A Concurrent Validation Study of

The United States Employment Service's

Validity Generalization Job Family Four Scores

by

David J. Hoover

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Management

APPROVED:

Robert M. Madigan, Ph.D., Cochair

Frederick S. Hills, Ph.D., Cochair

Kent F. Murrmann, Ph.D.

K. Dow Scott, Ph.D

Dianna L. Stone, Ph.D.

August 28, 1987

Blacksburg, Virginia

A Concurrent Validation Study of

The United States Employment Service's

Validity Generalization Job Family Four Scores

by

David J. Hoover

Robert M. Madigan, Ph.D., Cochair

Management

(ABSTRACT)

The United States Employment Service has implemented a recently developed testing program. The Validity Generalization (VG) Testing Program, adopted its name from the meta-analytic technique which cumulates the findings of test validation studies. For this testing program, predictors were developed for five job families based on a validity generalization study of 515 validation studies. The Employment Service claims that these predictors are valid and virtually all jobs are covered in the five job families.

This study is a direct test of the validity of one of the five predictors, Job Family IV Validity Generalization percentile scores. (The Employment Service estimates its true validity is .53.) Secondly, two potential moderators of that predictor's validity were investigated: jobs and locations. Three "computing and account recording" clerical jobs and two locations were examined. Finally, evidence of whether general abilities were better predictors of performance than specific abilities was examined, since the testing program's predictors are comprised of composite, general ability scores.

A concurrent validation study was conducted with 219 clerical bank employees. Two predictors, the General Aptitude Test Battery (GATB) and the American Bankers Association's test battery, were administered. Two criteria measures were obtained, supervisory ratings on the Descriptive Rating Scale for all subjects, and, objective measures - strokes per hour - for proof operators.

The observed validity for Job Family IV's predictor with the global DRS criterion was .16, Observed validity with a composite of DRS dimensions was .19. Corrected for attenuation, those

coefficients were .18 and .20 respectively. However, general cognitive ability measures appeared to be slightly better than the percentile scores at predicting performance.

While there was no statistical evidence of moderators, the relatively small effect size resulted in low power for the tests and may account for the results. Nevertheless, the evidence raised questions about the possible existence of situational moderators.

Finally, measures of general ability did not appear to predict performance better than measures of specific abilities.

Acknowledgements

In a project of this magnitude, many persons have contributed to its successful completion. As a Christian, I acknowledge the physical health and stamina, along with the innate abilities provided by God. Next, I acknowledge the strong emotional support provided by my extended family. In particular, my wife Judi was very patient through the long process. Without complaint, my children, Jonathan and Allegra, became more self-reliant as Dad was often at the office researching or writing. Both my parents, Jack and Jane Hoover, and Judi's parents, Dr. Walter and Ellen Craymer, provided many meals and overnight accommodations as well as encouragement. And, my Uncle Paul and Aunt May Petzold frequently took the extended family to dinner, giving us all a welcome break.

Next, I acknowledge the direction and encouragement provided by my committee members, Drs. Robert Madigan, Fred Hills, Kent Murrmann, Dow Scott, and Dianna Stone. Particularly, I want to recognize the many hours Bob Madigan invested in my academic development during the dissertation process.

Encouragement also came from my colleagues at Miami University. In particular, Dr. John Douglas, chair of the Management Department, was sensitive in his encouragement, not placing dysfunctional stress upon his ABDs. Drs. Don Bolon, my mentor, and Sandra Jennings were also

very supportive, covering classes while I defended this dissertation. I also want to thank former fellow ABD, Dr. Kay Snavely, for timely visits of encouragement.

A Virginia Tech ABD, Jerry Fox, was most supportive, giving me access to his office, tips on the computer system and software, and making my visits to Blacksburg as comfortable as possible during the College of Business's construction phase. Lastly, I want to thank Mike and Vicki Powell for opening both their hearts and home to their old friend from Ohio, particularly during my extended visits this summer.

Table of Contents

List of Illustrations

List of Tables

CHAPTER 1: Introduction

The harsh economic by-products of international competition are forcing business, labor and government to address the relative decline in American productivity. These by-products include: falling demand for products "made in the USA," plant closings, the trade deficit, structural unemployment, shrinking tax bases, and increased demand for governmental entitlement programs. As the United States' technological supremacy in the marketplace wanes, the importance of human resources to organizational and national productivity becomes increasingly apparent.

One approach to productivity improvement entails improving allocation of human resources. The United States Employment Service, created under the Wagner-Peyser Act (1933) to facilitate better allocation of the nation's human resources, has recently re-examined the use of testing in selection. As a result of that research, a new testing program was developed. If implemented nationally, the USES has "conservatively estimated" the utility of its VG testing program could be $50-100 billion (USES test research report no. 47, 1983). The objective of the "Validity Generalization" (VG) testing program is to improve productivity at both the individual business and national levels by improving the quality of selection decisions. The major focus of this study is to investigate the validity of this VG testing program.

This study is timely because the Employment Service is in the process of implementing its VG testing program nationally. Consequently, many employers are faced with policy decisions

concerning participation in the program. However, those who have sought evidence of the program's effectiveness have discovered that contrary to usual practice no direct tests of the validity of VG testing have been conducted.

Published research on validity generalization has consisted of only two types: meta-analytic studies that cumulate extant validation studies (an indirect or residuals approach), and Monte Carlo simulations. Thus, this field study addresses a void in the validity generalization literature, providing preliminary evidence of the VG testing program's validity.[1] Consequently, results of this study are of interest to human resources practitioners.

Results of this study are also of interest to academics since the VG program has been developed around two controversial assumptions. First, there are conflicting views on the extent to which general ability test validity can be generalized. The "old" view is that generalization is very restricted (Uniform Guidelines, 1978, Guion, 1965). That view has been called the situational specificity hypothesis because adherents believe that subtle (some suggest they are imperceptible) differences in job and work location can result in a test that is typically valid actually being invalid in a particular situation. On the other hand, the "new" view is that ability test validities can be generalized across all jobs and organizations (Hunter, Schmidt & Jackson, 1982; Schmidt & Hunter, 1977, 1978, 1980, 1981, 1984). That view is called validity generalization, from which the testing program obtained its title. Second, there are conflicting views over whether specific abilities are better predictors of work performance than general abilities measures. The specific abilities model holds that the better test items sample job related behaviors, the more closely the test will predict work performance. In contradiction, the general abilities model holds that job performance is best predicted by measures of general cognitive ability, because jobs are learned abilities and cognitive ability determines the learning process. The equations used to calculate VG percentile scores are composed of weighted general abilities score components.

As background to the VG testing program, a brief overview of selection testing is provided in the next section. Meta-analysis, the technology upon which Validity Generalization is based is

---

[1] "Validity generalization" is both a name given to an area of research focused on the generalizability of psychological test findings and the name of a government testing program.

introduced in the second section. The third section describes the VG testing program. Research questions related to the VG testing program are raised in the final section.

Selection Testing

Employment tests are used by employers to match an individual applicant's abilities and level of motivation with the requirements and rewards of a job. In essence, organizations attempt to predict an applicant's future job success. A central question in employment testing has been, and continues to be, the degree to which inferences based on test results contribute to the accuracy of predicting job success, i.e. is the test valid? Evidence of employment test validity can take a number of forms, but statistical analysis of the relationship between test scores and measures of job success (criterion-related validation) is the preferred approach. Criterion-related validation studies provide statistical evidence of the strength and direction of the relationship between the two variables. Thus, inferences can be drawn about whether test scores predict job success. Use of valid test scores increases the probability of "correct" selection decisions over procedures based on less valid information. Better decisions impact subsequent productivity (Dreher & Sackett, 1983).

While use of valid tests in selection has been criticized, in many cases there are no better alternatives (Ability Testing, 1982; Hunter & Hunter, 1984; Tenopyr, 1981a). Furthermore, research indicates that for entry level jobs, no alternative selection procedure matches the validity of standardized abilities tests, with the possible exception of biodata blanks (Hunter & Hunter, 1984; Reilly & Chao, 1982; Tenopyr & Oeltjen, 1982).

Employment testing in the United States began at the turn of this century. During both world wars, national interest was focused on development and use of tests to place manpower in the war effort. Following World War II, employment testing became popular in the private sector. By 1964, a National Industrial Conference Board survey indicated that 80 percent of the respondents used testing in selection and/or promotion decisions. However, a piece of legislation enacted that

same year, Title VII of the Civil Rights Act, precipitated a reversal of that trend. Section 703 (h) provided that it shall not be "an unlawful employment practice for an employer to give and to act upon the results of any professionally developed ability test," as long as there was no intent to illegally discriminate. However, subsequent court interpretation of that provision and enforcement agency guidelines have been very narrow. In the landmark decision, Griggs v. Duke Power Co. (1971), the Supreme Court indicated that where professionally developed tests had an adverse impact on protected classes of persons, those tests would have to be validated. Essentially, the court had based its decision on the "old view." In response to various judicial interpretations of Civil Rights legislation, employers reassessed use of employment testing. Due to stringent technical and burdensome administrative requirements and their associated costs (Employee Selection, 1983), testing was frequently dropped from selection processes (Ability Testing, 1982; Baker & Terpstra, 1982; Petersen, 1974; Tenopyr, 1981b).[2] If, however, selection procedures replacing testing have lower validity, then organizational and national productivity suffers.

The Employment Service claims that its VG percentile scores are valid for all jobs in all locations (even though there is no direct evidence) because meta-analytic studies have so indicated. Based on that premise, the employment service contends that further validation is unnecessary and validity can be generalized to new situations. That assertion contradicts national policy (Madigan et al., 1986) since the Uniform Guidelines (1978) recommend local validation studies to establish test validity for each job in each situation.

To summarize, local validation studies have demonstrated that ability tests have typically been valid for selection decisions. However, local validation studies have indicated that the same type tests for similar jobs, in similar situations were found to be invalid. The differences in interpretation for the extreme variance in validity coefficients have not been resolved. Yet, the VG testing program is based on the controversial "new" interpretation. That interpretation is based on evidence from a stream of meta-analytic research which is the topic of discussion in the next section.

---

[2] Perhaps the significant reduction in number of published validation studies (Monahan & Muchnisky, 1983) is also related to this more recent trend (Boehm, 1982).

Meta-analysis

The seedbed from which the VG testing program emerged is meta-analytic research. In developing the new testing program, one of the foundational Employment Service studies was a meta-analysis of General Aptitude Test Battery (GATB) validation data. Development of the meta-analytic technology was also the catalyst for reviving the situational specificity controversy. Consequently, this important technique is introduced in this section, to place the results of meta-analytic studies discussed in Chapter Two in context.

Meta-analysis is a technology for cumulating the finding of numerous studies. Particularly because the evidence from local validation studies was frequently contradictory, the question of how to determine the meaning of vast amounts of evidence arose. Until meta-analysis, summary studies often counted the number of studies that found statistically significant validity coefficients and those that did not. The finding with the highest total won. It became obvious that there were qualitative differences between studies. All results were not equal. Evidence should be weighed according to study characteristics.

Schmidt and Hunter (1981) developed a technique for dealing with qualitative deficiencies in validation studies so that the population validity mean and standard deviation could be estimated. In this technique corrections are made for statistical artifacts because such artifacts reduce the quality of observed results. Seven sources of statistical artifacts were identified.

- Sampling error - a result of the non-representativeness of a relatively small sample size;
- Differences between studies in criterion reliability.
- Differences between studies in test reliability.
- Differences between studies in range restriction.
- Differences between studies in amount and kind of criterion contamination and deficiency.
- Computational, typographical, and transcription errors.
- Slight differences in factor structure between tests.

In VG studies, corrections are made for only the first four sources of artifacts since currently it is virtually impossible to correct for the the last three sources (Hunter, Schmidt, & Jackson, 1982).

The Schmidt-Hunter meta-analytic technique involves three basic steps.

- Based on a Baysian prior distribution, researchers first estimate the amount of variance in validity coefficients due to sources of statistical artifacts.
- The estimated error variance is subtracted from the observed variance, providing an estimate of residual variance.
- Finally, based on the magnitude of residual variance, they determine whether or not situational moderators exist, i.e. "whether or not the situational specificity hypothesis is supported." Since there are three sources of artifacts for which no correction is made, the Schmidt-Hunter decision rule is: if the correction of four sources of statistical artifacts accounts for 75% of the observed variance, reject the situational specificity hypothesis. In such a case, observed variance would be deemed illusory.

Findings of meta-analytic research provided impetus for the development of the VG testing program which is introduced in the next section.

The US Employment Service VG Testing Program

The Employment Service has recently introduced a free applicant testing service to employers. While the test instrument, the General Aptitude Test Battery (GATB) is not new, the method in which its scores are used is entirely new.

The GATB consists of twelve timed tests designed to measure nine specific aptitudes. These nine specific abilities have historically been used in various combinations called Specific Aptitude Test Batteries (SATBs) to predict performance for a narrow group of jobs. Based on a recent study of the GATB's dimensionality, those nine scores are combined into three general ability scores for the VG testing program. Another study indicated that the general abilities were better predictors of job performance than the specific abilities. Those three general abilities are: cognitive, perceptual, and psychomotor (USES test report number 44, 1983).

A third study investigated whether jobs moderate validity (USES Test Research Report No. 45 1983). Whereas the traditional view held that job tasks moderate validity, cumulation of 515 GATB validation studies indicated that job/task differences could be accommodated by grouping all jobs into just five broad job families. And, since jobs in the 515 validation studies were judged to be a representative sample of the 12,016 jobs in the Dictionary of Occupational Titles (DOT) the findings were generalized to all jobs. Thus, five job family scores are computed for each VG testing program applicant and only the one which is appropriate for a job order is reported to a prospective employer.

Results of a fourth study led to the adoption of percentile scores as the predictor. The issue investigated was fairness. While estimates of performance based on the GATB were accurate for both minorities and non-minorities, the difference in average scores between non-minorities and blacks was substantial for all three general abilities. Those differences would result in limited employment opportunities for blacks if the GATB general ability scores were used in selection decisions (USES test research report no. 46, 1983). While job relatedness would be an adequate legal defense for such use of the GATB, the social objective of Civil Rights legislation - to provide proportionate employment opportunities for minorities - would be undermined. Consequently, a compromise was struck in formulation of the VG testing program Job Family scores.

From its historic database of GATB validation studies, the USES developed separate VG score distributions for minorities and non-minorities. Percentiles were then identified for each of the job families within those distributions. When applicant job family raw scores are calculated, they are compared to the appropriate historic distribution to obtain a percentile score. This procedure adjusts the scores of minorities to eliminate adverse impact. If employers "hire from the top down," as the USES urges, valid decisions should result that are racially proportionate (USES test report no. 43, 1983).

Summary

To summarize, the VG testing program is intended to significantly improve the allocation of human resources in our nation through offering a valid testing service at government expense. Improved allocation of human resources should result in increased organizational and national productivity. To reach that objective, the Employment Service is urging employer participation. However, there is no direct evidence of this new predictor's validity. Therefore, the first research question addressed here is whether VG percentile scores are valid predictors of work performance.

Moreover, from a scientific perspective, the question of the generalizability of employment tests across such large groupings of jobs and across all work locations remains to be resolved. Therefore, the second research question is whether the validity of VG scores is moderated by job or work location? In other words, this question addresses the issue of transportability. Specifically, the effects of differences in job and location on validity coefficient magnitude are examined.

Finally, the issue of whether general or specific ability tests are better predictors of work performance is investigated. Recall that the VG program replaces Specific Aptitude Test Batteries (SATBs) with general ability scores. Therefore the third research question is whether general ability scores are better predictors of job performance than specific ability scores.

To address these research questions, test scores from two test batteries were collected on 219 bank employees on three different clerical jobs in three locations of one organization. Supervisor ratings and objective measures of performance were also collected. These data provide the basis for addressing the three basic questions noted above. The research literature relevant to these research questions is reviewed in chapter two. Chapter three details the methodology used in this study, Chapter four reports the findings, and the implications of these results are discussed in Chapter Five.

Chapter 2: Literature Review

In the previous chapter, the conflict between the traditional and recent interpretations of validation study findings was introduced. The fact that the VG testing program is based on the new interpretation of those findings, gave rise to three research questions which were specified at the end of Chapter One. Literature reviewed in this chapter was selected on the basis of its relevance to those research questions and is presented in corresponding sections. The chapter concludes with those research questions restated as research hypotheses.

The first research question is: Are VG percentile scores a valid predictor of job performance? Since the VG testing program is based on GATB scores, and direct tests of VG percentile scores have not been published, evidence of the validity of basic ability tests is reviewed as they relate to clerical occupations. Both traditional validation studies and meta-analytic studies are reviewed.

The second research question is: Do differences between jobs within a job family and/or differences in work location (situation) moderate the abilities-performance relationship? The central issue with jobs as a moderator is how broadly they should be classified for prediction of performance. For the USES, the central issue with the possibility of work location moderators is whether general ability test validities are near-zero in some situations. Perhaps the most controversial aspect of the validity generalization literature is its conclusion that there are no work location moderators of general ability test validity. Part of that controversy may be related to ambiguous terminology.

Since the term 'moderator' has been used imprecisely, its meaning is clarified prior to reviewing literature related to job and work location moderators. Evidence related to moderator effects is presented, along with criticisms of the VG technology.

The third research question addresses a classic issue in selection research: Are general abilities better predictors of performance than specific abilities? The literature on this issue, both theoretical and empirical, is voluminous. The VG testing program is based on the view that general abilities are the better predictor. On the other hand, in its extreme form the traditional view is that the more similar test questions are to job behaviors, the better the prediction of performance. Discussion here will be limited to a summary of the Employment Service study and findings (USES Test Report No 44, 1984).

