2002

Comparing job component validity to observed validity across jobs

David Charles Morris

COMPARING JOB COMPONENT VALIDITY TO OBSERVED VALIDITY ACROSS JOBS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Psychology:

Industrial/Organizational

---

by

David Charles Morris

June 2002

COMPARING JOB COMPONENT VALIDITY TO OBSERVED

VALIDITY ACROSS JOBS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

---

by

David Charles Morris

June 2002

Approved by:

[redacted]

Kenneth S. Shultz, Chair, Psychology

2/22/02

Date

[redacted]

Matt L. Riggs, Loma Linda University

[redacted]

Calvin C. Hoffman, Alliant University

ABSTRACT

Five-hundred and eighteen observed validity coefficients based on correlations between commercially available test data and supervisory ratings of overall job performance were collected on 89 different job titles. Using Dictionary of Occupational Title Codes, Job Component Validity (JCV) estimates based on similar job titles residing in the PAQ Services database were collected and averaged across the General Aptitude Test Battery test constructs (G, V, N, S, P, Q). A bare bones meta-analysis was conducted on observed studies by test construct and 95% CI were calculated. Corresponding averaged JCV estimates were then compared to the 95% CI's for each test construct. Averaged JCV estimates fell within the 95% CI for each test construct except "G". A second study calculated JCV battery validity estimates for a cognitive (G, V, N) and perceptual (S, P, Q) test-battery. Results indicated an increase in validity for both batteries and serves as an alternative to relying on the highest, single JCV estimate as the best estimate of the observed battery validity in practical settings.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

The Industrial/Organizational Psychologist (TIP) published an interview asking prominent researchers in the field of Industrial/Organizational Psychology the question, "What are the most important issues facing organizations and their people that need to be addressed" (Church 1998)? Among the several areas mentioned, a prominent theme was the changing structure and design of work.

Cascio (1995) brought attention to the evolving structure of business and how the individual job is affected. For example, the shift from a manufacturing to a customer service driven economy has led to a decrease in the number of jobs existing independently within an organization. It is becoming more common to see jobs function interdependently as a result of intact work teams assembled for the purposes of specific projects and then disbanding when the project is finished. Cascio (1995) recommended that present efforts in constructing valid selection procedures move beyond the use of job-based predictors because of the constantly changing nature of the work.

This trend has created increased demands on those involved with personnel selection. Kevin Murphy noted in his interview in TIP, "As jobs become more fluid and ambiguous, the idea of tailoring selection systems to the specific content of the job will become less useful" (Church, 1998, p. 96). He continued by stressing the importance that we [those in the personnel selection field] must focus less on the specific content ingrained in the individual job, and focus more heavily on the overarching constructs required to perform these jobs.

This universalistic view has not always been supported. Early test validation research supported the notion that jobs, while similar in nature, were quite different due to extraneous influences. These extraneous contextual and environmental factors were thought to be responsible for the variation in observed validities between jobs.

Situational Specificity

In 1966, Edwin Ghiselli wrote an article on the generalization of validity. His article reported the validity of commonly administered tests used for personnel selection. To his surprise, he noticed a large amount of variability in observed validity coefficients among jobs thought to be similar in nature. Ghiselli noted that although it is not expected that validity studies will produce the exact same results, it is expected they be similar. After assessing numerous validation studies covering a wide variety of jobs, he concluded that they [validity coefficients] were worlds apart.

This variation in validity coefficients led to the belief that there were subtle but important differences between seemingly identical jobs. Differences or "moderators" were thought to vary from organization to organization. Factors such as organizational climate, management philosophy, and reward structure were considered unique from one setting to the next. The inability to detect such differences, using the current methods for studying jobs, constrained practicality from generalizing results from one setting to another, and forces practitioners to conduct an empirical study in each setting. In 1976, Guion commented on the inability to solve the problem of validity generalization, "The inability to generalize across studies prevents the development of general principles in personnel selection as well as taking the field of Industrial/Organizational Psychology from a mere technology to a science" (Pearlman, Schmidt, & Hunter, 1980, p. 375).

Validity Generalization

As early as 1952, Lawshe (p. 31) suggested that job-test-criterion relationships were generalizable. If mean, uncorrected validity coefficients were .40 or more, then the chance of finding a valid correlation in a single study seemed good. However, the idea of validity generalization lay dormant through the mid-1970s until new statistical procedures were designed to measure the variation between studies (Guion, 1998, p. 368).

