Employment testing: The U.S. Job Service is spearheading a revolution

R M Madigan
Dow Scott, Loyola University Chicago
D L Deadrick
J Stoddard
Employment Testing: The U.S. Job Service is Spearheading a Revolution

Considering validity generalization in employment decisions

By Robert M. Madigan, K. Dow Scott, Diana L. Deadrick and Jil A. Stoddard

For most of the past 20 years, the use of psychological tests in selection and promotion has diminished steadily. Faced with the difficult and often expensive task of developing evidence of test validity, employers frequently have opted to discontinue or forego testing. This trend is now being reversed. Surprisingly, impetus for this resurgence in testing is partially attributable to a federal agency, the U.S. Employment Service (USES). The validity generalization (VG) testing program currently being implemented by USES through local Job Service offices in many areas of the country sharply contradicts prevailing beliefs regarding the use of employment tests. The central tenets of the program are that standardized ability tests are fair and valid predictors of performance for all jobs, and that such tests provide employers (and the nation) with a powerful tool for improving work force productivity. This article provides an overview of the VG program, its relationship to current employment testing regulations, the potential benefits to employers of using VG, and some issues associated with the program.

Validity Generalization

The term “validity generalization” is drawn from an extensive and controversial stream of research launched in the 1970s by John E. Hunter (Michigan State University) and Frank L. Schmidt (then of the U.S. Office of Personnel Management). Their conclusion that standardized tests of cognitive (mental) ability are generally valid for employment selection decisions provides the basis (and the name) for the USES program. A brief overview of their research and findings as they apply to the USES program is provided here.

The research by Hunter and Schmidt can be described as a reinterpretation of findings from previous investigations of the validity of employment tests. Thousands of validation studies have been conducted over the past 50 years, but the findings have been inconsistent. Similar tests for similar (or identical) jobs in different settings often yielded widely varying validity coefficients (correlations between test scores and job performance). This apparent inconsistency led to the generally accepted conclusion that the validity of an employment test for a particular job is specific to the situation. As a result, employers usually have been required to develop evidence of the validity of any tests used, regardless of findings elsewhere.

Hunter and Schmidt tested an alternative explanation for the historical inconsistency in test validation studies. Noting that the number of workers (sample size) in most previous studies was relatively small, and that the results of such studies are potentially influenced by a variety of technical deficiencies, they reasoned that inconsistent findings could be primarily attributable to statistical errors. Using a meta-analysis research method that provides a means to correct or adjust for such sources of error,
Schmidt, Hunter and associates developed a persuasive body of evidence that the validity of ability tests for employment screening is relatively stable across jobs and organizations. Subsequent studies also strongly indicated that the validity of ability tests is generally higher than that of other selection procedures such as interviews, reference checks, experience ratings, etc.

In one study, Hunter cumulated the results of 515 studies of the validity of the General Aptitude Test Battery (GATB) carried out over a 45-year period by the Employment Service. Three general abilities (cognitive, perceptual, and psychomotor) derived from GATB scales were found to be valid predictors of job proficiency for all jobs. Moreover, although there was considerable variation in the validity of these three abilities across jobs, the differences were adequately accommodated by grouping all jobs into five broad job families reflecting a hierarchy of job complexity. Since this job sample was representative of the 12,000 jobs included in the Dictionary of Occupational Titles, the findings in this study were generalized to all jobs.

In a series of related studies, Hunter and the Employment Service addressed the topic of the fairness of psychological tests. The issue of test fairness actually consists of two distinct questions: (1) is the test fair in the sense that it is an accurate estimate of job performance ability for all applicant sub-groups? and (2) does use of the test adversely (disproportionately) limit job opportunities for minority group members?

Research evidence pertaining to question (1) is overwhelmingly positive. Virtually all studies conducted over the past 15 years by numerous researchers have concluded that ability tests are fair to minority groups. In fact, with respect to the GATB, an analysis of 51 studies showed that use of the GATB to predict job performance was likely to overestimate, rather than underestimate, the job performance of blacks.

The answer to question (2) depends upon the extent of the difference in average test scores between nonminority and minority members on the particular tests. Significant differences in ability test scores between groups are nearly universal; the GATB is no exception. Although these differences vary by minority group and type of ability, the difference in average scores between nonminorities and blacks is substantial for all three general abilities. For example, if the cutoff score on cognitive ability was set at the average of the majority group, only 25 percent of black applicants would pass. The figures for perceptual abilities and psychomotor abilities are 25 percent and 41 percent, respectively. Thus, use of the three general ability scores based on the GATB could be expected to produce a racially unbalanced work force.

In summary, the validity generalization program is based on the following research findings:

1. GATB scores are valid predictors of job proficiency although validity varies somewhat by type of job.
2. GATB scores fairly estimate job abilities for all groups.
3. Use of unadjusted GATB scores will adversely affect job opportunities for minorities.