The Validity of Ability Tests for Clerical Jobs

The literature pertaining to test validation is vast, encompassing the full spectrum of jobs in relationship to a wide range of predictors and criteria. This section focuses specifically on studies pertaining to clerical occupations. With regard to predictors, the focus was narrowed to tests of basic abilities, particularly cognitive ability and psychomotor ability tests (since Job Family 4 percentile scores are calculated from a combination of such measures) and clerical ability. With regard to criteria, the focus was narrowed to proficiency criteria[3] since that was the issue of interest to the host organization and most summaries of the literature have classified validation studies on that basis.

Ghiselli (1966, 1973) reviewed the findings of a large number of local validation studies by sorting them into job families and reporting average validities based on test and criterion type. In his first review (1966) average observed validities for ability tests and proficiency criteria across all combinations of tests and jobs ranged from -.40 to .80. Validity coefficients for cognitive ability

[3] Proficiency criteria are measures of work performance which do not include training success.

with proficiency criteria scores averaged .23. Average validities were reported for two subsets of clerical jobs that are related to the jobs included in the current study: 'bookkeepers and cashiers' (cognitive ability, .22, psychomotor ability, .21); 'bookkeeping machine operators' (cognitive ability, .21, psychomotor ability, -.01).

In a larger review, Ghiselli (1973) weighted validity coefficients according to sample size in calculating the mean. [4] For clerical occupations in which cognitive ability tests and proficiency scores were used, mean validity was somewhat higher (.28) than reported in the first study. For psychomotor abilities tests validity was .16.

Another approach to summarizing validation study results is the meta-analytic (VG) technique described in Chapter One. An early validity generalization study re-examined data from a 1977 study of clerical occupations (Schmidt, Hunter, Pearlman, and Shane, 1979). Validity coefficients of 3,300 test-criterion combinations were cumulated. Tests were grouped into five categories using an unexplained system "adapted from" Ghiselli and Dunnette. Each of the categories represented a test type or an ability factor found in the psychological literature. The five test types were:

- Verbal Ability,
- Quantitative Ability,
- Cognitive Ability,
- Perceptual Speed, and
- Psychomotor Ability.

The clerical jobs were all classified in one of two DOT clerical occupational divisions:

- Computing and Account-Recording (Occupational Division 21), and
- Stenography, Typing, Filing, and Related Occupations (Occupational Division 20).

Since the jobs examined in the current study, as reported in Chapter Three, are in Occupational Division 21, further discussion of the 1979 study is confined to that classification.

---

[4] Although no mention was made of the number or sample size of individual studies, the total number of subjects was over 10,000.

Within the "computing and account-recording" category, 58 studies (ave. N = 92) investigated the relationship of cognitive ability scores and job proficiency scores. The 90% credibility value was .22. (Ninety percent of observed validity coefficients should be .22 or above.) After correction for statistical artifacts, the estimated true validity was .49. For the sample of 131 validity coefficients (ave. N = 91) of psychomotor ability tests with proficiency scores, the credibility value was .12 and the corrected validity coefficient was .29.

In another study, Pearlman, Schmidt and Hunter (1980) examined 3,368 validity coefficients of clerical jobs and tests, obtained from 698 independent samples, where different predictors and criteria had been studied. Two thirds of the samples came from unpublished studies. Jobs were coded according to DOT classifications:

- Stenography, Typing, Filing occupations (DOT groups 201-209)
- Computing and Account-Recording occupations (DOT groups 210-219)
- Production and Stock Clerk occupations (DOT groups 221-229)
- Information and Message Distribution occupations (DOT groups 230-239)
- Public Contact and Clerical Service occupations (DOT groups 240-248)

Ten test types and two classes of criteria were represented. Test types were:

- Cognitive Ability
- Verbal Ability
- Quantitative Ability
- Reasoning Ability
- Perceptual Speed
- Memory
- Spatial/Mechanical Ability
- Psychomotor Ability
- Performance Tests, and
- Clerical Aptitude

The two classes of criteria were: job proficiency and training success. For a set of 47 studies in the "computing and account-recording" category, (ave. N = 94), mean observed validity for cognitive ability tests with proficiency criteria was .23. After correcting for statistical artifacts, the estimated

true validity was .49. For a similar set of 97 studies (ave. N = 86) where psychomotor ability tests were used to predict proficiency, mean observed validity was .14. After correcting for statistical artifacts, the estimated true validity was .30.

To summarize, the findings of Ghiselli and the meta-analytic studies concerning mean observed validity are similar. Table 1 presents a summary of the studies discussed above. In the first column are four groupings of clerical jobs. The type of ability test used as predictor is indicated in column two. Whether the approach was traditional or meta-analytic is indicated in column three. Column four reveals the year of the study with results reported in columns five through seven. Where the job classifications were comparable, there was at most a .04 difference between estimates of mean observed validity, with the meta-analytic estimate on the lower end. Validity of cognitive ability tests was between .22 - .26 for computing and account-recording clerks. Validity of Psychomotor ability tests were between .12 - .16. One major difference, however, between the traditional summary and the meta-analytic is the latter's corrected coefficient. After correcting for sampling error, unreliability in the predictor and criterion measures, and restriction in range for the predictor and criterion measures, (cf. discussion in Chapter Three), estimated true validity for cognitive ability tests for computing and account-recording clerks was .49. Estimated true validity for psychomotor tests was .29 - .30. (A second major difference regarding estimates of the variance of validity is discussed below). Accumulated evidence suggests that cognitive ability tests are typically valid predictors of proficiency criteria, with observed validity for computing and account-recording clerks in the .20-.30 range, and corrected validity in the upper .40s range. The VG testing program Job Family 4 percentile scores are a composite of cognitive ability and psychomotor ability which is intended to improve prediction. Therefore, the estimate of a .53 corrected validity and "worst case" of .33 for the Job Family 4 composite seems plausible (USES test research report No. 45, 1983, p. 43).

Table 1. Abilities-Proficiency Validity for Clerical Jobs

Clerical Jobs	Test Type	Study Type	Yr	Average Validity	SD	Corrected Validity
All	C	T	1966	.23	n/a	n/a
	C	T	1973	.28	n/a	n/a
	P	T	1966	.21	n/a	n/a
	P	T	1973	.16	n/a	n/a
Bookkeeper	C	T	1966	.22	n/a	n/a
	M	T	1966	.21	n/a	n/a
Machine Operator	C	T	1966	.21	n/a	n/a
	P	T	1966	-.01	n/a	n/a
Computing etc.	C	T	1973	.26	n/a	n/a
	C	MA	1979	.22*	.17	.49
	C	MA	1980	.23	.19	.49
	P	T	1973	.16	n/a	n/a
	P	MA	1979	.12*	.13	.29
	P	MA	1980	.14	.13	.30

C = Cognitive ability; P = Psychomotor; T = traditional; MA = meta-analytic
*Credibility Value

Moderators of Validity

In this section, the topic of moderators in the context of general ability employment tests is introduced. The relevance of the topic to this study is then discussed, and empirical evidence related to two potential moderators is presented.

The term moderator has been used in a selection context to refer to an individual or environmental characteristic that is differentially related to the ability-performance relationship. Academics who were disappointed in the magnitude of observed validity coefficients hoped that moderators would be useful in explaining more performance variance. As early as 1956, Ghiselli called for the investigation of moderator variables in order to improve prediction of work performance. Other industrial-organizational psychologists joined in the call (Guion, 1976; Schneider, 1978).[5] One approach to identify moderators was to integrate the selection and organizational behavior literatures (Porter, 1966). Theoretical justification for such integration was derived from the performance model.

Performance is generally modeled as a function of ability combining multiplicatively with motivation (Vroom, 1964; Lawler, 1971). Since abilities are relatively stable and motivational states are not, selection research had focused on the more stable ability-performance relationship (Dunnette, 1973). Organization behaviorists, on the other hand, had focused on predicting behavior based on organizational properties and processes like management or leadership style (Fleishman & Harris, 1962; Tannenbaum, 1968; Fiedler, 1967; House, 1971), reward systems (Lawler, 1966), or task characteristics (Theologus & Fleishman, 1976). Such situational variables were used as a surrogate for internal motivational states, i.e. as activators or suppressors of motivation (Schneider, 1978). Since a validity coefficient reflects the relationship of only one of the model's two determinants of performance, it follows that motivational differences across situations

---

[5] Stone & Hollenbeck (1984) suggested that such efforts were misdirected. The validity coefficient "cannot be directly interpreted as an index of the degree to which scores on one variable can be accurately predicted on the basis of scores on another variable." Accuracy of prediction, indexed by the standard error of the estimate, is a function not only of the value of the correlation coefficient, but also of the variability of performance scores.

should moderate the ability-performance relationship. However, while theory suggested the existence of moderators, very little direct research has been conducted to verify them (Schneider, 1978). Furthermore, only a small portion of studies on moderators deal directly with the ability-performance relationship, and of those, only a fraction are relevant for this study since the term 'moderator' has been used ambiguously.

The concept of moderation is complex. At the conclusion of his discussion of moderators, Guion (1976) stated: "We have become confused by the topic." More recently, however, certain issues have been clarified (Arnold,1982; Stone & Hollenbeck, 1984; Arnold, 1984). Arnold's (1982) distinction between moderation of the "degree" and "form" of relationship is useful for identifying research findings that are relevant to this study. To explain his distinction, first, the indices of "degree" and "form" are identified. Then, technical definitions of those terms are given, followed by a definition of 'moderator.'

The correlation coefficient between two variables, a dependent variable Y and an independent variable X, is the index of the "degree" of relationship whereas the regression coefficient of Y on X is an index of the slope or "form" of relationship. In other words, the "degree" of relationship is specified by the validity coefficient and the "form" refers to the slope of the relationship. If the degree and/or form of two variables are constant across values of some third variable Z, the X-Y relationship is constant with regard to Z. However, if the form or degree vary with values of Z, then Z is called a 'moderator.' The variable Z stands for a characteristic that in validation research has been used to subgroup subjects in the population of interest. Early subgrouping typically was done on the basis of race and sex. More recent subgrouping has been related to situational variables.

Since variables associated with either differences in correlation coefficients or differences in slopes across subgroups have been identified as moderators in the literature, the meaning of the term is ambiguous. Actually, different information is obtained from the two analyses. The question addressed by assessing the "degree" of the relationship across subgroups is whether X accounts for as much variance in subgroup E as in subgroup F. On the other hand, the question addressed by the "form" of the relationship is whether X and Z interact in determining Y. Where moderation of degree obtains, moderation of form does not necessarily obtain, and vice versa. Therefore, on

the basis of an analysis of degree of relationship, the conclusion could be that there is no moderator while on the basis of the analysis of form, the conclusion could be that a moderator exists.

In selection research, both correlation coefficients and slopes have been compared, albeit not for the purpose of identifying moderators in order to improve prediction. For example, regarding the issue of test bias, the 'degree' of relationship underlies the concept of differential validity whereas the 'form' of relationship is the focus of differential prediction (Cleary, 1968; Schmidt & Hunter, 1974). Validity Generalization studies, on the other hand, have been solely concerned with moderators of the degree of performance. As discussed in Chapter One, the traditional explanation for findings of extreme variance in observed validity coefficients is that there are situational moderators. Proponents of that view hold that the true validity for an ability test may be zero. On the other hand, proponents of the new interpretation argue that the extreme variance is artifactual and hold that all professionally developed cognitive ability tests are valid for all jobs, i.e. that there are no near-zero true validity coefficients for the ability-performance relationship. Thus, the relevant issue for this study is, are there any near-zero true validity coefficients, an issue related to the indices of moderation or the degree of relationship.

Commenting on the possibility of motivational moderators of performance, Schmidt et al. (1985) argued that situational factors that affect employee motivation do not affect validity. While mean performance may be suppressed, there is still variance around the mean. Consequently, since the correlation coefficient is a standardized measure, the magnitude of the coefficient would be unaffected. To translate that perspective into Arnold's terminology, the form of the relationship may be moderated where the degree is not. Another theoretical possibility exists: that both form and degree may be moderated, i.e. there may be an interaction effect and mean differences across subgroups associated with different levels of motivation. Thus, while Schmidt et al's scenario is possible, it neither exhausts the possible outcomes nor rules out the possibility of moderation of the form of the relationship. Whether there are moderators of the ability-performance relationship is an empirical question. Research evidence on two potential moderators is discussed below: job and work location.

Job Moderators

Traditionally it has been thought that jobs moderate the ability-performance relationship, i.e. that the validity of abilities tests are different across jobs (McCormick, 1978). For example, numerical ability was thought to be more valid for an accountant than for a cashier or more valid for a cashier than an auto mechanic.

A field study conducted by Colbert & Taylor (1979) directly addressed the issue of job effects on the ability-performance relationship. Their results suggested that job families within the clerical occupation moderate the ability-performance relationship. Using Position Analysis Questionnaire (PAQ) data, they derived thirteen potential clerical job families within a large insurance company. Of interest here are the three job families which contained a substantial number of entry-level jobs. The first job family (JF III; N = 219) was composed of upper level typing and secretarial jobs. The second contained lower level clerical jobs (JF IV; N = 504). Lower level keyboard operator jobs, and middle level clerical jobs were found in the third job family, (JF V; N = 67).

Three commercially developed basic aptitude tests (verbal, numerical, and visual speed and accuracy) plus a typing test were administered to 5,399 applicants. A total of 2,204 people were hired into the three job families in 18 months. (Only typing test results were considered in the hiring decision.) Approximately 30% (661) were hired into the first job family, 60% (1322) into the second, and 10% (220) into the third. The criterion was a composite of three-month performance dimension ratings which "differed somewhat" between job families.

To test the usefulness of these job families for predicting work performance, regression equations were developed and cross-validated. It was hypothesized first that different predictors would be valid for different job families. Regression equations were derived based on approximately two-thirds of the cases in each job family and were cross-validated with the remaining cases. Verbal and numerical tests were found predictive of performance within Job Family III. Numerical and visual speed and accuracy tests were predictive of Job Family IV. None of the tests were found to be valid at the .05 significance level for Job Family V. It was next hypothesized that prediction equations developed on one job within a job family would be found to cross validate at a statis-

tically significant level when applied to other jobs within the same job family. To test this hypothesis, only JF III and IV were used. One job in JF III which accounted for 64% of the sample was used to derive a regression equation as was a job in JF IV that accounted for 48% of its sample. Those equations were cross-validated with the respective job family sample residuals. The predictors found valid in the previous analyses were confirmed for the respective job families.

Finally, it was hypothesized that prediction equations would be found to yield greater amounts of prediction error when applied to a cross-validation sample drawn from a different job family than would be yielded if the cross-validation sample were drawn from the same job family. To test the last hypothesis the absolute value of prediction errors resulting from each validation operation were computed and subjected to a balanced 2 x 2 fixed effects ANOVA. Factor A was the job family from which the regression equations were derived and factor B was the job family on which the equations were cross validated. A significant interaction between factors A and B indicated that regression equations derived from different job families resulted in differing degrees of prediction error. Prediction error was greater for equations applied across job families than for equations applied within job families. This finding suggests that dividing the clerical occupation into several job families enhances prediction.

Different conclusions were reported in a meta-analytic study conducted by Schmidt, Hunter, & Pearlman (1981). They used virtually the same data base as the 1980 clerical study discussed earlier. (A few minor job classifications were added, totalling nearly 3,400 validity coefficients for clerical occupations.) Clerical job classifications were the same as in the 1980 study. Results indicated that within-family standard deviation (SD) of corrected validity coefficients is of the same magnitude as the SD pooled across the total group. On that basis, the researchers concluded that there were no situational moderators, and that neither task nor behavioral differences acted as moderators. Granting the accuracy of those findings, one broad clerical family would seem appropriate.

The issue of job moderators was also investigated by the Employment Service (USES Test Research Report No. 43, 1983 and USES Test Research Report No. 45, 1983). The USES sought

to discover whether validities vary by job. They considered five methods of job analysis to determine whether job differences affect validity coefficients. Methods used were:

- DOT estimated mean ability ratings
- Functional Job Analysis (FJA)
- Occupational Aptitude Patterns (OAPs)
- Position Analysis Questionnaire (PAQ, 1972)
- Position Analysis Questionnaire (PAQ, 1977)

Those job analysis methods were compared for the effect their classification schemes had on the validity of abilities tests.[6] It was determined that one job dimension, complexity was associated with validity in all five methods. As complexity decreased, the contribution of psychomotor ability to validity increased. Since the complexity dimension was common to all methods, the USES's Functional Job Analysis (Fine, 1974) was selected because of the comprehensive analyses available in the Dictionary of Occupational Titles (1977). In FJA, jobs are rated on three factors: "data" (information, facts, ideas, and statistics); "people" (clients or co-workers); and "things" (machines or equipment) respectively. A modified "data" and "things" categorization scheme was used to establish five complexity levels across all jobs. Those levels were used to classify jobs into families as indicated in Table 2.

In summary, empirical evidence indicated that jobs moderate the ability-performance relationship. Two large studies showed improved prediction of job success through subgrouping jobs. Colbert and Taylor's field study suggested that grouping jobs into families within the clerical occupational group improved prediction. However, the USES study indicated that a much broader grouping, encompassing all clerical occupations, was appropriate. Thus, a question remains as to the optimal grouping of jobs for prediction purposes.

---

[6] Pearlman (1980) argued that for the purpose of employee selection, grouping jobs on the basis of enhanced prediction of job performance is more appropriate than first establishing job families on some other basis, like job tasks or behavioral requirements.

Table 2. Validity Generalization Job Families

JOB FAMILY	NAME	DOT CODE	COMPLEXITY LEVEL
1	Setting up	Things = 0	1
2	Feeding, offbearing	Things = 6	5
3	Synthesizing	Data = 0,1	2
4	Analyzing, compiling	Data = 2,3,4	3
5	Copying, comparing	Data = 5,6	4

Work Location Moderators

In his summary of direct research on work location moderators, Schneider (1978) found research on only two categories of potential work location moderators: organizational climate and incentive systems. Both were found to moderate the ability-performance relationship when experimentally manipulated in laboratory studies. In one field study further support was found for work climate as a moderator. Forehand (1968) divided 120 government executives into Group-centered or Rules-centered work climates. Eight cognitive ability tests were administered and peer assessment ratings of innovative behavior were obtained for the criterion. In the Group-centered climate, all but one test was found valid at the .05 level of significance, ranging from .26 to .45, whereas in the Rules-centered climate none were statistically significant.