In 1976, Schmidt, Hunter, and Urry suggested that most of the variation found between validity studies was due to statistical artifacts. They demonstrated that low power studies (small N, typical in local validation studies) when corrected for sampling error, accounted for approximately 75% of the variance among studies. According to Schmidt et al. (1976) proponents of situational specificity falsely believed statistical significance tests controlled for sampling error. They did not realize that sampling error alone causes wide variations in observed validity coefficients, even if conducted within the same setting.

Two studies reported in Schmidt, Law, Hunter, Rothstein, Pearlman, and McDaniel (1993) demonstrated the effect sampling error has on small sample studies (Schmidt

& Hunter, 1983; Schmidt, Ocasio, Hillary, & Hunter, 1985). They found that observed validities from studies within the same settings varied to the same degree and magnitude as did validities from studies collected across settings. When corrected for sampling error alone, most if not all the variance was accounted for.

The role of statistical significance tests in controlling for sampling error is still widely misunderstood today. In the section that follows, weaknesses in significance testing will be discussed and an alternative method to significance testing which avoids such pitfalls will be described.

Problems with Statistical Significance Tests

In their book, Hunter and Schmidt (1990, pp. 23-27) provide an excellent overview on the properties of statistical significance testing. In a Monte Carlo simulation, 30 correlations measuring the same constructs were compared. Testing each coefficient for statistical significance, only 19 of the 30 correlations were found to be significant. These results, typical of what is found in actual studies, often lead to the conclusion in traditional review articles that more research is needed. The search for potential "moderators" is recommended to understand why significance is found in certain situations and not in others.

In the above example, Hunter and Schmidt (1990, p. 28) point out that each of the 30 studies was based on the same population correlation of .33. Random sample sizes, centering around 40, were generated for each study. Depending on a study's sample size, correlations would depart in varying degrees from the population value of .33.

In their example, the largest and smallest correlations came from studies with very small sample sizes. Studies based on larger sample sizes tended to center more closely around the population correlation. If all of the studies were based on the same population correlation, why did only 19 of the 30 studies result in a significant finding?

Hunter and Schmidt (1990, p. 29) report that the significance test was developed in response to the problem of sampling error. A common misunderstanding is that the test guarantees an error rate of 5% or less. An error rate of 5% or less is guaranteed only when the null hypothesis is true. Given the fact that the null hypothesis is true and our results lead us to reject it, then we have committed a Type I error. The significance test is designed to control for this error. That is, when the null hypothesis is true, significance testing limits falsely rejecting it to 1 in 20 studies. However, if the null hypothesis is false, and our results lead us to accept it (Type II error), the error rate may be as high as 95%.

The problem with significance testing is knowing when the null hypothesis is true or false for a given study. If it is not known, how can a researcher be sure his or her error rate is 5% or some value as high as 95%? Hunter and Schmidt (1990, p. 31) suggest there is only one-way in guaranteeing an error rate of 5%: Abandon the significance test and use confidence intervals.

Confidence Intervals

Confidence intervals are more appropriate than significance testing for two reasons: (a) The interval is correctly centered on the observed value rather than on the hypothetical value of the null hypothesis; and (b) It paints a more accurate picture of the uncertainty of small sample size studies (Hunter & Schmidt, 1990, p. 32).

In their Monte Carlo simulation, Hunter and Schmidt (1990, p. 32) provided two examples that demonstrated the first reason. First, they compared two studies with the same correlation coefficient, but with different sample sizes. One study found a significant relationship while the other did not. Second, they compared two studies with the same sample sizes, but with different correlation coefficients. Again, one study found a significant relationship while the other did not. Relying on significance testing, conflicting results indicated that the relationship between the hypothesized variables were significant in some settings, but not in others.

When the same studies reported confidence intervals, the intervals of all four studies centered around the true population value of .33. Hunter and Schmidt (1990, p. 32) point out that the use of confidence intervals did not contradict the results from using significance tests. In the two studies where significance was not obtained, this was also indicated by the confidence intervals where the range included $\rho = 0$, or the null hypothesis. However, using the confidence interval also showed the overlap in values shared between studies. This overlap also included the true population value of .33.

Hunter and Schmidt (1990, pp. 32-33) commented on the large range of values the confidence intervals reported for a particular study. Sometimes the range of values would span a 50-point spread. This sense of uncertainty in knowing what the true population value is, demonstrates the second reason why confidence intervals are more appropriate.

