate their true ability? This is a question that can be answered based upon the specific GATB research reported below on differential validity and test fairness.

Differential Validity

The differential validity hypothesis states that the GATB will be less valid for minorities than for whites. It is tested by applying statistical tests between observed validities. Most reviews in the industrial psychology professional literature report differences between subgroups occurring about 5 percent of the time; that is, at only the chance level of frequency. Over the past 10 years, the U.S. Employment Service completed 51 validity studies with enough minority applicants to test the differential validity hypothesis. Each study produced nine opportunities to observe differential validity by correlating each aptitude or ability with a measure of job performance. Out of 459 opportunities to observe differential validity, only 31 significant differences occurred, which is 6.75 percent. This difference is trivial since 5 percent is the chance level.

Thus there is no differential validity in predicting job performance with the GATB. This mirrors the evidence for the field as a whole, which indicates that employment tests are equally valid for all groups.

Test Fairness

Even if validity coefficients are statistically equal for minority and non-minority groups, tests are likely to be perceived unfair, especially if average test scores are lower for minorities. The theory behind this perception is that minorities systematically miss culturally biased items. It would follow from this hypothesis that test scores underestimate minorities' true ability to perform the job. Underprediction of minority job

performance is the core of the most commonly accepted model of test fairness. This model defines a test as unfair to a minority group if it predicts lower levels of job performance than the group actually achieves. Thus the factors causing lower test performance would not be reflected in actual job performance.

Table 14

Overprediction Using the Specific Aptitudes of the GATB

		Partial Correlation	Observed Overprediction	Corrected Overprediction
G	General Learning Ability	.10	.23	.12
V	Verbal Aptitude	.14	.31	.20
N	Numerical Aptitude	.13	.28	.16
S	Spatial Perception	.14	.31	.17
P	Form Perception	.14	.30	.19
Q	Clerical Perception	.16	.34	.24
K	Motor Coordination	.19	.39	.38
F	Finger Dexterity	.17	.67	.59
M	Manual Dexterity	.18	.31	.28

Just the opposite situation occurs with the GATB. Job performance levels are not underpredicted by test scores. To the contrary they are overpredicted. Table 14 shows the extent of overprediction using cumulative GATB results for each of the nine aptitudes. All overpredict to a varying degree. Table 15 shows the extent of overprediction for the composite scores GVN, SPQ, and KFM. The extent of overprediction is smaller for the ability composites than for the specific aptitudes.

Table 15

Overprediction Using the GATB Ability Composites

		Partial Correlation	Observed Overprediction	Corrected Overprediction
GVN	Cognitive	.11	.25	.18
SPQ	Perceptual	.10	.22	.13
KFM	Psychomotor	.16	.33	.30

Preliminary calculations have shown that composite abilities will have negligible overprediction for jobs to which they are highly relevant and large overprediction where they have low relevance. Therefore once the correct ability composite is used for a given job, differences in average job performance between groups with the same test scores disappear. To summarize, the GATB research shows that the average ability differences between minority and majority groups are reflected in job performance and are therefore real differences, not pseudo-differences caused by tests, and that ability composites are more accurate predictors of job performance than specific aptitudes.

CONCLUSION

Historically two key factors restrained employer use of the U.S. Employment Service for listing job orders. One was the limited ability of local Job Service offices to refer those applicants most likely to be productive workers. Instead local offices referred only those candidates qualifying above minimum selection cut-offs without knowing where applicants stood in relation to others taking the test. The second major factor limiting the Job Service testing program was the lack of occupational coverage. Job success could only be predicted for roughly 400 occupations of a total of 12,000 included in the Dictionary of Occupational Titles.

Under validity generalization, it is now possible to predict occupational success for all applicants in all jobs in the U.S. economy. The newly developed technology makes optimal test use a reality and provides more useful test information to employers.

The economic impact or benefit of optimal test use under validity generalization on workforce productivity can have profound effects on individual employers as well as the total economy. Looking at the total perspective economic gains along the order of 50 to 100 billion dollars can and should be realized through the Job Service testing program. In addition, the data collected over the past 35 years on the General Aptitude Test Battery (GATB) substantiates the fact that it is a fair and valid personnel selection device.

REFERENCE NOTES

1. Hunter, J. E., and Schmidt, F. L. Fitting people to jobs: the impact of personnel selection. In E. A. Fleishman (Ed.) Human Performance and Productivity. Hillsdale, N.J.: Lawrence Erlborough Associates, 1982.

2. Schmidt, F. L., and Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

3. Schmidt, F. L., and Hunter, J. E., and Pearlman, K. Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 1981, 66, 166-185.

4. Hunter, J. E. The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance. Report to the U.S. Employment Service, 1982.

   Hunter, J. E. Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB). Report to the U.S. Employment Service, 1982.

   Hunter, J. E. 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. Report to the U.S. Employment Service, 1981.

   Hunter, J. E. Fairness of the General Aptitude Test Battery (GATB): ability differences and their impact on minority hiring rates. Report to the U.S. Employment Service, 1981.

5. Schmidt, F. L., Hunter, J. E., McKenzie, R., and Muldrow, T. The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 1979, 64, 609-626.

6. Mack, M. J., Schmidt, F. L., and Hunter, J. E. Estimating the productivity costs in dollars of minimum selection test cut-offs. Washington, D.C.: U.S. Office of Personnel Management, in press.

7. U.S. Employment Service. Section III: Development Manual for the USES General Aptitude Test Battery. U.S. Department of Labor, 1970.

APPENDIX

DATA (4th digit)	PEOPLE (5th digit)	THINGS (6th digit)
0 Synthesizing	0 Mentoring	0 Setting up
1 Coordinating	1 Negotiating	1 Precision Working
2 Analyzing	2 Instructing	2 Operating-Controlling
3 Compiling	3 Supervising	3 Driving-Operating
4 Computing	4 Diverting	4 Manipulating
5 Copying	5 Persuading	5 Tending
6 Comparing	6 Speaking-Signaling	6 Feeding-Offbearing
	7 Serving	7 Handling
	8 Taking Instruction-Helping	