The existence of work location moderators has been a major focus of meta-analytic studies. However, in contrast to the direct tests for moderation reviewed by Schneider, Schmidt-Hunter utilized two approaches. Their primary approach was meta-analytic, an indirect or residuals approach. Three meta-analytic studies addressed the issue of portability of test validity across locations for clerical jobs. To lend supporting evidence, two within-setting studies were also conducted.

With cognitive ability tests and proficiency criteria, Schmidt, Hunter, Pearlman, and Shane (1979) found that corrections for statistical artifacts accounted for only 57% of observed validity variance for 'computing and account-recording' jobs. That is far below the 75% decision rule.

Thus, according to their a priori rule, there was insufficient evidence to reject the situational specificity hypothesis for this category of jobs. Nevertheless, in several other clerical classifications, corrections did account for 75% or more of the variance, thus, Schmidt et al. asserted that the validity for all clerical jobs is generalizable. They contended that differences between studies in the amount and kind of criterion contamination and deficiency probably accounted for the additional 18% unexplained variance from 'computing and account-recording' studies.

In another study of clerical jobs, Pearlman, Schmidt and Hunter (1980) found that 51% of the variance in validities for 'computing and account-recording' jobs was accounted for by correcting for artifacts. Of 32 predictor-criterion relationships reported, only 16 met the 75% rule. Nevertheless, the researchers optimistically inferred from the half that did satisfy their decision rule that validity was generalizable for all clerical jobs. In reporting the results of their study, the researchers maintained that their findings strongly confirmed the validity generalization theory. Nonetheless, the tenuous nature of that conclusion cannot be denied.

Results from a third meta-analytic study, however, tend to support the notion of validity generalization even though the findings do not conform to the Schmidt-Hunter decision rule. Schmitt et al. (1984) used the Schmidt-Hunter technique to examine the effect of research design, criterion used, type of selection instrument used, occupational group studied, and predictor-criterion combination on the level of observed validity coefficients. Validation studies were taken from Journal of Applied Psychology and Personnel Psychology between the years 1964 and 1982. Since earlier VG studies reported that the predominant source of statistical artifacts was sampling error, corrections in Schmitt et al.'s study were made only for sampling error. To cumulate a sufficient number of validity coefficients (53) from studies that examined the relationship of general cognitive ability to proficiency criteria, seven occupations classifications were grouped together. Average sample size was 144. Mean observed validity was .22. Corrections for sampling error accounted for only 12% of the variance in validity coefficients. Thus, the Schmidt-Hunter decision rule would determine that the null hypothesis (that validity variance is zero) must be rejected, moderators exist. However, observed validity variance for these studies was small. There are several plausible reasons. Relatively large sample sizes would reduce sampling error, and the psychometric qualities of meas-

ures used may have been more reliable than typical validation studies which would reduce that statistical artifact and could enhance effect size. Counter to the Schmidt-Hunter decision rule's conclusion, two findings tend to support the generalizability of validity: the relatively high mean observed validity and the small variance about that mean. Consistent with the VG testing program tenets, one could argue that the small variance observed between validity coefficients from these large sample studies is due to job family difference. Thus, this evidence appears to support the transportability of ability test validity.

Technical criticisms of VG technology arising from theory and simulation studies suggest that meta-analytic conclusions concerning the absence of situational moderators are tenuous. Algera et al. (1984) challenged the power of the Schmidt-Hunter test to detect moderators. The power of a statistical test refers to its ability to reject the null hypothesis if the null hypothesis is in fact false. The Schmidt-Hunter null hypothesis is that true test validity variance is zero. When a test fails to reject the null, Schmidt-Hunter have concluded that the null is affirmed. Their univariate test, however, has lower power than a multivariate test. Consequently, the null hypothesis is favored since the univariate test may have had insufficient power to detect an existing relationship. In response, Algera et al. (1984) suggested use of a more powerful multivariate test, the test for the homogeneity of correlations (Hays, 1973; Kraemer, 1979). Evidence from simulation studies discussed below lends support to this criticism of power deficiency.

Three separate Monte Carlo simulations found that the 75% decision rule failed to detect low-to-moderate true validity moderators. In the first of these studies, Osburn et al. (1983), generated attenuated bivariate distributions using a FORTRAN program with a random number generator and a regression procedure to introduce the desired correlations. One hundred such distributions of N = 5,000 were generated. To estimate the power of the Schmidt-Hunter 75% rule, 500 replications in which the true validity variance was actually greater than zero were generated for each of a number of specified conditions. Power calculations were made on three levels of sample size: 50, 100, and 200; three levels of number of studies: 25, 50, and 100; two levels of population true validity: .25 and .50; and four levels of true validity variance: .005, .012, .022, and .034. Those levels were designated as "small," "moderate," "large," and "extreme" respectively. All possible

combinations were run for number of studies equal to 50, but only selected combinations were run for number of studies equal to 25 and 100.

Results indicated that although the 75% rule protected well against Type I errors (rejecting the null hypothesis when it is true), it was inconsistent in protecting against Type II error (failing to reject the null hypothesis when it is false). With anything less than samples of $N = 200+$, no matter how many studies were included in the analysis, the 75% rule failed to detect small-to-moderate true validity variance. Power actually decreased as the number of studies increased if their sample size was less than 200 and/or the variance of the true validities was low to moderate.

Ladd & Cornwell (1986) also addressed the accuracy of validity variance estimates based on the Schmidt-Hunter technique. To investigate the accuracy of VG findings, these researchers conducted a massive simulation which included measurement unreliability, sampling error, and restriction in range. First, they generated normally distributed predictor and criterion variables with a mean reliability of .80. Population validity coefficients were then calculated at eleven levels, rho = 0.0, .05, .10, .15, .20, .25, .30, .35, .40, .45, .50. From those populations, samples of individual cases were taken, with sample sizes from 20-300 (mean = 60). Restriction in range was also simulated. Three thousand validity coefficients were generated for each value of rho, with sample size, criterion reliability, and restriction in range varied independently. They also manipulated the number of studies included in meta-analyses: K = 6, 12, 18, 24, 30, 48, 72, 96, and 120. A total of 3,069 meta-analyses were computed, each having a unique combination of number of studies and correlation coefficients. Ten replications performed on each of those meta-analyses resulted in 30,690 individual meta-analyses.

Results documented the reliability of both the estimated true mean and variance. While estimates of the true mean were "acceptably reliable," the estimates of true variance were not. "Only in near-ideal situations of large N and K, are they reliable enough to warrant attention." These results supported Osburn et al's (1983) finding of power deficiency in the moderator detecting ability of the Schmidt-Hunter technique. Due to that deficiency, the authors suggested that "researchers may be wise to continue to search for moderators even though a meta-analysis of their literature indicates that none exists."

A third Monte Carlo study (Sackett et al., 1986) assessed the relation of three factors to the power of meta-analytic studies:

- the number of studies included in the analysis (4, 8, 16, 32, 64, 128);
- the average sample size of those studies (50, 100, 200); and
- the size of the true differences in population correlations (.10, .20, .30).

The simulation assumed that half the studies in the meta-analysis were in one population and the other half in a second population. For each combination of factors, researchers generated 1,000 sets of sample correlations, and performed 1,000 meta-analyses according to the Schmidt-Hunter procedure. These analyses were performed under five conditions. First, the data contained no measurement error - providing upper limits for the power of meta-analysis conducted on given sample size and number of studies. Second, unreliability of measurement of both predictor and criterion varied from study to study, without correction for attenuation. Third, both variables contained measurement error, but reliability was constant across studies and no correction was made for attenuation. Fourth, observed correlations were corrected for attenuation prior to meta-analysis. And fifth, true reliabilities were used for only a subset of studies and the rest were corrected according to Hunter et al.'s (1982) artifact distribution methods.

This study led to a particularly interesting finding. Correcting for measurement error, either by correcting each correlation coefficient or by artifact distribution methods, did not affect power. While those corrections did improve the accuracy of population validity estimates, they were irrelevant to the issue of power. The study showed, however, that power is reduced considerably as reliability drops from 1.0 to .8 to .6. For example, where the true population difference in validity was .20, the power of Schmidt- Hunter analyses entailing 32 studies with sample sizes of 100 dropped from .945 to .756 when reliability of measures dropped from .80 to .60.

The ability to correct for known measurement error appears to be no substitute for more accurate (reliable) measurement. Based on that finding, Sackett et al. conclude that meta-analysis has its limits: "it cannot consistently detect the presence of moderator variables under some of the conditions investigated." Therefore, "statements attributing observed variation across studies to

statistical artifact should be made with more caution than is evident in much of the current meta-analysis literature."

A final criticism has to do with the boldness of inferences drawn in VG studies. According to James et al. (1986), VG research has fallen prey to the logical fallacy of "affirming the consequent." To infer that a causal theory actually and uniquely explains the data when a good fit exists between predictions from a causal theory and empirical data is a logical fallacy. Other causal theories may explain the same data equally well. When the proportion of observed variance accounted for by statistical artifacts is 75% or greater, the most that should be inferred is that VG ("cross-situational consistency") furnishes a useful explanation for the observed validity variance. However, VG studies typically infer that "the situation specificity hypothesis is rejected."

Based on a contrived set of validity coefficients, James et al. (1986) presented two plausible alternative explanations of their data. (The variance in validity was designed to illustrate a condition in which multiple conclusions could be drawn regarding causes of the validity variance.) Each explanation began with a set of untested assumptions. The key "what if" assumption of the validity generalization procedure was: What if the population correlation is assumed to be the same over studies. Using the Hunter et al. (1982) equations, mean observed validity was .50 and estimated variance due to sampling error was 75% of the observed variance. Thus, the VG procedure found that true variance is zero. The key "what if" assumption of the situational specificity procedure was that a unique population validity underlies each situation (study). Using the Gulliksen (1950) equation to estimate "reasonable limits" of the error of measurement for a population for each of the validity coefficients, possible values of the 30 true validities were calculated. Those values ranged from a low of .04 to a high of .84. Thus, the situational specificity procedure found that the validity coefficients came from different populations.

There is nothing unusual in having two viable explanations for data in causal, or confirmatory, analyses. "When confronted with conflicting models, the objective is to ascertain if one or more of the models might be disconfirmed by additional tests" (James et al., 1986, p. 443). Meanwhile, the situational specificity hypothesis remains "alive and well."

In summary, meta-analytic findings regarding the absence of moderators appear to be premature. First, meta-analytic findings have been mixed. Second, the appropriateness of the Schmidt-Hunter decision rule is questionable. In Schmitt et al.'s meta-analysis of quality validation studies (1984) evidence seemed to support validity generalization, whereas the decision rule rejected it. Third, criticism of the technique's power to detect moderators except with quality studies has been raised in theory and supported by simulation studies. Thus, caution has been urged in accepting conclusions that "moderators do not exist." Fourth, according to James et al. (1986), VG findings that support the null hypothesis that "true validity variance is zero" should be treated as one plausible explanation of current empirical evidence. On the other hand, procedures assuming and confirming situational specificity remain plausible. Consequently, the situational specificity vs. validity generalization issue remains unresolved.

Taking a different approach to the situational specificity issue, two within group studies were conducted (Schmidt & Hunter, 1984; Schmidt Ocasio, Hillery, & Hunter). It was posited that if the setting, job, organization, criterion, and applicant pool do not vary, observed test validity will not vary.

In the first study, Schmidt and Hunter examined results of a series of validity studies which had been conducted on one organization's stenographer applicants in each of four consecutive years. Sample sizes ranged from 39-49 applicants. Observed validity ranged from .08 to .19, with two Minnesota Clerical tests yielding both statistically significant and non-significant results. That appears to contradict the prediction of situational specificity. While the situation was held constant, variance in observed validities between studies was comparable to the variance between studies conducted in different settings. However, an alternate explanation for these results is that significant changes in the organization or in employee motivation over time was reflected in the observed validity coefficients.

In the second study, however, temporal effects were controlled. Numerous small sample studies were generated from a single large-sample study ($N = 1,455$). Although the organization, job, test, criterion measures, applicant pool, time-period and sample size were held constant ($N = 30$ & $N = 60$), substantial variability was found across studies in observed validity coefficients, signif-

icance levels and traditional conclusions about the presence or absence of validity. Since subjects were randomly assigned to subsamples, situational variables should not account for the variance. On the other hand, results of meta-analysis indicated that between study variance in validity coefficients was due to sampling error.

These within-setting (single location) studies provide a form of direct evidence for the VG contention that statistical artifacts account for a significant portion of variance in validity coefficients.

In conclusion, evidence indicates that jobs do moderate validity. The question remaining to be answered is what is the optimal breadth of job families. With regard to work location, only two studies provided direct evidence of location effects. On the other hand, meta-analytic studies have provided mixed results. The findings of no location effect may be a function of the technique's lack of power. On the other hand, within-setting findings suggest that statistical artifacts do account for a substantial portion of correlation coefficient variance. Thus, extant empirical evidence does not resolve the issue of whether work location moderators exist. Rejection of the situational specificity hypothesis for clerical jobs based on current evidence, particularly using the Schmidt-Hunter decision rule, seems premature. Further evidence is needed to disconfirm the situation specificity hypothesis (that jobs and work locations moderate validity), the validity generalization hypothesis (that work location does not moderate validity), or both hypotheses. More meta-analyses of large sample studies from independent sources like Schmitt et al.'s (1984) would be informative. New large sample validation studies which could be cumulated at some point in the near future would overcome the Schmidt-Hunter procedure's power deficiency. In the meantime, direct tests of the validity of the Employment Service's testing program which is based on meta-analytic findings of small sample studies would add a new stream of evidence. New approaches appear to be necessary inasmuch as most validity generalization studies have lacked sufficient power to detect moderators, if they exist.

Specific versus General Abilities

A foundational theoretical question pertaining to the USES's testing program is whether prediction of performance is best achieved on the basis of specific or general abilities. The two basic perspectives can be summarized as follows:

1. Specific ability theory - job performance is best predicted by tests that are most similar to the job in terms of behaviors sampled (Wernimont & Campbell, 1968). The path diagram in Figure 1 depicts this perspective.

2. General ability theory - job performance is best predicted by tests of general intelligence, i.e. cognitive ability, because jobs are learned as abilities in their own right, the learning of which is governed by general cognitive ability (Humphreys, 1979). Relationships in this theory are depicted in Figure 2.

These two perspectives view the causal relationships of general intelligence differently. In the specific abilities model, intelligence (cognitive ability) determines verbal and numerical abilities and other specific abilities that in turn are the key determinants of work performance. Thus, performance of specific types of clerical work is caused by related, specific abilities. On the other hand, the general abilities model indicates that general intelligence is the direct cause of performance skills as well as the specific verbal and numerical abilities. The implications for employment testing are great. If the general abilities model is correct, general test batteries like the GATB would provide the best prediction of performance, whereas if the specific abilities model is correct, tests customized to specific jobs would enhance prediction of work performance.

Several forms of evidence supporting the general abilities theory were found in studies conducted for the USES (test report number 44, 1983). Correlations between validity coefficients from 515 GATB validation studies indicated that specific abilities tend to cluster into general, more inclusive abilities. That finding was supported by a confirmatory factor analysis. Job performance

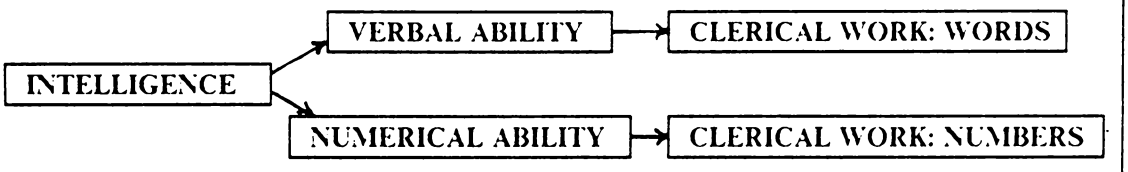

Figure 1. Specific Abilities Model (USES Test Report No. 44, 1984).

appeared to be related to three general abilities: cognitive (GVN), perceptual (SPQ), and psychomotor (KFM). However, SPQ was highly correlated with both GVN and KFM. Therefore, it appeared that the three general abilities span only two dimensions: cognitive and psychomotor. Multiple regression analysis of criterion scores on the three general abilities indicated the dominance of cognitive ability in predicting performance providing further support for the general abilities theory. From these regression analyses, equations were developed for use in calculating the USES's VG scores, cf. Table 3. USES test report number 45, 1983).

Table 3 displays the regression weights obtained from regression analyses and the derived formulas for all five levels of job complexity, discussed above. The first box from the left indicates the regression weights when all three basic abilities were considered and displays the multiple R (validity coefficient). In the second box, results of a second regression analysis where perceptual ability (SPQ) was dropped are displayed. In the last box are the equations chosen to calculate VG scores for each job complexity level. Job Family 4 which is of interest in this study is at complexity level three. Therefore, to calculate the VG score for clerical jobs, cognitive ability (GVN) is weighted .73 and added to psychomotor ability weighted .27. Estimated validity of that predictor is .53. Note that in the JF4 formula, which is allegedly the best predictor of clerical ability, test scores of perceptual ability (SPQ) which includes clerical perception (Q) are not included. On the other hand, results of meta-analytic studies suggest that more specific abilities [7] may be better predictors of performance than more general abilities.[8]

---

[7] Specific abilities refers to the nine separate abilities measured by the GATB.

[8] General abilities refers to the three measures of general ability.