The GATB

The General Aptitude Test Battery was put into use in 1947 by state employment services offices after an extensive developmental research effort. It has been the focus of a continuing program of research to refine the tests and validate their use for vocational counseling and employee selection. Hence the GATB has a broad research record for many occupational classifications, and is probably the best validated multiple aptitude test battery in existence for employment applications.

The test battery consists of 12 timed tests (parts) - eight paper and pencil tests and four apparatus tests. Each part requires performance of familiar tasks such as name comparisons, arithmetic computations and reasoning, form perceptions, pegboard manipulations, etc. The 12 parts yield eight specific aptitude scores: verbal aptitude, numerical aptitude, spatial aptitude, form perception, clerical perception, motor coordination, finger dexterity, and manual dexterity. In addition, a general intelligence or learning ability score is derived from the arithmetic reasoning, vocabulary, and spatial perception test scores. These nine aptitudes measure basic abilities or capacities to learn various jobs.

Past research on the GATB had concentrated on developing norms (standards) for specific occupations and identifying patterns of aptitudes for occupational families. Norms were developed by identifying the specific aptitudes most relevant to particular occupations, and determining the degree to which these aptitudes related to proficiency in job performance or success in occupational training. A minimum or “cutting” score for each aptitude was set at the level that would effectively screen out a majority of the potentially unsuccessful workers for that occupation. The combination of aptitudes and minimum scores that comprised a norm differed for each occupation.

Use of GATB data in the VG program differs from the historical approach in three key respects. First, specific aptitude scores are combined into the three general ability scores noted above. These ability scores are then weighted and combined to produce overall aptitude scores for each of the five broad job families that Hunter had previously identified. Thus, minimum or critical scores on any specific aptitude are eliminated.
Second, the overall score for each job family is converted to a percentile score to facilitate the use of a top-down hiring strategy (hire the best qualified), as opposed to use of the test scores to screen out the unqualified. Third, separate rankings (within-group percentiles) are computed for nonminorities, blacks, and hispanics. In effect, this procedure adjusts the GATB scores of minorities to eliminate adverse impact.

The Job Service Program
Implementation of VG in local offices of the Job Service currently varies within and between states. For example, local Job Service offices in Michigan tested 73,600 applicants during 1985, and over 26,000 VG-selected applicants were referred to employers. Other states are just beginning to implement the program, but at least 25 states offer VG testing in some locations. A number of localities are following the “full implementation” model piloted in Roanoke, Virginia, and recommended by USES. In this approach, testing replaces interviews by local office personnel as the basic tool for all applicant screening and referral activities. As a consequence, significant operational changes are required, both in the nature and mode of service to applicants, and in the role of the local office staff. To achieve maximum efficiency, the Roanoke office tailored all internal procedures around VG. Because the GATB is most efficiently administered in groups, individual applicant registration and interviewing were replaced by group registration (intake) and testing sessions. In the daily intake sessions, new applicants are given assistance in completing the necessary forms, and an orientation to program procedures and the nature of the GATB is provided. All applicants are encouraged to take the GATB, which is administered daily at a centrally located testing center.

As of January 1986, 80 percent of applicants were being tested. Results of the testing are entered into the local office records (computer files), and applicants are sent a report of their percentile rank for each of the five job families. These percentiles are interpreted as an indicator of the applicant’s relative suitability (aptitude) for jobs of that type. When a job order is received, the job family to which it belongs is determined and a file search is initiated to identify and rank applicants for referral to the job. Other criteria established by the employer and preferences of the applicant also are considered in making referrals. For instance, an employer might require a specific type of training or experience, or an applicant’s availability for certain hours or types of work could be restricted. Referrals would then be based on percentile ranking within these constraints.

Implementation of the VG program significantly changes the role and activities of Job Service interviewers. Much of the time formerly spent on initial assessment interviews is devoted to matching applicants’ records with employer requirements. Interviewers are encouraged to become more knowledgeable of the particulars of the jobs they fill and the employers they serve. In order to facilitate this, Roanoke interviewers are designated as “account representatives” for specific employers, with responsibility for all transactions with that employer. According to officials in the Roanoke office, these changes have resulted in increases in the number and quality of candidates who are referred to employers. This improved responsiveness to employer needs has led to increased use of the job service by employers.

VG and the “Uniform Guidelines”
The VG program appears to challenge standards in the Uniform Guidelines for Employee Selection Procedures in a number of key respects. First, as noted above, the basic premise that the validity of ability tests is specific to particular situations is rejected. Employers using the VG program are informed that the VG testing program is valid and fair. Hence, they are relieved of the burdensome responsibility for conducting validation studies (of the GATB) in their organizations.