Suppose a researcher was interested in establishing a confidence interval to have the width of $\pm$ .05 around the population value of .33 (i.e., conducting a precision analysis). Hunter and Schmidt (1990, p. 33) computed the minimum sample size to be approximately 1,538. This minimum sample size of 1,538 demonstrates the point that Schmidt, Hunter, and Urry (1976) argued, and that proponents of situational specificity misunderstood. Significance testing leads to the misunderstanding that sample sizes derived from local settings are large enough for reliable and valid empirical studies. Schmidt, Hunter, and Urry (1976) point out that because of range restriction and criterion unreliability, sample sizes necessary to provide adequate statistical power for individual studies quickly become unfeasible for local validation.

If using confidence intervals demonstrates the uncertainty produced by small sample studies, then researchers have two choices: (a) Conduct large-sample single studies; or (b) Combine results across many small-sample studies. Hunter and Schmidt point out that given the limited resources available to practitioners in local settings, the only possible option is to combine results across many studies, hence an introduction to meta-analysis.

Meta-Analysis

If the confidence interval is the solution to statistical significance testing at the single study level, then meta-analysis is the solution to traditional review methods for comparing results across studies (Hunter & Schmidt, 1990, p. 31). By the mid-to-late 1970s, cynicism grew to a peak regarding the inability of the social and behavioral sciences to provide definitive answers to pressing issues. Funding sources were being drastically cut, and the public as well researchers themselves started to question whether the field was capable of generating definitive solutions (Hunter & Schmidt, 1990, p. 37).

In an attempt to explain why the social and behavioral sciences were faced with this dilemma, Hunter and Schmidt (1990, pp. 36-37) provide a sequence of events that research in a new area typically follows. First, there are a number of questions that social and behavioral scientists set out to answer. Large numbers of primary studies are conducted, hypotheses are tested, and results are reported. Using traditional review methods, results are compared and conflicting results are found. A second phase of research is initiated aimed to study the causes for such differences (a.k.a., search for moderators). Traditional review methods again compare the results from these second phase studies and again, more conflicting results are reported. Hunter and Schmidt (1990, p. 37) make the point that at some time, the need is not for more primary research articles, but for some means of making sense out of what has already been accumulated. Methods of meta-analysis were designed to answer such questions.

A meta-analysis examines independent research studies based on the same or similar hypothesis. Each study may be considered a replication of the others. Although they may use different measures under different conditions, they are nevertheless concerned about the underlying relationship between the same constructs (Guion, 1998, p. 373).

In what is referred to as a "bare bones" meta-analysis, where the only correction made is for sampling error, each independent study is weighted by its sample size. Guion (1998, p. 374) reported this as the most accurate estimate of the population value. Studies based on larger sample sizes are more reliable estimates of the population correlation and are thus given more weight. Once the average correlation value is computed, the variance across studies is determined by taking the average squared deviations of sample correlations from the mean and weighting the squared deviations by their sample size. The question now becomes when the studies are corrected for sampling error, how much variance is left over among studies? As a rule of thumb, Hunter and Schmidt (1990) indicate if sampling error alone accounts for $\geq$ 75% of the variance, the remaining variance consists of trivial differences (which can usually be accounted for by other statistical corrections such as range restriction and attenuation) and the results can be justified as being generalizable.

To date, several hundred meta-analyses have been conducted in the field of the social and behavioral sciences. In the field of Industrial/Organizational Psychology, research spanning 85 years has been accumulated and summarized using meta-analysis in the area of personnel selection (Schmidt & Hunter, 1998). As reported in Guion (1998, p. 376) Schmidt and Hunter (1981, p. 1128) said, "Professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs."

Although there is an overwhelming amount of evidence to support the generalization of test validities, and that the test validities are not situationally specific, the processes governing the transition to a new way of thinking often lag behind the development of new technology.

Uniform Guidelines on Employee Selection Procedures

In 1978, the Uniform Guidelines on Employee Selection Procedures were developed to provide employers with a uniform set of principles. These procedures were to serve as Guidelines for test use and other selection procedures and as a basis for employment decisions. In the scope of the present study, only the standards set forth by the guidelines regarding the validation of selection procedures will be discussed.