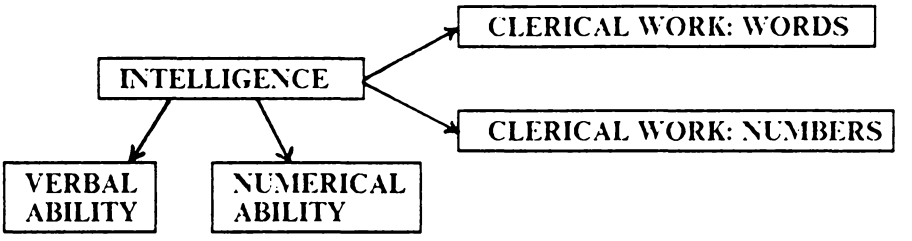

Figure 2. General Ability Model (USES Test Report No. 44, 1984).

For 'computing and account-recording' clerical jobs, Pearlman, Schmidt and Hunter (1980) found that mean observed validity for Q, clerical perception (.26), was higher than that of G (.23), a composite measuring cognitive ability. After correction for statistical artifacts, estimated true validity for clerical perception was .53, (as high as Job Family 4 scores), whereas estimated true validity for G was .49. On the other hand, this study reported the highest estimated true validity (.63) for a more specific component of cognitive ability, "reasoning ability." In another meta-analytic study, Schmitt et al. (1984) found that mean observed validity of large sample studies was higher for specific aptitude tests (.27) than for general mental ability tests (.25).

To summarize, there is evidence supporting both views of the relationship of general and specific abilities to work performance. Since the VG testing program is based on general ability scores, if specific abilities are better predictors than general abilities, the accuracy of Job Family 4 percentile scores for predicting clerical work performance may be lower than specific ability scores. Since there is conflicting evidence on which model is correct, the validities of both general and specific measures of ability are reported.

Summary and Research Questions

The research of Schmidt & Hunter suggests that cognitive abilities are valid predictors for all jobs. The basic premise of the VG testing program is that cognitive ability measures are valid for all jobs and for jobs in Job Family 4 (including clerical jobs), validity can be enhanced by a com-

Table 3. Development of VG Scores from Abilities Combinations

COMPLEXITY LEVEL	REGRESSION WTS							VG FORMULAS			
	GVN	SPQ	KFM	R	GVN	KFM	R	GVN	SPQ	KFM	R
1	.40	0.19	.07	.59	.52	.12	.57	.59	+ .30	+ .11	.59
2	.75	-.26	.08	.60	.58	.01	.58	1.0			.58
3	.50	-.08	.18	.53	.45	.16	.53	.73		+ .27	.53
4	.35	-.10	.36	.51	.28	.33	.50	.44		+ .56	.50
5	.16	-.13	.49	.49	.07	.46	.49	.13		+ .87	.49

posite of cognitive ability and psychomotor ability. Therefore: the first set of hypotheses address the issue of the validity of Job Family 4 scores as predictors of job performance for three 'computing and account-recording' clerical jobs.

H1a: Job Family 4 percentile scores are valid predictors of work performance.

H1b: Job Family 4 scores have higher validity than the measure of cognitive ability for 'computing and account-recording' clerks.

An Employment Service study found that the validity of cognitive and psychomotor ability tests is consistent across clerical jobs. Other research suggests that job differences within the clerical occupational category moderate validity. Moreover, the research of Schmidt-Hunter suggests that the ability-performance relationship is consistent across work locations, whereas, some studies have provided direct evidence of situational moderators. Therefore, the second pair of hypotheses address the questions of job and work location effects on the validity of ability tests.

H2a: Within the clerical occupation the relationship between cognitive and psychomotor ability and work performance is consistent across jobs, i.e., job differences do not moderate the test-performance relationship.

H2b: The relationship between cognitive and psychomotor ability and work performance is consistent across work locations, i.e. work locations do not moderate the ability-performance relationship.

The USES VG testing program is based on the premise that work performance is best predicted by general abilities. This view is not consistently supported by research findings. Therefore, the last set of hypotheses test the proposition that the correlation between measures of general abilities and performance will be stronger than correlations between specific abilities and performance.

H3a: The validity coefficient for general cognitive ability will be higher than for the more specific clerical, verbal or numerical abilities.

H3b: The validity for general psychomotor ability will be higher than for the more specific measures of motor coordination, finger or manual dexterity.

Chapter 3: Methodology

In the previous chapter, literature pertaining to three research questions identified in Chapter One was reviewed and those questions were formulated into research hypotheses. This chapter begins with an overview of the context and the basic research strategy of the present study. The subjects, measures, and procedures used in collecting and analyzing the data are then described, and the chapter concludes with a brief discussion of the limitations of this study.

Overview

This study was conducted at the request of a bank that was interested in exploring the feasibility and potential utility of employment testing. The bank had been experiencing substandard performance from new employees. Management suspected that one possible solution to the problem was to improve their selection process. That process consisted of screening application forms for minimum qualifications, checking references, conducting personal interviews, and, finally, credit checks. It was thought that the use of test data would improve the quality of employment decisions and thus improve organizational performance and possibly reduce turnover.

Two ability test batteries were chosen for evaluation. The first was developed specifically for banks and was supplied by the American Bankers Association to its member banks. The second was the General Aptitude Test Battery (GATB) provided by the Employment Service. The local office of the Employment Service was soliciting employer participation in the VG Testing Program. Since the Job Family percentile score was a new predictor, no direct evidence of validity, was then available. Conveniently, the interests of the bank, the Employment Service and the author meshed in this study.

Due to time and cost constraints, a concurrent validation strategy was used. This form of criterion-related validation involves collecting test scores and performance scores on job incumbents at approximately the same time. Such a strategy meets professional standards (Principles for the validation and use of personnel selection procedures, 1980; Standards for educational and psychological testing, 1985). Test and performance scores were collected on 219 current employees within a period of three and one half weeks.

Sample

At the time of the study (summer 1985), the host organization, a bank with multiple branch banks and operations centers, was one of the largest banks in the mid-Atlantic area. The sample was drawn both from operations centers in two metropolitan areas located in geographically-distinct regions of the state and from branch banks in the second region. The first area is the state capital which is in close proximity to several other major cities, while the second city is situated within a rural region and is home to the bank's corporate headquarters. Operations centers process branch bank transactions centrally for several regions, and official records are maintained at that site. The branches are typical for a commercial bank, providing many financial services to customers.

Three clerical jobs were studied: bookkeeper, proof-machine operator, and teller. Bookkeepers and proof-machine operators worked at operations centers. Tellers worked in seventeen

branch banks located in the second metropolitan area. For these jobs, the bank had job descriptions which had been formally updated within two years of the study. To assess their current accuracy, incumbents were observed and several of them were interviewed at each location. Supervisors were also given a copy of the appropriate job description and asked to identify any deficiencies or inaccuracies. None were identified. Industry wide, the content of these jobs has remained relatively stable. The following Dictionary of Occupational Titles (1977) classifications accommodated the bank's job descriptions.

- 210.382-018 Bookkeeper (Clerical) II - Keeps one of a set of financial records: Verifies and enters details of transactions... Summarizes details on separate ledger...Balances books and compiles reports...

- 217.382-010 Proof-machine Operator (Fin. Inst.) - Operates proof machine to sort, record, and prove records of bank transactions...Depresses keys to sort items...Positions items in machine to be endorsed...Totals tapes and locates and corrects errors...Attaches tape to sorted batches...Proves deposits, checks, debits, and credits listed on batch sheet.

- 211.362-018 Teller (Fin. Inst.) - Receives and pays out money, and keeps records of money and negotiable instruments...Receives checks and cash for deposit...Enters deposits...Issues receipts...Cashes checks... Places holds on accounts...Orders supply of cash...

As verified by an Employment Service official, the Dictionary of Occupational Titles (DOT) codes placed each job in Job Family 4 of the VG testing program.

Even though all three jobs are in Job Family 4, they differ on a number of dimensions. Those dimensions are: tasks diversity, the level of customer contact, and money handling. Task diversity is greatest for tellers, followed by bookkeepers, with proof operators having the least. Tellers have direct customer contact. Bookkeepers have some indirect contact with customers who call with questions, whereas proof-machine operators have no customer contact. A major responsibility for tellers is settling their drawer. However, neither bookkeepers nor proof operators handle bank

monies. On the other hand, the position of proof operator entails significantly more psychomotor activity than do the positions of bookkeeper or teller. Approximately 80-90% of a proof operators time is spent sitting at a keyboard, inputting data from cancelled checks as quickly as possible. Interestingly, Colbert and Taylor (1978) placed these jobs into two separate job families: bookkeeper and teller in Job Family 1, and proof-machine operator in Job Family 5. As mentioned below, the American Bankers Association also places these jobs into two job families.

Table 4 presents the demographic characteristics of the subjects and sample sizes for each job and location. Approximately 89% of the 219 subjects were female and 26% were blacks. At location 1, there was a significantly higher percentage of blacks than location 2. For example, 37% of the bookkeepers at location 1 were black compared to 20% at location 2. On average, proof-machine operators were nine years younger than the rest of the sample and there was less age variance than with the other jobs. The sample was educationally homogeneous, most were high school graduates but very few were college graduates. Proof-machine operators had significantly less tenure than the others, but, the variance was high for all three. However, a notable difference in tenure between locations is apparent. For example, on average, bookkeepers at location 2 have more than twice the tenure of those at location 1.

Measures

Predictors

As noted above, both the General Aptitude Test Battery (GATB) and the American Banking Association's Test Battery for Entry Level Positions in Banking (ABA) were administered to all sub-

Table 4. Sample Characteristics

	N	GENDER		RACE		AGE		EDUCATION		TENURE	
		M	F	B	W	MN	SD	MN	SD	MN	SD
BKKR	90	5	85	26	64	30.6	12.2	12.8	1.2	24.5	26.1
Loc1	46	4	42	17	29	30.1	13.5	13.0	1.3	16.0	18.8
Loc2	44	1	43	9	35	31.1	10.8	12.5	1.1	33.3	29.7
PROOF	69	17	52	16	53	21.1	3.5	13.0	1.2	11.1	12.2
Loc1	32	11	21	14	18	22.5	4.0	13.2	1.3	9.5	6.7
Loc2	37	6	31	2	35	19.8	2.4	12.8	1.2	12.5	15.4
TELLR	60	1	59	4	56	29.4	9.6	12.5	.9	20.0	18.6
TOTAL	219	23	196	46	173	27.2	10.4	12.8	1.2	19.0	21.2

jects. The GATB[9] has been used and constantly refined by the United States Employment Service since 1947. It consists of twelve timed tests, eight paper-and-pencil tests and four apparatus tests which measure nine factor analytically derived basic abilities (Manual for the USES General Aptitude Test Battery, III, 1970). Industrial and organizational psychologists recognize it as one of the most carefully developed and thoroughly researched abilities test batteries (Guion, 1965; Tests in Print III, 1983). In the volumes of Tests in Print, 566 studies of the GATB are referenced. The USES, with apparently good reason, claims it is the best validated test battery in existence for use in occupational selection (USES Test Research Report No. 43, 1983).

Seven of the nine GATB scales [10]are of interest to this study, general intelligence (G), verbal ability (V), numerical ability (N), clerical perception (Q), motor coordination (K), finger dexterity (F), and manual dexterity (M). Clerical perception is of interest since the three jobs studied were clerical. Two of three general ability composite scores are used in computing Job Family 4 percentile scores. Cognitive ability (GVN), is comprised of general intelligence, verbal ability, and numerical ability. Psychomotor ability (KFM) is comprised of motor coordination, finger dexterity, and manual dexterity. Job Family 4's raw score composite (JFR) is composed of GVN and KFM. Finally, Job Family 4 percentile scores (JF4) are derived from racial distributions of the raw score composites and the end product of the VG testing program. JF4 scores are the only test result received by employers.

The Technical Manual for the GATB reports test-retest reliabilities for most aptitudes are between .80-.90. For general learning ability (G), verbal ability (V), and numerical ability (N), however, reliabilities with adults are often greater than .90. On the other hand, lower reliabilities are generally reported for finger (F) (.69-.80) and manual (M) (.70-.80) dexterity.

In 1971, the American Bankers Association (ABA) began a nationwide study of entry level clerical positions to provide a technically sound and legally defensible test battery to their member banks. Development proceeded in three phases. The first phase consisted of extensive job analysis of clerical positions in the banking industry. A "sophisticated task and job attributes questionnaire"

---

[9] The proprietary nature of both the GATB and ABA test batteries preclude their inclusion in this report.

[10] Cf. Table 5

was distributed to ABA member banks. Several hundred banks participated, providing data on over 1,100 jobs. Through factor analysis, 28 job classifications were identified and grouped into seven major job families, as follows:

- Equipment Operating - Proof machine operator, Encoding machine operator, Computer operator, and Coin machine operator;
- Clerical Service - Secretary, Stenographer, Receptionist, and Telephone operator;
- Clerical and Machine Operation - Clerk typist, File clerk, Messenger, and Statement clerk;
- Collector - (only one collector job);
- Equipment Operation and Arithmetical - Control clerk (data processing), Stock clerk, and Transit clerk;
- Clerical and Arithmetical - Credit clerk and Vault Attendant; and
- General Clerical - Teller, General clerk, Bookkeeping clerk, and Management trainee.

Note that proof-machine operator is in Job Family One (JF1) while teller and bookkeeper are in Job Family Seven (JF7). That grouping is consistent with Colbert & Taylor's (1978) findings mentioned above.

In the second phase (1975), the ABA administered an experimental test battery to approximately 10,000 entry level applicants in 441 banks. However, only 8,799 tests were complete enough to analyze. The average sample size from participating banks was approximately 20. Fourteen test scores were factor analyzed and reduced to four abilities scales: language (LANG), quantitative (QUANT), and clerical (CLER) abilities, and finger dexterity (DEX). Although the quantitative and language abilities loaded as one cognitive factor (COG), they were retained as separate scales due to improved validities across job families using the specific abilities measures. Reliability estimates (Cronbach Alphas) reported in the Technical Manual for the four tests all exceeded .89.

The third phase entailed gathering criterion measures. Supervisor judgments of acceptability, rehire status, and overall evaluation along with an objective measure, length of service, were collected in 1975 and 1976. Usable data covered about 4,700 persons who had been hired and worked six full months. Next, validation analyses were conducted. The validity for the entire sample (across clerical jobs plus management trainee) was "very low." However, when final validation analysis was conducted on the basis of the seven job families developed in phase one, job differences

were found to be a strong moderator of predictor-criterion relationships. The following validities (multiple R) were reported:

- Bookkeeper: .40, $p < .01$
- Proof operator: .36, $p < .01$
- Teller: .28, $p < .01$

In its final form, the ABA test battery (cf. Appendix C) consists of seven brief paper-and-pencil tests designed to measure four abilities, (see Table 5).

Since the ABA test battery is relatively new, little information is published regarding the construct validity of its four scales. Thus, inter-correlations of scores from both batteries were calculated to provide a form of multi-trait multi-method (MTMM) analysis. Campbell and Fiske (1959) developed the MTMM technique to provide evidence of construct validity. Tests designed to measure a specific trait should correlate highly among themselves, i.e. demonstrate convergent validity. On the other hand, tests designed to measure one trait should not correlate highly with tests designed to measure unrelated traits. Thus, if the GATB and ABA tests for verbal ability, quantitative ability, clerical ability, and finger dexterity measure the same constructs, they should correlate highly, whereas, other correlations should be lower, evidencing discriminant validity.

Performance Measures

Three measures of employee performance were initially obtained for this study. Subjective ratings from the company's performance appraisal system and ratings on the Descriptive Rating Scale were obtained for all jobs. An objective measure, inputs per hour, was collected on proof operators. However, due to missing data and lack of uniformity in administration of the bank's performance appraisal, that measure was dropped.

Table 5. Predictor Measures

GATB ABILITIES

Symbol	Name	Tests
G	General Intelligence	Vocabulary + Arithmetic Reasoning + Three Dimensional Space
V	Verbal Ability	Vocabulary
N	Numerical Ability	Computation + Arithmetic Reasoning
S	Spatial Ability	Three Dimensional Space
P	Form Perception	Tool Matching + Form Matching
Q	Clerical Perception	Name Comparison
K	Motor Coordination	Mark Making
F	Finger Dexterity	Assemble + Disassemble Washers
M	Manual Dexterity	Place + Turn Board Pegs
GVN	Cognitive Ability	Composite of G + V + N
SPQ	Perceptual Ability	Composite of S + P + Q
KFM	Psychomotor Ability	Composite of K + F + M
JFR	Job Family 4	Raw score composite of GVN & KFM
JF4	Job Family 4 Percentile Scores	Scores from race distributions

ABA TEST BATTERY ABILITIES

Symbol	**Name**	**Tests**
COG	Cognitive Ability	Language + Quantitative Ability
LANG	Language Ability	Spelling + Vocabulary
QUANT	Quantitative Ability	Arithmetic + Mathematical Reasoning
CLER	Clerical Ability	Number Matching + Name Matching
DEX	Finger Dexterity	Mark Triangle in Circles

Table 6. Performance Measures

DESCRIPTIVE RATING SCALE (DRS)

Symbol	Name
y1-QUANT	Quantity of Work
y2-QUAL	Quality of Work
y3-ACC	Accuracy of Work
y4-KNOW	Job Knowledge
y4-SKILLS	Job Skills
y6-GLOBL	Overall Performance
y7-COMPST	(y1 + y2 + y3 + y4 + y5)

OBJECTIVE PERFORMANCE

Symbol	Name
IPH-3mo	Monthly average of inputs/hour, 3rd month after testing
IPH-6mo	Monthly average of inputs/hour, 6th month after testing

Descriptive Rating Scale

The Descriptive Rating Scale (DRS), an Employment Service instrument, is a conventional graphic rating instrument with scales intended to measure five performance dimensions plus an "overall" (global) rating (cf. Appendix B). The performance dimensions assessed are quantity of work, quality of work, accuracy of work, job knowledge, and job skills. Each dimension and the global assessment are measured with five point single item scales. [11] Three additional questions are included in the DRS. The first two enable assessment of the quality of supervisor ratings: 1) extent of observation of subordinate performance, and 2) length of working acquaintance with ratees. The last question (reason for termination) was not relevant to this study. The DRS was used in some of the 515 validation studies cumulated in the Employment Service's meta-analytic study of GATB results. In 425 of those studies, the criterion was supervisor judgments of proficiency.