Second, because the three general abilities are believed to be valid predictors of success for all jobs, the detailed job analysis required by the guidelines is unnecessary. The analysis need only be sufficient to classify jobs into their appropriate job family. Since the worker function ratings in the occupation code of the Dictionary of Occupational Titles (DOT) provided the original basis for grouping jobs into five families, and the 515 jobs in the study were representative of the 12,000 jobs in the DOT, the basic job analysis for most jobs in the U.S. economy already exists.

Third, the VG program uses test scores to rank applicants. Research by USES has shown that ability test performance is directly related to job performance, i.e., as test scores increase there is a corresponding increase in average job performance. This finding appears to provide adequate justification for a “top-down” selection strategy. On the other hand, the thrust of the guidelines has been to “encourage” use of the minimum score (low cutoff) method.

Finally, the use of within-group percentile scores for non-minorities, blacks, and hispanics appears to be proscribed by the disparate treatment provision of the guidelines. Separate rankings are clearly a case of subjecting different racial groups to different standards (albeit for the purpose of eliminating adverse impact). The guidelines appear to allow use of different standards only in situations where a test is potentially unfair.
In short, the VG program is consistent with the ultimate objective of the EEOC, to increase employment opportunities for minorities. In fact, if the VG program is used to supplant less objective screening procedures, the overall selection ratio of minorities should increase over current levels. However, this is accomplished by means of a race-conscious hiring procedure based on premises that have heretofore been explicitly rejected by the EEOC. It is conceivable that the VG program, or more specifically, the procedures and findings of Hunter and Schmidt, will be defined as new, “professionally acceptable techniques” (under Sec. 14 of the guidelines). However, the silence of the EEOC to date and the continuing debate over validity generalization within the academic community suggest that this is unlikely.

**Benefits to Employers**

Few employers would challenge the proposition that the productivity of good workers significantly exceeds that of poor workers. The difference in productivity between top and bottom workers can vary widely by job or organization, but research findings suggest the ratio is about two to one for the typical job. Since valid ability tests increase the probability of selecting successful employees (better than other known selection devices), it follows that the use of valid tests to rank applicants provides an economic benefit to employers.

Documentation of the dollar value of productivity improvement attributable to testing has been difficult and expensive to develop in the past. Attempts to translate productivity differences among employees into dollar terms required detailed cost accounting procedures, and this approach was only applicable to a limited range of jobs. However, reasonable (conservative) estimates are now readily calculated for most jobs by means of recently developed procedures. Hence, analyses of the monetary value (utility) of VG testing to individual employers will be forthcoming as the program is implemented more widely.

Application of this type of analysis to federal hiring data for 1980 indicates that optimal use of the GATB for all hiring decisions in that year would have had a value to the government of over $1 billion.

**Surprisingly little discussion or debate over the consequences of nationwide implementation of the VG program has surfaced in the professional or popular press.**

or approximately 16 percent of total wages. Although an estimate of this magnitude naturally invites skepticism, it is actually based on conservative assumptions regarding the validity of VG test scores and variance in worker productivity. Furthermore, the figure escalates substantially if the continuing effects of improved selection (average tenure) are included in the calculation. Thus, the economic benefits of testing are likely to be much larger than was previously believed.

To some employers, a promise of increased worker productivity through better selection of qualified applicants will be viewed with misgiving or disinterest. Productivity improvement often is primarily a function of technology, and in some instances the contribution of workers to total productivity is virtually impossible to determine. However, cost reduction is of interest to all employers, and there is solid logic and evidence to support the use of VG to reduce costs. For instance, the validity of the GATB (and other ability tests) for predicting successful performance of applicants in various types of training programs is well established. Improved selection of trainees results in fewer failures, which translates directly into savings in training and employment costs. Similar reports of the effectiveness of VG have been received by the authors from Roanoke area employers, and are frequently reported in other VG pilot project areas. For example, a study by Philip Morris found a significant increase in the rate of training success for employees hired via a selection procedure incorporating VG over that of two comparison groups.

The VG program offers other possible monetary and intangible benefits to employers. For example, one Roanoke area employer attributes a dramatic reduction in turnover of reduced administrative costs by using the Job Service as a sole source of job applicants. Fewer applicants are processed by the company and the quality of referrals has often increased, as evidenced by a reduction in the number of referrals per placement for many of these employers. Furthermore, the close working relationship between employer and Job Service personnel engendered by this procedure facilitates responsiveness to employer needs by the local Job Service office.

While the potential benefits appear to be significant, employers must recognize that the VG program is not a panacea. Test information is only one factor in employment decisions. The organization also must improve validity and fairness in other aspects of the selection decision process. Moreover, ability is only one component in the employee performance equation. Management practices often determine whether those abilities are properly developed, and whether the employee's energies are focused and tapped. In other words, tests might be valid predictors of job performance and training success, but the extent to
which organizations capitalize on improved selection procedures depends to a large degree on how well they manage their human resources.