Use of subjective ratings in validation studies has long been accepted. A majority of published studies have employed performance ratings as criteria (Landy & Farr, 1980; Lent, et al., 1971; Monahan & Muchinsky, 1983). A more preferable method to operationalize criterion constructs is with hard (objective) measures. However, for many jobs there are no hard data available. Consequently, the only option is to use subjective measures. Various techniques have been employed to remove the possibility of various rater bias, (BARS, BOS, etc.), but, no systematic research has examined the relative effectiveness of those various techniques (Bernardin & Beatty, 1984). Thus, a graphic rating scale is as acceptable as any other subjective criterion instrument.

In the study reported here, it was initially believed that each DRS performance dimension constituted a separate criterion measure. Thus, an examination of the relationship between predictors and various facets of performance was originally planned. For example, it was thought that psychomotor ability might predict 'quantity of work' better for proof operator than the other two jobs because of the job requirements, or, that 'job knowledge' might be more closely associated with cognitive ability. However, based on results of tests for dimensionality (reported in Chapter

---

[11] Lissitz & Green (1975) found that reliability increased up to five scale points.

Four) that plan was abandoned. An additional criterion was formed by adding subjects' DRS dimension scores. Although this composite could have been formed in several ways, the various alternatives are comparable to the unit weighting procedure used here (Cascio, 1982).

Psychometric information regarding the DRS was not available from the Employment Service or general literature.[12] To estimate the reliability of the DRS ratings, interrater and intrarater reliabilities were obtained. Interrater reliabilities were calculated for all tellers and location 2 proof operators from independent ratings provided by two superiors. These raters were believed by the bank to be equally knowledgeable of subordinate performance. Since pooled judgments are likely to be more accurate than either judgment separately, the average of the ratings for these jobs was used as the primary criterion in this study. However, equally knowledgeable raters were unavailable for any bookkeepers or proof operators at location 1. Therefore, in those cases an intrarater estimate of reliability was used. Each supervisor provided two sets of independent ratings on each subject. The first set was used as the criterion in this study since supervisors may have been more conscientious in the first instance, and there is no theoretical benefit from combining parallel ratings where it is assumed rater biases would be relatively constant. Table 7 summarizes the type of reliability estimate and the DRS measure for each job.

Inputs Per Hour

Objective measures also were obtained for proof-machine operators. Mechanical devices on proof-machines record the number of strokes per hour made by a proof-machine operator. The measure is called "inputs per hour" (IPH). These machines also compute balances. If an entry does not balance, the machine will not accept another entry until the imbalance error is corrected. Consequently, the objective measure of strokes per hour is a measure of both quality and quantity of performance. Reliability is excellent for this measure. Since the strokes per hour are mechanically recorded, human error can only occur as the inputs are transcribed to a record sheet. Because

---

[12] Ghiselli (1973) observed that subjective performance appraisal reliability typically ranged from .60 - .80.

Table 7. DRS Criteria

Job	Reliability Estimate	Criteria Operationalizations
Bookkeeper	Intrarater	First Ratings
Proof - Loc 1	Intrarater	First Ratings
Proof - Loc 2	Interrater	Average Ratings
Teller	Interrater	Average Ratings

those records are used to figure incentive bonuses, the mechanical counters receive regular maintenance. Both employer and employee verify records for accuracy. Interviewed employees were unaware of errors in those reports.

When proof machines need repair or are inoperable due to supply shortages, the machines are turned off. Thus, the time meter which is controlled by the machine switch does not run and the IPH is automatically adjusted in such circumstances. Since a proof operator's incentive pay would be adversely affected if the machine's timer were left running, this source of error (Thorndike, 1949) is highly improbable.

The most important concern with the IPH measure is the variance in individual performance. In a study of over 1000 keypunch and proof-machine operators, Klemmer & Lockhead (1962) noted significant differences in the daily error rates of individual operators. Further, the rate of errors was independent of mean production level. Consequently, short term averages could differ significantly from long term. Thus, monthly averages of strokes or items per hour were used in the analyses. Two different months were selected for analysis because, due to high turnover many proof operators were recently employed. Three months after testing, all but eight proof operators had completed their probationary period. Yet, with passage of each week, the sample was reduced due to turnover. Six months after testing, by definition, all subjects had completed their probationary period. Therefore, IPH averages were collected for the third and sixth months.

Moderators

Two potential moderators were examined. Job was treated as a three level categorical variable. As noted above, the three jobs studied were: bookkeeper, proof-machine operator, and teller. Work location was treated as a two level categorical variable. Operations center 1 and operations center 2 were similar in terms of the jobs studied and the work environment.

Data Collection Procedures

All bookkeepers and proof-machine operators in both geographic locations completed both the GATB and ABA test battery between July 17 and July 25 1985. The ABA battery was administered first, followed two days later by the GATB. In order to reduce disruption of work in the operations centers, eleven testing sessions were conducted on location. From four to nineteen persons took the tests at one time. In the branch banks, a different arrangement was required. Since work flow varies considerably, it was not feasible to test tellers during working hours. Consequently, the host organization paid tellers and required them to take both test batteries on Saturday, August 10, 1985. Refreshments were provided at two breaks, one in the morning and one in the afternoon. The GATB was given in the morning, followed by lunch which was also provided by the company. The paid luncheon not only provided a break between tests, it also evidenced company support for the project.

The ABA battery was administered by the researcher. For larger sessions, the bank provided an assistant, and a volunteer assisted with the large teller group. Prior to testing, the assistant and volunteer were trained to monitor the testing and answer subjects questions. Test administration was in accordance with ABA administrative procedures. To ensure anonymity of test results, answer booklets were coded. After the testing was completed, Personnel staff hand scored the coded

tests. However, only the researcher knew the code to identify subjects. A random scoring audit detected only two minor errors in the scoring of 55 tests.

GATB testing was conducted by Employment Service (ES) staff, on company premises in each geographic region. The ES provided a test administrator and 1-4 assistants, depending on the number scheduled for testing. Assistants monitored the test, answered questions, and demonstrated the proper procedures for the board tests. After testing sessions, ES staff transported the completed test forms from the testing sites to their repective local offices where they were machine scored.

Three potential sources of random error in test scores were identified which are related to test administration: conditions, testing procedures, and employee motivation. Inconsistency in test administration is the first potential source. Differences in the testing environment, such as noise level, lighting, temperature, comfort of chairs, and table space can differentially affect individual performance levels between test administrations. With one possible exception, the administration of both test batteries were equivalent with respect to environment. The possible exception was the session at which 46 tellers were tested. Since the room was more crowded than at any other administration, those test scores could reflect environmental contamination.

A second potential source of random error in administration concerns differences in procedure, such as extent of instructions and explanation of the tests, equivalence of test forms, and/or time allowed for tests. Although three different ES staff persons administered the GATB, they closely followed written ES procedures. Thus, test score variance attributable to this source should be minimal. That is also the case with the ABA testing since there was only one administrator.

The most troublesome potential source of error is the motivation of test takers. The literature suggests that incumbent motivation is not equivalent to that of job applicants taking employment tests (Guion, 1965). It is impossible to create a testing situation for incumbents identical to that experienced by applicants seeking employment, but a number of steps were taken to ensure employee motivation and control for any dysfunctional factors. First, incumbents were strongly encouraged to do their best on the test. Top management and immediate supervisors alike cooperated with the researcher in encouraging incumbents. Second, at an introductory meeting, the researcher offered to provide confidential feedback on test results to interested individuals. The feedback was

billed as informational and potentially useful for self-development. Third, in an attempt to reduce low-level performance from employees "going through the motions," it was also announced that the names of all persons who "minimally passed" the test would be pooled and fourteen names drawn for $50 savings bonds. "Minimally passed" was never defined, and all employees were included in the drawing. Finally, to alleviate employee fear of being tested, assurances were given that individual scores would be released neither to the company nor to anyone else, and that test results would in no way affect job status.

As soon as test scores were received from the VEC, employees were informed that they could personally receive their scores. Their response was overwhelming, approximately 80% of the subjects obtained their scores. This response suggests acceptable levels of personal motivation to perform well on the tests. Another possible indicator of positive motivation was the essentially normal distribution of the VG percentile scores (cf. Chapter 4).

On the criterion side, the DRS was completed by teller supervisors and one other management employee at the bank branches. At all branches, two persons deemed to have equal familiarity with teller performance rated each teller, thus providing a basis for estimates of interrater reliability. However, as previously mentioned, for bookkeeper and proof operator supervisors, two administrations were necessary. With the exception of proof operators at operations center 2, two raters with knowledge of the subjects' performance were not available.

All supervisors received training in small groups of three or less prior to completing the DRS. Sessions lasting approximately one hour began with a general review of performance appraisal principles, including examples of rater biases. The DRS questionnaire was discussed item by item and raters practiced rating a subordinate on all scales. It was recommended, however, that all subordinate subjects be rated on one performance dimension at a time to reduce halo error. (By design, the format facilitated that procedure.) Each supervisor then received a form with the names of subjects who were their subordinates. A period for questions concluded the training session. Supervisors were to complete and return the forms within one week.

The first set of DRS. forms was distributed to bookkeeper and proof operator supervisors before administration of the predictors. After a three week interval, a second set was distributed to

those same supervisors.[13] Two factors were considered in determining the interval between DRS administrations. The first was rater memory. If the interval was too short, the second rating might be a function of a supervisor's desire for consistency and ability to remember earlier ratings. The second consideration was that true performance can change over time. If the interval was too long, employee performance could substantially change, warranting different ratings. These competing considerations suggest conflicting intervals, i.e., long v. short. Allowing three weeks between administrations was deemed appropriate.

To maintain the independence of both ratings, the fact that there would be two administrations was kept secret. Raters might otherwise have kept a copy of their first evaluation to copy responses onto the second. Therefore, only top management was initially informed. When each supervisor received the second rating form, a brief explanation of its purpose was given: to evaluate the instrument, not the raters. In addition, each supervisor was provided the mean, maximum, and minimum scores awarded to his/her subordinates in the first wave. While that could tend to inflate the reliability coefficient, that procedure should have provided an estimate of the upper bound of instrument reliability.

Analyses

The research question of primary interest was, are VG percentile scores valid predictors of performance? Therefore, the first research hypothesis was

H1a: Job Family 4 percentile scores are valid predictors of work performance.

This can be restated as:

---

[13] Since no one in the organization had access to predictor scores, the second wave was not a "contaminated criterion."

OH1a: A positive, statistically significant, correlation exists between Job Family 4 percentile scores and all measures of performance for all three jobs.

To test operational hypothesis 1A, a Pearson product-moment correlation was calculated on the whole sample for VG job family 4 percentile scores with all criterion measures. The null hypothesis is that the correlation coefficient is zero. If the validity coefficient is positive and statistically significant, the null is rejected. Two other methods of reporting findings have been used in the literature: 1) correcting the correlation coefficient for attenuation and restriction in range (Block, 1963, 1964; Bobko, 1983; Carmines & Zeller, 1979; Greener & Osburn, 1979; Gross & Fleischman, 1983), and, 2) indicating the confidence interval (Hunter et al., 1982). Correction for attenuation in validity coefficients is appropriate only for unreliability in criterion measures. The predictor as it is formulated is of interest, not an estimate of its constructs if they were perfectly measured. A confidence interval indicates a range of validity coefficients that 95 times out of one hundred would include the population validity coefficient. As reported in Chapter Four, the interval is used to compare results of this study with estimates of true validity reported in the literature. Finally, to determine if the assumption of linearity was appropriate, the eta statistic was calculated (Cohen & Cohen, 1983).

The Employment Service study indicated that a composite score of general cognitive and psychomotor abilities is the best predictor of performance for Job Family 4. Therefore,

H1b: Job Family 4 scores have higher validity than cognitive abilities scores.

This can be restated as:

OH1b: The validity coefficient of Job Family 4 scores correlated with proficiency scores is greater than the validity coefficient of cognitive ability scores correlated with proficiency scores.

The second set of hypotheses concern situational moderators of validity. Two three-level variables, job and work location, were investigated as potential moderators. Regarding the possibility of job differences moderating the "degree" of the abilities-performance relationship, it was hypothesized:

H2a: Jobs within clerical occupations do not moderate the abilities-performance relationship.

In operational terms:

OH2a: The validity coefficients for Job Family 4 with all DRS criteria are stable (equivalent) for bookkeepers, proof operators, and tellers.

The appropriate statistical test is of the homogeneity of correlation coefficients (Arnold, 1982; Algera et al., 1984; Cohen & Cohen, 1983)

$$Z = \frac{z_1 - z_2}{\left[\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right]^{1/2}}$$

where: Z = the standard score (Z) value, and z sub1 and z sub2 = the Fishers z transforms for each job sample. Pearson product-moment correlations computed on subsamples by job were compared for their equivalence using the test for the homogeneity of correlation coefficients which indicates whether there is evidence that one of the correlation coefficients was obtained from the sample of a different population. If there is no such statistical evidence, then the conclusion is to fail to reject the hypothesis that the correlation coefficients are equal.

OH2b: The validity coefficients for abilities-performance relationships are stable (equivalent) across work locations 1 and 2.

The identical statistical analysis applied to test for job effects was used to test for work location effects.

The third set of hypotheses pertain to the relative magnitude of validity coefficients for specific and general abilities. Conceptually, the final set of hypothesis stated that:

H3a: Across criteria, the validity for measures of cognitive ability will be higher than for verbal or numerical ability.

This may be restated as:

OH3a: The validity coefficients for cognitive ability (GVN or COG) across criteria are greater than the validity coefficients for verbal (V or LANG) or numerical (N or QUANT) ability and greater than the validity coefficient for clerical (Q or CLER) ability.

The second hypothesis related to this topic was:

H3b: The validity for psychomotor ability will be higher than for motor coordination, finger or manual dexterity.

Thus:

OH3b: The validity coefficients for psychomotor ability (KFM) across criteria are greater than the validity coefficients for motor coordination (K), finger dexterity (F), manual dexterity (M) or dexterity (DEX).

Pearson product-moment correlations were run for the specified predictors with all criteria. Then, the magnitudes of average validities across DRS scales were subjectively compared.

Results of these analyses are reported in Chapter Four and discussed in Chapter Five.

Limitations

The primary limitations of this study are related to its sample size, and the performance measures used because they impact the power of the test for validity. Limitations of the power of typical validation studies, particularly due to small sample size, have recently resulted in less confidence being placed in findings of one local validation study (Guion, 1987). In this study, sample size limitations within jobs and locations increased the probability of sampling error and reduced the probability of finding a statistically significant relationship. Consequently, the ability of this study to detect moderators was limited.

Power is related to statistical significance and is the ability to reject the null hypothesis (that the validity coefficient equals zero) when the null hypothesis is false. Three parameters: sample size, effect size and the alpha level impact the power of a validation study. The alpha level is traditionally set at .05. Effect size is an unknown prior to conducting a study, but can be estimated based on the literature. Therefore, in planning research, sample size is manipulated in order to obtain desired power. According to the Employment Service estimate of $r = .53$, the estimated n-size of this study

appeared to be appropriate to achieve power between .80 - .90 during the feasibility study. Unfortunately, the size of observed effects was significantly lower.

Assessment of the quality of the DRS was limited by loss of the anticipated bank performance appraisal scores. In this study, ability to assess the validity of the DRS was limited. Moreover, since an objective criterion was available only for proof operators, the job moderator issue was tested only with the subjective criterion instrument.

To summarize, the limitations of this study are related to the power of the tests conducted. First, smaller sample sizes resulting from subgrouping subjects into job and location categories reduced the probability of finding a statistically significant relationship when the population validity is above zero. Second, the criterion instruments may not have sufficiently captured performance differences, which would result in a reduced effect size. That, in turn, would decrease the power of the tests, the probability of finding a statistically significant relationship.

Chapter 4: Findings

In the preceding chapter, the methodology for this study was described. This chapter first reports the findings regarding the psychometric qualities of the measures and then presents the results of the hypothesis tests.

Psychometric Qualities of Measures

The psychometric qualities of the GATB have already been addressed in Chapter Three. While the reliability of the ABA test battery was also reported, evidence of construct validity was not available. Since the construct validity of the GATB measures is well documented, the correspondence between ABA measures of constructs similar to GATB measures was investigated. Finally, in this section, the reliability estimates and assessment of the dimensionality of the DRS scales are reported.

To investigate the construct validity of the ABA scales, evidence of convergent and discriminant validity was sought. Convergent validity should exist between measures intended to capture the same construct. In this case there were four parallel measures, but the scale labels are

somewhat misleading. The ABA's language ability (LANG) is composed of two tests, vocabulary and spelling. On the other hand, the GATB verbal ability (V) is only a vocabulary test. So, while the constructs of LANG and V appear nominally equivalent, the direct comparison should be between the ABA vocabulary test and GATB verbal ability. The same is true with the two clerical ability scales. The ABA clerical ability scale (CLERICL) is composed of name matching and number matching. However, the GATB clerical perception (Q) tests only name matching. Therefore, the comparison should be between the single ABA name matching test and the GATB Q. The two batteries also approach the assessment of dexterity differently. The ABA dexterity scale measures finger dexterity with a paper and pencil test. However, the GATB finger dexterity measure is an apparatus test. On the other hand, the GATB motor coordination test is a paper and pencil instrument similar to the ABA DEX. Therefore, the more appropriate comparison is between the latter two measures. One direct comparison of ability measures was possible. The ABA's quantitative ability (QUANT) consists of two tests, math reasoning (word problems) and arithmetic problems which closely parallel the GATB N scale's two tests, arithmetic reasoning (word problems) and computation (arithmetic problems). Intercorrelations of the four measures are displayed in Table 8.

Correlations between measures of a similar construct should be higher than between those of dissimilar constructs. Thus, the correlations on the diagonal in Table 8 should be higher than the off diagonal values. With the exception of the correlation of DEX and K, that is what was found. A plausible explanation for that low dexterity correlation is the restriction in range of DEX scores. A high percentage of the subjects placed pencil marks in all 320 boxes within the five minutes allotted. Subjects who completed the task ceased working while the slower subjects continued marking boxes during the remainder of the test time. In summary, the findings reported in Table 8 provide some evidence of the construct validity of the ABA language, quantitative, and clerical ability tests.

The remaining psychometric considerations pertain to the criteria. Eight criterion measures were used in this research. As indicated in Chapter Three, seven were subjective measures obtained from the DRS instrument and one (IPH) was an objective measure of proof-operator performance.

Table 8. Construct Validity of the ABA Test Battery

ABA	General Aptitude Test Battery			
	V	N	Q	K
VOC	.54	.40	.18	NS
QUANT	.50	.73	.32	NS
NAME	.32	.31	.51	.34
DEX	NS	NS	.20	.42

NS = not significant at alpha = .05.
N = 210 for all correlations

Two aspects of the DRS scales were investigated, their reliability, and dimensionality. Last, the relationship of the objective measure to the subjective measures was examined.

As indicated in Chapter Three, two methods of estimating DRS scale reliabilities were used. Interrater reliabilities were calculated with two independent supervisor ratings for tellers and proof-machine operators in location 2. Intrarater reliabilities were calculated with two separate DRS ratings from individual supervisors on bookkeepers and location 2 proof-machine operators. Reliability estimates are reported in Table 9.

Interrater reliabilities for Proof-machine operators ranged from .85 for the composite scores to .64 for the quality dimension. For Tellers, interrater reliabilities ranged from .82 for the accuracy dimension to .35 for the quantity dimension. The low reliability for proof operator DRS quality scores may be related to differences in standards between teller supervisors and branch managers. Competitive pressures on the bank had resulted in new expectations of tellers to aggressively sell the bank's services. Since managers were assessed on number of new accounts opened, etc., such activities could have been weighted more heavily by branch managers than by supervisors who were more accountable for the accuracy of transactions. At the time of this study, teller performance standards were being vigorously debated. Differences in rater standards may be reflected as well in the reliabilities for skill diversity and in the global assessment. Even with the divergence in super-

Table 9. DRS Scale Reliabilities

DIMENSION	INTERRATER				INTRARATER			
	Proofs-L2		Tellers		Bookkeepers		Proofs-L1	
	N	r	N	r	N	r	N	r
QUANT	37	.74	57	.35	90	.91	64	.83
QUAL	37	.64	57	.63	90	.78	64	.79
ACC	37	.78	57	.82	90	.80	64	.76
KNOW	37	.77	57	.67	90	.82	64	.75
SKILL	0	*	57	.52	90	.74	0	*
GLOBL	37	.69	57	.53	90	.77	64	.71
COMPST	37	.85	57	.73	90	.91	64	.83
Average		.75		.61		.81		.80

All correlations significant at $p < .01$.
* Dimension not assessed for proof operators.

visor standards, though, the average of teller interrater reliabilities across criteria was .61.[14] The average of interrater reliabilities across criteria for proof-machine operators was higher, .75. On the other hand, intrarater reliability estimates ranged from .71 for the global estimate of proof operator performance to .91 for the quantity of work of bookkeepers. The average of these intrarater reliabilities was about .80, which places them in the upper range of reliabilities observed by Ghiselli (.60-.80) in amassing his large collection of validation studies.

Evidence of DRS dimensionality was also sought. As displayed in Table 10, intercorrelations of the five dimension scales indicate that two measures, quality and accuracy of work, share much common variance since their correlation is .82. That common variance may be due to the fact that the DRS is a standardized instrument, not tailored to the jobs studied. For these clerical jobs, it is difficult to conceptualize the difference between the two constructs. The four other dimension intercorrelations are between .60 and .74, which may suggest that they capture slightly different dimensions of perceived performance. However, to assess dimensionality, Nunnally (1960) suggested computing a Cronbach's alpha. If alpha is less than .64, evidence suggests more than one dimension. For the five DRS dimensions alpha was .92 which indicates a very high degree of consistency

[14] This falls within typical performance criterion reliabilities and equals the .60 which is assumed in Schmidt-Hunter's validity generalization studies (Pearlman, et al., 1980).

of ratings across the five dimensions. This evidence suggests that only one dimension is captured by the DRS instrument. The high degree of correlation of dimension scores with both the global and the composite scores also suggests that the dimension tapped is an "overall" rating.

Although the sample of proof operators after three and six months was small, the relationship of inputs per hour to subjective measures was examined because there is little literature on such comparisons. Nevertheless, as typically reported in that sparse literature, this objective measure appears to measure a different aspect of performance than the subjective measures. The strongest relationship (.61) was with quantity of work, as anticipated, which provides some evidence of construct validity for that DRS measure.

Hypothesis Tests

The distributions of observed predictor scores are summarized in Table 11.
The mean of Job Family 4 scores was 58.5, slightly above the theoretical mean of 50, and the standard deviation was 25.1, somewhat less than the theoretical 34, which indicates some restriction in range. However, virtually the full range of possible scores was obtained and tests for skewness and kurtosis indicated that these scores were approximately normally distributed. Specific ability test scores from both test batteries were also approximately normally distributed. Since correlational analysis is robust with regard to violations of the normalcy assumption (Cohen & Cohen, 1983), these predictors are suitable for such analysis.

The distributions of observed criterion scores are summarized in Table 12.
Quality was the only criterion dimension where the full range of scores was not utilized. No one who had survived the probationary period received the lowest rating for quality of work. Mean dimension scores were generally around 3.5 which is higher than the theoretical 2.5 mean and standard deviations were below the theoretical 1.7 standard deviation. This suggests some restriction in range, which is not out of the ordinary for these types of measures.

Table 10. Criterion Scale Intercorrelations

	N	Y1	Y2	Y3	Y4	Y5	Y6	Y7
QUANT	211							
QUAL	211	.70*						
ACC	211	.68*	.82*					
KNOW	211	.65*	.72*	.74*				
SKILL	174	.60*	.71*	.71*	.69*			
GLOBL	211	.70*	.81*	.80*	.77*	.75*		
COMPST	174	.84*	.91*	.92*	.88*	.77*	.87*	
IPH-3mo.	34	.32*	.33*	.32**	.46*	n/a	.44*	.38*
IPH-6mo.	26	.61*	.36**	.54*	.55*	n/a	.53*	.55*

*p = .05; **p = .10

In the objective measure for proof-machine operators, there was more than a 50% difference between the high and the low performance scores (more than 1400 IPH difference). The standard deviations were approximately 20% of the mean scores which was approximately the same as with the DRS dimensions, suggesting that the reduced variance in DRS ratings may reflect actual performance variance rather than rater bias. Tests for skewness and kurtosis suggested that both the input per hour scores and DRS ratings were relatively normally distributed. As with the predictors, the evidence suggests that the criterion scores are appropriate for correlation analyses.

Hypothesis One

The first set of hypotheses deals with the validity of Job Family 4 percentile scores for predicting work performance. Table 13 presents the observed validity coefficients for the total sample, correlating Job Family 4 with each DRS criterion measure. Also calculated were the corrected correlation coefficient and the 95% confidence interval.

Five of the seven validity coefficients were valid at the .05 level of statistical significance. The remaining two criteria had positive relationships with JF4 and were statistically significant if the alpha level is relaxed to the .10 level. Observed validities ranged from a low of .11 to a high of .21. Due

Table 11. Descriptive Statistics for Predictors

ABILITY	N	MEAN	STD	MIN	MAX
JF4	219	58.5	25.1	2	99
G	219	97.7	16.6	11	150
V	219	98.6	15.2	60	143
N	219	99.9	15.7	54	145
S	219	99.1	19.7	58	163
P	219	116.5	19.1	27	163
Q	219	120.9	16.4	81	185
K	219	116.1	19.2	0	157
F	219	103.0	22.7	31	158
M	219	116.7	22.1	-55	186
LANG	210	49.1	11.7	20	80
QUANT	210	18.5	6.8	4	34
CLERICL	210	108.3	28.9	36	178
DEX	210	285.9	48.0	111	320

to the unidimensionality of the DRS ratings, the intended investigation of separate predictor-criterion dimension relationships (mentioned in Chapter 3) was aborted.

Corrected validity coefficients ranged from .12 to .23 which is significantly below the Employment Service's worst case estimate of .33, mentioned in Chapter Two. To estimate true validity, the observed validity coefficient was corrected for attenuation using Spearman's (1904) correction formula (Lee, Miller & Graham, 1978). Following that procedure, observed validity was divided by the square root of the criterion reliability. To provide conservative estimates, the highest reliability estimate for each criterion measure, whether interrater or intrarater reliability, was used (cf. Table 11). However, since there was little evidence of direct restriction in range for test scores or criterion measures, no correction was made for that potential artifact.

It must be noted, that neither the corrected validity coefficients nor the upper limit of the confidence intervals approached the Employment Service's estimated JF4 true validity of .53 or the .49 corrected mean indicated in VG studies of abilities tests for clerical positions.

Hypothesis 1b relates to the VG testing program claim that the Job Family 4 composite of general cognitive ability and psychomotor ability (JF4) is a better predictor of performance than the more specific cognitive ability scores (GVN) for jobs examined in this study. Table 14 compares

Table 12. Descriptive Statistics for Criteria

CRITERIA	N	MEAN	STD	MIN	MAX
QUANT	211	3.53	.76	1	5
QUAL	211	3.54	.78	2	5
ACC	211	3.51	.87	1	5
KNOW	211	3.59	.85	1	5
SKILL	174	2.87	.94	1	5
GLOBL	211	3.39	.76	1	5
COMPST	174	17.00	3.66	6	25
IPH-3mos	37	2032	436	1112	3234
IPH-6mos	28	2120	398	1354	2800

the validity of JF4 scores and its raw score composite (JFR) with two GATB measures of cognitive ability (GVN and G) across criteria. JFR was included to determine if there is any effect on validity from the testing program's use of percentile scores

For none of the criteria is JF4 validity the highest and the average validity coefficient for JF4 is lower than for the other three. With all but the skill diversity criterion, JFR has the highest validity coefficient. Thus, if these relatively small differences are meaningful, the evidence suggests that prior to transformation, the composite of GVN and KFM may be a better predictor of work performance, than VG percentile scores. Moreover, the percentile scores may be a slightly poorer predictor than general cognitive ability alone. If any true differences exist between these predictors, perhaps the raw score composite is a better predictor than composite percentile scores. However, there is no evidence to support the hypothesis that VG percentile scores for job family 4 are better than cognitive ability for predicting performance.

Hypothesis Two

Moderators of the abilities-performance relationship were dealt with in the second set of hypotheses. The null hypothesis was that validity coefficients for the three jobs studied were homogeneous. A finding that statistically significant validity coefficients were not equivalent across jobs,

Table 13. JF4 VALIDITY for TOTAL SAMPLE

Criteria	N	Observed Validity	Corrected Validity	95% C.I.
QUANT	211	.11**	.12	-.03 - .25
QUAL	211	.13**	.15	-.01 - .27
ACC	211	.17*	.19	.03 - .31
KNOW	211	.21*	.23	.07 - .35
SKILL	174	.15*	.17	.01 - .29
GLOBL	211	.16*	.18	.02 - .30
COMPST	174	.19*	.20	.05 - .33

* p = .05; ** p = .10

would suggest that the abilities-performance relationship was moderated by jobs. In Table 15, observed validity coefficients for the three separate jobs across criteria are displayed.

Since no two validity coefficients were statistically significant at the .05 level for any criterion,[15] to accommodate an assumption of the test for the homogeneity of correlation coefficients - that the coefficients are statistically significantly greater than zero - alpha was relaxed to p.10. Under those conditions, the test was applied to KNOW for bookkeepers and tellers. The Z score was 1.06. Since the critical value at the .05 level is 1.65, the null hypothesis that the correlations were homogeneous was not rejected.[16] Due to lack of statistical significance, it was inappropriate to test the other pairs of coefficients for homogeneity. Thus, there was no statistical evidence in these data that jobs moderated JF4 validity.[17] However, when average DRS validities were compared across jobs, the evidence suggests that average validities were equal for bookkeeper and proof operator, but,

---

[15] The major limitation to detecting statistically significant relationships was the relatively small sample sizes of the job subgroups.

[16] Note that as n gets smaller, a greater difference between validity coefficients is required to reject the null hypothesis that the correlation coefficients are not equal.

[17] The reader may be interested to know that moderated regression technique also failed to detect moderation (form).

Table 14. VALIDITY of COMPOSIT v. GENERAL ABILITY

Criteria	N	JF4	JFR	GVN	G
QUANT	211	.11**	.16*	.11**	.11**
QUAL	211	.13**	.17*	.17*	.16*
ACC	211	.17*	.20*	.19*	.17*
KNOW	211	.21*	.24*	.23*	.20*
SKILL	174	.15*	.16*	.19*	.21*
GLOBL	211	.16*	.21*	.19*	.16*
COMPST	174	.19*	.23*	.23*	.25*
DRS ave.		.16	.20	.19	.18

* p = .05; ** p = .10

substantially higher (.09 or 64%) for teller. That raises questions about the actual equality of validities across the three jobs. If the sample size were larger and the test more powerful, would a significant difference be detected?

Hypothesis 2b is concerned with potential moderation of validity by work location. Observed validity coefficients for the two work locations are reported in Table 16.

No statistical tests were conducted for the homogeneity of correlation coefficients across locations because no coefficients were statistically significant for location 2. However, several observations are worth noting. First, observed validity coefficients were more stable across locations for bookkeeper than for proof operator. The large difference in results between the two locations for proof operators (DRS ave. difference of .26) suggests that even though the sample sizes were small, location moderated the validity. In location 1, the environment for proof operators was more turbulent than in location 2. Location 1 turnover was higher, while departmental efficiency and morale were lower. A plausible explanation for location moderation (if it existed) is that ability is more critical to performance for low tenure employees, whereas, motivation is more critical to performance for high tenure employees in such a situation. Second, since the differences between locations appear to vary by job, a job by location interaction is possible.

Table 15. JF4 Validities by Job.

Criteria	Bookkeeper (n = 90)	Proof (n = 64)	Teller. (n = 57)
QUANT	.17**	.09	.05
QUAL	.06	.15	.23**
ACC	.11	.14	.29*
KNOW	.18**	.18	.35*
SKILL	.15	***	.17
GLOBL	.13	.15	.27*
COMPST	.16	.15	.27*
DRS ave.	.14	.14	.23
IPH-3mo.		.35**	
IPH-6mo.		.40*	

* p = .05; ** p = .10; *** This criterion not measured for proofs.

Table 16. JF4 Validities by Work Location

Criteria	Location 1			Location 2		
	Tot	Bkkr	Prf	Tot	Bkkr	Prf
N	(73)	(46)	(27)	(81)	(44)	(37)
QUANT	.20**	.22	.12	.05	.05	.07
QUAL	.17	.07	.25	-.01	-.06	.06
ACC	.18	.07	.30	.03	.09	.03
KNOW	.26*	.12	.45*	.07	.20	-.02
SKILL	.14	.11		.12	.12	
GLOBL	.22*	.12	.33**	.02	.10	.00
COMPST	.22*	.13		.11	.11	
DRS ave.	.20	.12	.29	.06	.09	.03

* p = .05; ** p = .10

Hypothesis Three

The third set of hypotheses address the question of whether general abilities are better predictors of performance than more specific abilities. Hypothesis H3a states that the validity coefficients for cognitive ability across criteria are greater than the validity coefficients for verbal, numerical, or clerical ability. Thus, this hypothesis involves specific comparisons of cognitive ability as measured by GVN and COG with various measures of more specific abilities. If this hypothesis is true, then one would expect to find higher validity coefficients for GVN and COG than for clerical ability (Q or CLER). Table 17 displays the validity coefficients for measures of both general and specific abilities, across all criteria.

Table 18 displays the validities of GATB and ABA cognitive ability measures, both general (GVN and COG) and more specific, with two global criteria, the clinical judgment (Y6) and the mechanical composite (Y7).