A standard caveat pertaining to the use of employment tests also applies to the VG program. To state that ability tests have high validity for predicting successful job performance is not to imply that they are perfect predictors. On the contrary, it merely suggests that the probability of mistakes in assessing the relative abilities of applicants is reduced. Predictions for particular applicants will sometimes be wrong, but over the long run the proportion of "correct" decisions will exceed that of decisions made without benefit of test information. Hence, organizations with a high level of employment activity will more quickly realize the benefits of VG screening.

**Issues**

Surprisingly little discussion or debate over the consequences of nationwide implementation of the VG program has surfaced in the professional or popular press (perhaps the program has not yet captured sufficient attention). Whatever the reason, VG testing is likely to renew or reinvigorate debate over some basic policy issue.

For example, the use of within-group percentiles to rank applicants appears to be at odds with the position taken by other federal agencies. Not only could separate rankings for nonminorities, blacks, and Hispanics constitute disparate treatment, but this procedure also raises the more emotion-laden issue of preferential treatment. Given equivalent qualifications (in terms of test performance) for a minority and a nonminority candidate, the minority candidate will be ranked higher. This procedure is a logical and reasonable compromise between the conflicting national goals of economic efficiency (hire the best qualified) and equal economic opportunity for members of minority groups. However, by explicitly using different standards, the program also appears to be inviting charges of reverse discrimination. It is interesting to speculate how the current leadership of the EEOC and the Justice Department might view this particular form of quota in view of their adamant opposition to preferential hiring under any conditions.

---

**Local Job Service offices in Michigan tested 73,600 applicants during 1985, and over 26,000 VG-selected applicants were referred to employers.**

A second issue pertains to the basic mission of the Job Service. The VG program very closely aligns the agency with employer perspectives and values. Obviously, job orders from employers are necessary to meet the needs of job seekers, and VG holds considerable promise for expanding the number and type of job orders placed with local offices. However, the emphasis on identifying the best qualified means some less qualified (in terms of test performance) individuals will be excluded from referral to jobs that they could perform successfully. The question boils down to one of the proper role of the public employment system. Should the Employment Service focus primarily on those most in need? (This was its function in the '70s, and it was not a marked success. Individuals most in need often are not competitive in the job market. As a result, employers found they could get better applicants elsewhere.) Or, should the Employment Service concentrate on helping employers make optimal use of the nation's human capital? An affirmative answer here raises a question of individual rights: Should a government agency be so closely identified with employer interests in decisions that affect people's life chances?

From the perspective of individual applicants, there is little doubt that widespread adoption of VG will directly affect the distribution of economic winners and losers. Some of the able, but chronically unemployed, will have opportunities that were previously denied them. (The Roanoke Job Service staff can attest to this fact.) On the other hand, top-down referral of applicants based on VG scores will exclude people who would have been considered if minimum standards were used. The consequences for the less able (and the developmentally disabled) could be severe. Moreover, extensive use of VG testing could exacerbate the "permanent underclass" problem in our society by creating a new class of stigmatized unemployed, the "low scorers" (some of whom will be erroneously labelled).

If the VG program is to be successfully implemented on a national level, problems of this genre will need to be addressed. While such problems will undoubtedly prove difficult to resolve, they must be kept in perspective. The previous approach based on interviewing was less effective and was plagued with similar and equally formidable problems. On the other hand, VG testing holds the promise of significantly contributing to productivity improvement, one of the keys to stemming the flow of jobs offshore. If this is borne out, the net effect of nationwide implementation of VG on job opportunities could be positive.

**Conclusions**

From the perspective of individual employers, the VG program appears to be one of those rare "no lose" opportunities. It offers a number of potential benefits with little perceptible risk or cost. As noted above, the extent to which a particular employer can realize
benefits from the program depends on a number of factors.

However, from the evidence to date it appears that the majority of employers would be well advised to take two actions. First, they should carefully re-examine their beliefs and/or practices with respect to the use of ability tests in employment decisions. If employment activity is high, they could be missing an opportunity to significantly improve the quality of selection decisions.

Second, employers should re-evaluate their use of the Job Service. They might be under-utilizing a valuable resource. With the advent of the VG program, many local Job Services offices can provide a comprehensive applicant screening service at no cost.

References
1. Readers unfamiliar with this research can find a thorough summary and extensive references in the USES Test Research Reports 45-49 available through the Job Service officials in their area. Also, Personnel Psychology, Winter 1985, provides an overview of many of the issues raised with respect to VG in a question-and-answer format.
5. Ibid.