Table 17. VALIDITY COEFFICIENTS

PREDICTOR	QUANT	QUAL	ACC	KNOW	SKILL	GLOBL	COMPST	IPH-6mo
GVN	.11**	.17*	.19*	.23*	.19*	.19*	.23*	.35*
G	.11**	.16*	.17*	.20*	.21*	.16*	.25*	.45*
V	.06	.10	.14*	.19*	.17*	.12**	.19*	.20
N	.12**	.18*	.19*	.22*	.12	.21*	.19*	.25
Q	.10	.11	.07	.11	.09	.14*	.10	.23
KFM	.17*	.07	.11	.11**	.14**	.13*	.16*	.29
K	.11	.05	.05	.09	.02	.13*	.06	.46**
F	.14*	.04	.06	.07	.18*	.07	.17*	-.12
M	.16*	.07	.14*	.10	.14**	.11	.15**	.36**
COG	.11	.23*	.20*	.26*	.14**	.21*	.21*	.24
LANG	.06	.18*	.15*	.20*	.07	.15*	.14**	.24
QUANT	.15*	.25*	.24*	.28*	.24*	.25*	.28*	.10
CLER	.15*	.17*	.19*	.21*	.22*	.23*	.19*	.48*
DEX	.07	.05	.07	.07	.08	.07	.08	.28

*p < .05 **p < .10
N = 211 except for IPH-6mo, where n = 28

Table 18. AVERAGE GENERAL COGNITIVE & SPECIFIC ABILITY VALIDITIES

	GEN. COG.		VERBAL		NUMERICAL		CLERICAL	
GATB	.19*	.23*	.12**	.19*	.21*	.19*	.14*	.10
ABA	.21*	.21*	.15*	.14**	.25*	.28*	.23*	.19*

First, with the GATB the highest validity is .23 for general cognitive ability (GVN) with the mechanical composite. However, with the clinical judgment, the validity (.19) is lower than the validity (.21) for numerical ability. So, evidence related to the ability theories is mixed with the GATB. On the other hand, with the ABA validity is highest for both criteria with numerical ability (QUANT), .25 and .28 versus .21 for both criteria with general cognitive ability. Thus, evidence with the ABA tends to support the specific ability theory. The nature of the test batteries could account for the difference in results across methods. The GATB are more general than ABA tests where test items are closer to problems typically encountered in a banking context. For example, the GATB clerical perception test consists only of name matching, whereas the ABA also contains a number matching test. For the three computing and account-recording jobs studied here, the relevance of number matching ability to work performance is obvious and may account for the superiority of that ability over the other less specific measures in predicting performance. Evidence from the objective criterion (for proof operators) lends support to this specificity argument. With IPH data (cf. Table 17), the validity for CLER was .48, whereas the validities with GVN (.35) and COG (.24) were lower. The evidence indicates that clerical ability was a significantly better predictor of objective performance than either of the general cognitive ability measures.

One possible explanation for the difference in results between the study reported here and the Employment Service study lies in the range of clerical jobs studied. Computing and account-recording clerks may encounter name and number matching tasks more frequently than typical clerks, which the Employment Service study sampled.

Hypothesis 3b states that psychomotor ability will be a better predictor of performance than motor coordination, finger dexterity, or manual dexterity. Table 19 displays the validities of GATB

and ABA psychomotor ability measures, both general (KFM) and more specific, with two global criteria, the clinical judgment (Y6) and the mechanical composite (Y7).

The evidence for psychomotor abilities provided by the GATB is not consistent with hypothesis 3b, which states that the validity for KFM will be greater than for K, F, or M. With the subjective criterion, validity with motor coordination is equal to that with the general psychomotor ability measure. With the mechanical composite criterion the validity for finger dexterity (.17) is slightly higher than for general psychomotor ability (.16). In addition, evidence of validities with the objective criterion suggest that there were significant differences between the general psychomotor ability validity (a non-significant .29) and the statistically significant coefficients of the more specific motor coordination (.46) and manual dexterity (.36). Thus, for the small sample of proof operators, the evidence supported the specific abilities model.

In summary, evidence from this study does not support the general abilities theory upon which the VG testing program is based. At best, with the subjective criterion, GATB measures provided weak support for the general abilities models and only with cognitive abilities. The strongest evidence supported the specific abilities model. Evidence with both subjective and objective criteria suggested that specific abilities were better predictors than general cognitive ability, or than general psychomotor ability, (cf. Tables 17, 18 and 19). objective criterion.

To conclude, hypothesis 1a, that JF4 is a valid predictor of work performance, was supported. But, 1b, that JF4 is a better predictor than general cognitive ability, was not supported. With hypothesis two, no statistical evidence of job moderation was found. However, average DRS scale validities suggested that validity was substantially higher for teller. Similarly, no statistical evidence of work location moderation was found. However, the substantial difference in average DRS scale validities for proof operator across locations suggested that work location was a moderator. Finally, the general abilities model was not supported by the data.

Table 19. AVERAGE GENERAL PSYCHOMOTOR & SPECIFIC ABILITY VALIDITIES

	PSYCHOMOTOR		MOTOR COORDINATION		FINGER DEXTERITY		MANUAL DEXTERITY	
GATB	.13*	.16*	.13*	.06	.07	.17*	.11	.15**
ABA			.07	.08				

Chapter 5: Conclusions and Recommendations

In the preceding chapter, results of the empirical tests of the validity of Job Family 4 percentile scores, tests for moderation of the ability-performance relationship, and evidence regarding the specific or general abilities models was reported. This chapter discusses some of those results, their implications and limitations, and offers some recommendations for future research.

Conclusions and Implications

The hypothesis of primary interest, that VG Job Family 4 scores are valid predictors of work performance for computing and account-recording clerks, was supported. Both subjective and objective criteria were positively and significantly related to Job Family 4 scores. The mean observed validity for the DRS performance dimensions was .16, and the corrected validity was .18. Validity estimates from this study, however, are relatively low and substantially lower than the Employment Service estimate of .53. Even the average upper boundary of the 95% confidence interval (.30) is significantly lower than that estimate. For similar jobs, traditional estimates of cognitive ability tests validity (.26) and meta-analytic estimates (.23) were higher (cf. Table 1).

While there are alternative explanations for these differences, the results here suggest that pure cognitive ability tests might be better predictors of work performance than Job Family 4 percentile scores. Average observed validity of general cognitive ability across performance criteria was .19, corrected to .21, whereas the mean observed validity for JF4 was .16, corrected to .18. General cognitive ability (G) also had a higher validity with the objective IPH-6mo measure (.45) than did JF4 (.40). The results also suggest that cognitive ability is a better predictor than psychomotor ability which had a mean observed validity of .13. If these results reflect reality, it raises questions about job categories. Placing the computing and account recording clerks studied here in the VG program's Job Family 4 did not enhance the magnitude of validity. Consequently, more research needs to address the issue of the best categorization of jobs for performance prediction purposes. Second, the results tends to reaffirm the predominance of cognitive abilities found by Schmidt-Hunter and colleagues. Third, there are implications for the VG testing program. Since the quality of prediction might be less than optimal, it appears that predictor formulas should be subject to further review as direct validation becomes available. Consequently, there is a need for more large sample direct studies to compare Job Family 4 percentile scores for computing and account-recording clerks against other predictors and/or composites from the GATB. Research should also be conducted in several other occupational groups within Job Family 4 or, as suggested in Chapter Three, on other DOT classifications of clerical jobs to discover if other direct tests of JF4 validity find significantly lower coefficients than estimated by the Employment Service. Similar validation studies should also be conducted in other four job families. Job Family 4 results may not generalize to other VG job families since complexity levels vary and most have significantly fewer DOT jobs in their classification, i.e. the jobs may be more homogeneous with regard to ability requirements.

For employers, direct evidence from large sample studies may provide more accurate estimates of validity for use in utility analyses. Results of this study suggest that the Employment Service estimates of effect size may be substantially inflated. Thus, the accuracy of decisions concerning participation in the VG testing program based on utility analysis may be improved by obtaining more direct evidence of the program's validity. As for the VG program's estimated national

utility of $50-100 billion, unless the effect size is greater for other jobs, that estimate may also be substantially inflated.

Testing of the second set of hypotheses failed to produce statistical evidence of job moderators within the clerical occupation. As discussed in Chapter Four, however, that result may have been a function of the power deficiency of the tests. Mean observed validity differences suggested that statistically significantly different coefficients might have been detected if sample sizes had been larger. If that apparent difference in the degree of validity for jobs reflects reality, then for purposes of performance prediction, more narrow classification of jobs than found in the VG testing program may be appropriate. Consequently, large sample studies that have the power to detect differences in the degree of validity across jobs, if it exists, should be conducted. Another potential approach would be to test results of this study of "computing and account-recording clerks" with large sample studies of jobs in other DOT clerk categories.

Statistical tests also failed to provide evidence of a work location moderator effect. In this case, there were no statistically significant validity coefficients for one work location. However, subjective comparison of the mean observed validities for proof-machine operators again suggests that if sample sizes were larger, significant differences in validity could have been detected across locations. Average validity for proof operators across DRS criteria in location 2 was near-zero, whereas in location 1 it was .29. One potential explanation for this difference relates to the work environment. At location 2 where validity was low, management philosophy could be characterized as results oriented, but supportive. After training, employees were encouraged to experiment to see what worked best for them, and unless problems developed, management did not interfere. Since there was no public contact, employees were permitted to dress as they pleased, and the atmosphere was very relaxed. On the other hand, management philosophy at location 1 was authoritative. Management had a high profile. The dress code was formal and strictly enforced. These differences could affect the motivation of employees. If turnover and tenure are any indication, location 1 had high turnover and low tenure, whereas location 2 had relatively low turnover and higher tenure. As suspected, average productivity was higher at location 2. How these differences could relate to validity difference is that higher levels of motivation at location 2 may compensate for ability defi-

ciencies. On the other hand, since motivation appears to be low in location 1, the ability-performance relationship is much more obvious. Again, these exploratory results should encourage larger sample studies that have the power to detect moderators.

As James et al. (1986) indicate, more empirical evidence is needed to address the situational specificity debate. Meta-analysis appears to be a useful tool for dealing with a large body of information. However, in light of evidence that meta-analytic estimates of the variance of validity coefficients are accurate only when n-sizes are over 200 (Osburn et al, 1983), a strategy to provide evidence on moderator effects is to cumulate results of a large number of large sample studies. Since most extant studies have smaller sample sizes, new initiatives must begin to conduct and publish validation studies with sample sizes exceeding 200. Future studies must also pay more attention to criterion development and assessment prior to conducting a validation study. Sackett et al. (1986) found that unreliability of measures significantly reduced effect size and consequently power.

Results of testing hypothesis three with GATB data were inconclusive. Contrary to the results of the Employment Service study across the full spectrum of clerical jobs, the GATB data obtained here supports neither the specific abilities nor the general abilities model. On the other hand, evidence from the more tailored ABA battery supported the specific ability model. Average validity for numerical ability (.24) was higher and for clerical ability (.19) was equal to the validity of general cognitive ability (.19 for both GVN & COG). An obvious possible implication of that finding is that the VG testing program's assumption that general ability measures predict work performance better than specific ability measures may be incorrect. Consequently, more direct large sample studies are needed to address this issue. An implication for employers with access to more customized test batteries like the ABA is that they may get better prediction in selection with a tailored battery.

Study Limitations

Limitations of the study reported here are discussed in terms of threats to internal and to external validity. The most serious limitation to this study was the power of several statistical tests, particularly those testing for moderators. Three parameters affect the power of a test: effect size, sample size, and alpha level. As discussed above, the alpha level is traditionally set at .05. Consequently, in determining the feasibility of this study estimates of the effect size drawn from the VG literature were used to determine if the sample at this site was large enough to achieve the power desired. The judgment was affirmative. However, since the effect size was significantly lower than anticipated, the power of the tests was significantly reduced. That resulted in many statistically non-significant validity coefficients for jobs and work locations across the DRS criteria. The other potential threats to internal validity are related to proof operators. Turnover reduced the IPH score sample size. While the n-size was small, there was nevertheless good differentiation and relative normalcy in the scores and high validities with the predictors. The other threat came from an inability to control for experience. Some proof operators work varied part-time hours, so length of service would not be appropriate. Furthermore, some had previous experience, but the quality and quantity of that experience was not assessed.

With regard to threats to external validity there was essentially one, which was related to sample characteristics. The implicit target population was the population of all computing and account-recording clerks in the United States. Rather than a carefully planned study to appropriately sample that population, this field study sample could be characterized as a "sample of convenience" (Cook & Campbell, 1979). Therefore, the generalizability of results to the target population must be tentative, pending further evidence. That means, the results may not generalize to other clerical jobs, other organizations, or perhaps other geographic regions of the country.

Bibliography

Arnold, H.J. (1982) Moderator variables: A clarification of conceptual, analytic, and psychometric issues. Organizational Behavior and Human Performance, 29, 143-174.

Arnold, H.J. (1984) Testing moderator variable hypotheses: A reply to Stone and Hollenbeck. Organizational Behavior and Human Performance, 34, 214-224.

Ability testing: Uses, consequences, and controversies. (1982) A.K. Wignor & W.R. Garner, eds. vols. I & II, Washington, DC: National Academy Press. Psychological Bulletin, 51, 201-238.

AERA-APA-NCME (1985) Standards for educational and psychological testing. Washington, DC: APA.

Albright, L.E., Glennon, J.R., & Smith, W.J. (1963). The use of psychological tests in industry. Cleveland: Allen.

Algera, J.A., Jansen, P.G.W., Roe, R.A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt- Hunter procedure. Journal of Occupational Psychology, 57, 197-210.

APA, Division of Industrial and Organizational Psychology (Division 14). (1980). Principles for the validation and use of personnel selection procedures (2nd ed.). Berkeley, CA: Author.

Baker, D.D., & Terpstra, D.E. (1982). Employee selection: Must every job test be validated? Personnel Journal, 61, 602-605.

Bechtolt, H.P., & Carroll, J.B. (1965). General Aptitude Test Battery. In O.K. Buros (Ed.), The sixth mental measurements yearbook (pp. 1023-1029). Highland Park, NJ: Gryphon.

Bemis, S.E. (1968) Occupational validity of the General Aptitude Test Battery. Journal of Applied Psychology, 52, 240-249.

Bennett, G.K. (1969). Factors affecting the value of validation studies. Personnel Psychology, 22, 265-268.

Bernardin, H.J., & Beatty, R.W. (1984). Performance appraisal: Assessing human behavior at work. Boston: Kent.

Block, J. (1963). The equivalence of measures and the correction for attenuation. Psychological Bulletin, 60, 152-156.

Block, J. (1964). Recognizing attenuation effects in the strategy of research. Psychological Bulletin, 62, 214-216.

Blum, M.L., Greene, E.B., & Taylor, H.R. (1953). General Aptitude Test Battery. In O.K. Buros (Ed.), The Fourth Mental Measurements Yearbook (pp. 686-693). Highland Park, NJ: Gryphon.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. Journal of Applied Psychology, 68, 584-589.

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. Applied Psychological Measurement, 4, 385-398.

Boehm, V.R. (1982). Are we validating more but publishing less? The impact of governmental regulation on published validation research - an exploratory investigation. Personnel Psychology, 35, 175-187.

Boehm, V.R. (1972). aNegro-white differences in validity of employment and training selection procedures: Summary of research evidence. Journal of Applied Psychology, 56, 33-39.

Bullock, R.J., & Svyantek, D.J. (1985). Analyzing meta- analysis: Potential problems, an unsuccessful replication and evaluation criteria. Journal of Applied Psychology, 70, 108-115.

Burke, M.J. (1984). Validity generalization: A review and critique of the correlation model. Personnel Psychology, 37, 93-116.

Callender, J.C., & Osburn, H.G. (1980). Development and test of a new model for validity generalization. Journal of Applied Psychology, 65, 543-558.

Callender, J.C., & Osburn, H.G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. Personnel Psychology, 66, 274-281.

Callender, J.C., & Osburn, H.G. (1982). Another view of progress in validity generalization: Reply to Schmidt, Hunter and Pearlman. Journal of Applied Psychology, 67, 846-852.

Callender, J.C., Osburn, H.G., Greener, J.M., & Ashworth, S. (1982). Multiplicative validity generalization model: Accuracy of estimates as a function of sample size and mean, variance, and shape of the distribution of true validities. Journal of Applied Psychology, 67, 859-867.

Campbell, D.T. (1959) Convergent and discriminant validtation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Campbell, J. & Pritchard, R. (1976) Motivation theory in industrial and organizational psychology. In Dunnette, M.D. (Ed) Handbook of industrial and organizational psychology. Chicago: Rand-McNally.

Carmines, E.G. & Zeller, R.A. (1979) Reliability and validity assessment. Beverly Hills: Sage.

Cascio, W.F. (1982) Costing human resources: The financial impact of bahavior in organizations. Boston: Kent.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. NY: Academic Press.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Colbert, G.A. & Taylor, L.R. (1978) Empirically derived job families as a foundation for the study of validity generalization. Study II. Generalization of selection test validity. Personnel Psychology, 31, 355-364.

Comrey, A.L., Froehlich, C.P., & Humphreys, L.G. (1959). General Aptitude Test Battery. In O.K. Buros (Ed.), The fifth mental measurements yearbook (pp. 695-700). Highland Park, NJ: Gryphon.

Cook, T.G. & Campbell, D.T. (1979) Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand-McNally.

Cook, T.G., & Leviton, L.C. (1980). Reviewing the literature: A comparison of traditional methods with metaanalysis. Journal of Personality, 48, 449-472.

Cornelius, E.T., III, Schmidt, F.L., & Carron, T.J. (1984). Job classification approaches and the implementation of validity generalization results. Personnel Psychology, 37, 247-260.

Dreher, G.F., & Sackett, P.R. (1983). Perspectives on employee staffing and selection: Readings and commentary. Homewood, IL: Irwin.

Dunnette, M.D. (1973). Performance equals ability and what? University of Minnesota, Department of Psychology, Technical Report No. 4009.

Dunnette, M.D. (1963). A modified model for test validation and selection research. Journal of Applied Psychology, 47, 317-323.

Employee selection: Legal and practical alternatives to compliance and litigation. (1983) E.E. Potter, ed. Equal Employment Advisory Council, monograph.

Feidler, F. (1967) A theory of leadership effectiveness. New York: McGraw-Hill.

Forehand, G.A. (1968). On the interaction of persons and organizations. In R. Tagiuri & G. Litwin (Eds) Organizational climate: Explorations of a concept. Boston: Harvard Business School.

Fine, S.A. (1974) Functional Job Analysis: An approach to a technology for manpower planning. Personnel Journal, 53, 813-818.

Fleishsman, E.A. & Harris, E.F. (1962). Patterns of leadership behavior related to employee grievances and turnover. Personnel Psychology, 15, 43-56.

Ghiselli, E.E. (1966). The validity of occupational aptitude tests. New York: John Wiley & Sons.

Ghiselli, E.E. (1972). Comment on the use of moderator variables. Journal of Applied Psychology, 56, 270.

Ghiselli, E.E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Greener, J.M., & Osburn, H.G. (1979). An empirical investigation of the accuracy of corrections for restriction in range due to explicit selection. Applied Psychological Measurement, 3, 31-41.

Griggs v. Duke Power Co. (1971) 401 U.S. 424.

Gross, A.L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. Applied Psychological Measurement, 7, 227-237.

Gross, A.L., & Kagen, E. (1983). Not correcting for restriction of range can be advantageous. Educational and Psychological Measurement, 43, 389-396.

Guion, R.M. (1987). Changing views for personnel selection research. Personnel Psychology, 40, 199-213.

Guion, R.M. (1965). Personnel testing. New York: McGraw-Hill.

Guion, R.M. (1967). Personnel selection. Annual Review of Psychology, 18, 191-216.

Guion, R.M. (1976). Recruiting, selection and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 777-828). Chicago: Rand-McNally.

Guion, R.M., & Cranny, C.J. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 67, 239-244.

Gullickson, H. (1950) Theory of mental tests. New York: Wiley.

Hawk, J.A. (1970) Linearity of criterion-GATB aptitude relationships. Measurement and Evaluation in Guidance, 2, 249-251.

Hays, W.L. (1973). Statistics for the social sciences. New York: Holt, Rinehart & Winston.

Hersey, P. & Blanchard, K.H. (1977) Management of organization behavior. Englewood Cliffs, New Jersey: Prentice Hall.

House, R. (1971) A path-goal theory of leadership Administrative Science Quarterly, 16, 321-338.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Hunter, J.E., & Schmidt, F.L. (1978). Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. Journal of Applied Studies, 68, 1-11.

Hunter, J.E., & Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M.D. Dunnette & E.A. Fleishman (Eds.), Human performance and productivity vol. 1 (pp. 233-284). Hillsdale, NJ: L. Earlbaum Associates.

Hunter, J.E., Schmidt, F.L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). Meta-analysis: Cummulating research findings across studies. Beverly Hills: Sage.

Hunter, J.E., Schmidt, F.L. & Pearlman, K. (1982). History and accuracy of validity generalization equations: A response to the Callender and Osburn reply. Journal of Applied Psychology, 67, 853-858.

James, L.R., DeMaree, R.G. & Mulaik, S.A. (1986) A note on validity generalization procedures. Journal of Applied Psychology, 71, 440-450.

Ladd, R.T. & Cornwell, J. (1986) The accuracy of meta-analysis estimates. Paper presented to the Society for Industrial and Organizational Psychology. Chicago, Illinois.

Lawler, E.E., III (1966) The mythology of management compensation. California Management Review, 9, 11-22.

Lawler, E.E., III (1971) Pay and organizational effectiveness: A psychological view. New York: McGraw-Hill.

Lawshe, C.H. (1985). Inferences from personnel tests and their validity. Journal of Applied Psychology, 70, 241-343.

Lee, R., Miller, K., & Graham, W. (1982). Corrections for restriction of range and attenuation in criterion related validations studies. Journal of Applied Psychology, 67, 637-639.

Lent, R.H., Aurbach, H.A., & Levin, L.S. (1971). Predictors, criteria, and significant results. Personnel Psychology, 24, 519-533.

Lissitz, R.W. & Green, S.B. (1975) Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.

Madigan, R.M., Scott, K.D., Deadrick, D.L., & Stoddard, J.A. (1986) Employment testing: the U.S. Job Service is spearheading a revolution, The Personnel Administrator, 31, 102-112.

McCormick, E.J. (1979) Job analysis: Methods and applications. New York: AMACOM.

McGregor, D. (1960) The human side of the enterprise. New York: McGraw-Hill.

Monahan, C.J., & Muchinsky, P.M. (1983). Three decades of personnel selection research: A state-of-the-art analysis and evaluation. Journal of Occupational Psychology, 56, 215-225.

Muchinsky, P.M. (1979). Some changes in the characteristics of articles published in the Journal of Applied Psychology over the past 20 years. Journal of Applied Psychology, 64, 455-459.

Nunnally, J.C. (1960) Tests and measurements. New York: McGraw-Hill.

Osburn, H.G., Callender, J.C., Greener, J.M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. Journal of Applied Psychology, 68, 115-122.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Peterson, D.J. (1974). The impact of Duke Power on testing. Personnel, 51, 30-37.

Porter, L.W. (1966). Personnel Management. Annual Review of Psychology, 17, 395-422.

Raju, N.S., & Burke, M.J. (1983). Two new procedures for studying validity generalization. Journal of Applied Psychology, 68, 382-395.

Reilly, R.R. & Chao, G.T. (1982) Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Sackett, P.R., Harris, M.M. & Orr, J.M. (1986) On seeking moderator variables in the meta-analysis of correlation data: a Monte Carlo investigation of statistical power and resistance to Type I error. Journal of Applied Psychology, 71, 302-310.

Sackett, P.R., & Wade, B.E. (1983). On the feasibility of criterion-related validity: The effects of range restriction assumptions on needed sample size. Journal of Applied Psychology, 68, 374-381.

Schmidt, F.L., Gast-Rosenberg, I., & Hunter, J.E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

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

Schmidt, F.L., & Hunter, J.E. (1978). Moderator research and the law of small numbers. Personnel Psychology, 31, 215-232.

Schmidt, F.L., & Hunter, J.E. (1980). The future of criterion-related validity. Personnel Psychology, 33, 41-60.

Schmidt, F.L., & Hunter, J.E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F.L., & Hunter, J.E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. Journal of Applied Psychology, 68, 407-414.

Schmidt, F.L., & Hunter, J.E. (1984). A within setting empirical test of the situational specificity hypothesis in personnel selection. Personnel Psychology, 37, 317-326.

Schmidt, F.L., Hunter, J.E., & Caplan, J.R. (1981). Validity generalization results for two job groups in the petroleum industry, Journal of Applied Psychology, 66, 262-273.

Schmidt, F.L., Hunter, J.E., McKenzie, R.C., & Muldrow, T.W. (1979). Impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64, 609-626.

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

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Shane, G.S. (1979). Further tests of the Schmidt-Hunter bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. (1976). Statistical power in criterion related validation studies. Journal of Applied Psychology, 61, 473-485.

Schmidt, F.L., Mack, M.J., & Hunter, J.E. (1984). Selection utility in the occupation of U. S. park ranger for three modes of test use. Journal of Applied Psychology, 69, 490-497.

Schmidt, F.L., Ocasio, B.P., Hillary, J.M. & Hunter, J.E. (1985) Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. Personnel Psychology, 38, 509-524.

Schmidt, F.L., Pearlman, K., & Hunter, J.E. (1981). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 33, 705-724.

Schmidt, F.L., Pearlman, K., Hunter, J.E. & Hirsh, H.R. (1985) Forty questions and answers about validity generalization and meta-analysis. Personnel Psychology, 38, 697-798.

Schmitt, N., Gooding, R.Z., Noe, R.D., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Schneider, B. (1978) Person-situation selection: A review of some ability situation interaction research. Personnel Psychology, 31, 281-297.

Smith, P.C. (1976). Behaviors, results and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of individual and organizational psychology (pp. 745-775). Chicago: Rand-McNally.

Snedecor, G.W. & Cochran, W.G. (1980). Statistical methods. 7th ed. Ames Iowa: Iowa State University Press.

Stone, E.F. & Hollenbeck, J.R. (1984) Some issues associated with the use of moderated regression. Organizational Behavior and Human Performance, 34, 195-213.

Tannenbaum, A.S. (1968). Control in organizations. New York: McGraw-Hill.

Tenopyr, M.L., & Oeltjen, P.D. (1982). Personnel selection and classification. Annual Review of Psychology, 33, 581-618.

Theologus, G.C. & Fleishman, E.A. (1976). Validation study of ability scales for classifying human tasks. Washington, DC: American Institutes for Research, TR No. 5.

Thorndike, R.L. (1949). Personnel Selection. New York: Wiley.

Thorndike, R.L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8, 63-70.

Uniform guidelines on employee selection procedures Federal Register, August 25, 1978, 285-316.

U.S. Department of Labor. (1970) Manual for the U.S.E.S. General Aptitude Test Battery. Section III: Development. Washington, DC: U.S. Employment Service.

U.S. Department of Labor. (1977) Dictionary of Occupational Titles (4th ed.). Washington, DC: U.S. Government Printing Office

U.S. Department of Labor. (1983) The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance for the U.S. Employment Service. U.S.E.S. test research report no. 44. Washington, DC: U.S. Employment Service.

U.S. Department of Labor. (1983) 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. U.S.E.S. test research report no. 47.

U.S. Department of Labor. (1983) Fairness of the General Aptitude Test Battery: Ability differences and their impact on minority hiring rates. U.S.E.S. test research report no. 46. Washington, DC: U.S. Employment Service. Washington, DC: U.S. Employment Service.

U.S. Department of Labor. (1983) Overview of validity generalization for the U.S. Employment Service. U.S.E.S. test research report no. 43. Washington, DC: U.S. Employment Service.

U.S. Department of Labor. (1983) Test validation for 12,000 jobs: an application of job classification and validity generalization analysis to the General Aptitude Test Battery. U.S.E.S. test research report no. 45. Washington, DC: U.S. Employment Service.

Wallace, S. (1965). Criteria for what? American Psychologist, 20, 411-417.

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

Weiss, D.J. (1972). General Aptitude Test Battery. In O.K. Buros (Ed.), The seventh mental measurements yearbook (pp. 676-679). Highland Park, NJ: Gryphon Press.

Wernimont, P.F. & Campbell, J. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

Zedeck, S. (1971) Problems with the use of "moderator" variables. Psychological Bulletin, 76 295-310.

Appendix A. Descriptive Rating Scale

Suggestions to Raters

We are asking you to rate the job performance of the people who work for you. These ratings will serve as a "yardstick" against which we can compare the test scores in this study. The ratings must give a true picture of each worker or this study will have very little value. You should try to give the most accurate ratings possible for each worker.

These ratings are strictly confidential and won't affect your workers in any way. Neither the ratings nor test scores of any workers will be shown to anyone in your company. We are interested only in "testing the tests." Ratings are needed only for those workers who are in the test study.

Workers who have not completed their training period, or who have not been on the job or under your supervision long enough for you to know how well they can perform this work should not be rated. Please inform the test technician about this if you are asked to rate any such workers.

In making ratings, don't let general impressions or some outstanding trait affect your judgement. Try to forget your personal feelings about the worker. Rate only on the work performed. Here are some more ideas which might help you:

- Please read all directions and the rating scale thoroughly before rating.

- For each question compare your workers with "workers in general" in this job. That is, compare your workers with other workers on this job that you have known. This is very important in small plants where there are only a few workers. We want the ratings to be based on the same standard in all locations.

- A suggested method is to rate all workers on one question at a time. The questions ask about different abilities of the workers. A worker may be good in one ability and poor in another: for example, a very slow worker may be accurate. So rate all workers on the first question, then rate all workers on the second question, and so on.

- Practice and experience usually improve a worker's skill. However, one worker with six months' experience may be a better worker than another with six years' experience. Don't rate one worker as poorer than another merely because of a lesser amount of experience.

- Rate the workers according to the work they have done over a period of several weeks or months. Don't rate just on the basis of one "good" day, or one "bad" day or some single incident. Think in terms of each worker's usual or typical performance.

- Rate only the abilities listed on the rating sheet. Do not let factors such as cooperativeness, ability to get along with others, promptness and honesty influence your ratings. Although these aspects are important, they are of no value for this study as a "yardstick" against which to compare aptitude test scores.

Questions

A. How often do you see this worker in a work situation?

- All the time.
- Several times a day.
- Several times a week.
- Seldom

B. How long have you worked with this worker?

- Under one month
- One to two months.
- Three to five months.
- Six months or more.

C. How much can this worker get done? (Worker's ability to make efficient use of time and to work at high speed.) (If it is possible to rate only the quantity of work which a person can do no this job as adequate or inadequate, use No. 2 to indicate "inadequate" and No. 4 to indicate "adequate.")

- Capable of very low work output. Can perform only at an unsatisfactory pace.
- Capable of low work output. Can perform at a slow pace.

- Capable of fair work output. Can perform at an acceptable pace.

- Capable of high work output. Can perform at a fast pace.

- Capable of very high work output. Can perform at an unusually fast pace.

D. How good is the quality of work? (Worker's ability to do high-grade work which meets quality standards.)

- Performance is inferior and almost never meets minimum quality standards.

- Performance is usually acceptable but somewhat inferior in quality.

- Performance is acceptable but usually not superior in quality

- Performance is usually superior in quality.

- Performance is almost always of the highest quality.

E. How accurate is the work? (Worker's ability to avoid making mistakes.)

- Makes very many mistakes. Work needs constant checking.

- Makes frequent mistakes. Work needs more checking than is desirable.

- Makes mistakes occasionally. Work needs only normal checking.

- Makes few mistakes. Work seldom needs checking.

- Rarely makes a mistake. Work almost never needs checking.

F. How much does the worker know about the job? (Worker's understanding of the principles, equipment, materials and methods that have to do directly or indirectly with the work.)

- Has very limited knowledge. Does not know enough to do the job adequately.
- Has little knowledge. Knows enough to get by.
- Has moderate amount of knowledge. Knows enough to do fair work.
- Has broad knowledge. Knows enough to do good work.
- Has complete knowledge. Knows the job thoroughly.

G. How large a variety of job duties can the worker perform efficiently? (Worker's ability to handle several different operations.)

- Cannot perform different operations adequately.
- Can perform a limited number of different operations efficiently.
- Can perform several different operations with reasonable efficiency.
- Can perform many different operation efficiently.
- Can perform an unusually large variety of different operations efficiently.

H. Considering all the factors already rated, and only these factors, how good is this worker? (Worker's all-around ability to do the job.)

- Performance usually not acceptable.
- Performance somewhat inferior.

- A fairly proficient worker.
- Performance usually superior.
- An unusually competent worker.

Appendix B. Predictor Intercorrelations

	1	2	3	4	5	6	7	8	9
1. G									
2. V	73*								
3. N	70*	58*							
4. S	64*	40*	43*						
5. P	39*	23*	28*	44*					
6. Q	41*	35*	34*	34*	52*				
7. K	15*	11**	10**	14*	27*	30*			
8. F	24*	16*	03	34*	43*	30*	31*		
9. M	14*	04	04	16*	30*	16*	35*	53*	
10. GVN	92*	87*	86*	56*	34*	42*	14*	16*	08
11. SPQ	62*	41*	44*	77*	83*	76*	30*	46*	27*
12. KFM	23*	13**	07	28*	44*	32*	67*	81*	83*
13. JF4	69*	70*	69*	51*	36*	39*	37*	38*	31*
14. JFR	88*	80*	77*	59*	46*	48*	38*	46*	40*
15. COG	75*	62*	64*	39*	26*	31*	13**	12**	05
16. LANG	63*	58*	48*	31*	18*	25*	12**	11	00
17. QUANT	73*	50*	73*	42*	31*	32*	12**	11**	12**
18. CLER	45*	35*	30*	33*	51*	48*	31*	41*	32*
19. DEX	12**	09	01	12**	28*	20*	45*	34*	36*

	10	11	12	13	14	15	16	17	18
10. GVN									
11. SPQ	56*								
12. KFM	16*	44*							
13. JF4	79*	53*	45*						
14. JFR	92*	65*	53*	85*					

15. COG	74*	41*	13**	55*	69*				
16. LANG	62*	31*	09	47*	57*	94*			
17. QUANT	73*	44*	15*	53*	69*	81*	56*		
18. CLER	41*	55*	45*	37*	52*	45*	41*	38*	
19. DEX	08	25*	48*	24*	26*	08	07	08	39*

*p < .05 **p < .10

Appendix C. ABA TEST BATTERY EXAMPLES

Cognitive ability is a factor composed of two abilities, language and quantitative. Language ability is measured by two tests, spelling and vocabulary. Quantitative ability is measured by two tests, arithmetic and mathematical reasoning. Instructions to and sample questions from those four tests follow.

I. LANGUAGE ABILITY

SPELLING - A list of words, some spelled incorrectly, is given. You are to write in the CORRECT spelling for any misspelled word or put a check if a word is correctly spelled. (40 in 8 minutes).

Samples: 1. invoice _____

2. morgage _____

3. thier _____

4. personell _____

5. medecine _____

VOCABULARY - Each word on the list is followed by five (5) words, only one of which means the same as the word on the left. Mark a check after that one word. (50 in 6 minutes).

Sample:

neglect - blame mind punish attend disregard

compile - assemble calculate render run declare

remit - accuse advise intensify send capture

II. QUANTITATIVE ABILITY

ARITHMETIC - (25 problems in 15 minutes)

Sample: 10

5

3

--

ans.

MATHEMATICAL REASONING - (10 word problems in 15 minutes)

Sample: If 10 percent interest is charged on a loan, the interest on a $2500 loan would be: _____ ans.

III. CLERICAL ABILITY

Clerical ability is measured by two tests, name (73 in 5 minutes) and number matching (100 in 5 minutes). Instructions and examples from the tests follow.

NUMBER and NAME MATCHING - on the next pages are lists of pairs of numbers and words or names. If the numbers or groups of words in a pair are exactly the same, put an X in the box to the right labeled "SAME." If they are not exactly the same, encircle the part of the right-hand column that differs from the left-hand column.

Samples: SAME

1. 16.423 ------- 16.423 __
2. $197.27 ------ $197.72 __
3. William Cavendish ----William R. Cavendish __

Dexterity is measured by one paper and pencil test. The instructions for that test follow.

IV. MOTOR COORDINATION

The ABA parallel to the GATB motor coordination (K) test is called Dexterity (DEX). The instructions for that test follow.

DEXTERITY - On the next page are rows of circles. Inside each circle draw a small triangle. Work as rapidly as you can, beginning with the top row, filling in each row before you go on to the next. (There are 20 rows of 16 circles for a total of 320 possible in five minutes.)

The vita has been removed from the scanned document