The Guidelines report three acceptable types of validity studies: (a) Criterion-related; (b) Content-related; and (c) Construct-related. In all three types of studies, the Guidelines strongly suggest that validity should be based on information about the job gathered from a job analysis. Any method of job analysis may be applied so long as it provides the information required for the specific validation strategy used. All three-validation types call for generally the same type of information. Validation should be based on important and observable work behaviors required for successful performance on the job.

In practice, mostly criterion-related validity methods are used for employee selection. Regarding content-validation, the Guidelines restrict the use of selection tests to those measuring the knowledge, skills, and abilities found necessary for successful performance on the job. The Guidelines specifically state that selection procedures based upon inferences about mental processes cannot be supported solely or primarily on the basis of content validity. Content-validation is not appropriate for demonstrating the validity of selection procedures which purport to measure traits or constructs, such as intelligence, aptitude, personality, commonsense, judgment, leadership, and spatial ability. Therefore, a content-validation strategy would be more typically used in conjunction with criterion-validation to serve as a supplement in the form of job-knowledge and job-sample tests. In regard to construct-validation, as stated in the Guidelines, it was then seen as a relatively new and developing procedure that presently lacked enough research to support its use in employment settings.

The Guidelines were developed in 1978, around the same time research on validity generalization (VG) began to surface in the professional literature. Therefore, the Guidelines do not specifically describe VG as an option for supporting test use. However, they state that the Guidelines are not intended to preclude the development and use of other professionally acceptable techniques with respect to the validation of selection procedures and that new strategies will be evaluated as they become accepted by the psychological profession.

Technical standards are given for each of the three types of validation strategies. Depending on the validation strategy followed, slightly different information is needed. The next few paragraphs will discuss how VG meets and exceeds most of the rules set forth by the Guidelines with regard to the technical standards set forth for criterion-related validity studies.

In the technical standards for criterion-related validity, the technical feasibility of the study is initially addressed. The first step is to determine whether or not the appropriate sample size can be collected in a given employment situation in order to provide a meaningful study. The Guidelines specifically state that in situations where jobs substantially share the same major work behaviors, those jobs may be grouped together in order to obtain adequate sample sizes. This is exactly what the VG literature has proposed. Where the Guidelines and VG differ is in what the appropriate statistic should be when studying the relationships between predictors of job performance.

The Guidelines state significance testing as the professionally accepted method for studying the relationships between variables. In their description of the power of significance testing, the Guidelines make the same assumption about control for Type I and Type II error that proponents of the situational specificity hypothesis make.

In its description of the operational use of selection procedures, the Guidelines state that other factors remaining the same, the greater the magnitude of the relationship (e.g., correlation coefficient) between performance on a selection procedure and one or more criteria of performance on the job, the more likely it is that it will be appropriate for the given employment situation. What the Guidelines do not address is how to appropriately study the magnitude of the relationship. The VG literature has shown that significance testing is not the most appropriate method for doing so and recommends replacing it with confidence intervals. Reporting confidence intervals would demonstrate the lack of statistical power and precision inherent among most employment situations and would further support the aggregation of similar jobs across employment settings.

Up to this point, the VG literature seems to meet and in some cases exceed the expectations set forth by the Guidelines. If this is true, then why is it that the most common methods of conducting validation studies still rely on reporting findings via statistical significance testing? One could make the argument that the formal quantitative training found among programs in the social and behavioral sciences simply are not teaching these concepts to their students. Included within this argument is that this lack of understanding forces those validating selection procedures to rely on the Guidelines verbatim. That is, if it is not specifically stated in the Guidelines, then the assumption is made that it not allowed, or worse yet, not possible.

Landy (1986) proposed that practitioners not use the Guidelines as a checklist where they are constrained to fit validation research into one of the three validation boxes (criterion, content, construct) stated in the

Guidelines. Instead, he proposed that practitioners treat validation as a form of hypothesis testing where the collection of data stemming from multiple methods and sources be used to support inferences based on predictive hypotheses. This multiple-method approach used to gather converging evidence from multiple sources adds to the credibility and confidence in supporting the operational use of a selection system (Hoffman & McPhail, 1998).

As reported in Hoffman and McPhail (1998) the major advantage of using VG findings to support test use in a new setting is the fact that little job analysis information is needed and that it does not require any additional validation. All that is needed is enough information to be able to match a particular job to similar jobs that were used during the initial validation study (Pearlman, Schmidt, & Hunter, 1980).

From a litigation perspective, VG's greatest advantage in the sense of the limited amount of job analytic work needed also poses as its greatest disadvantage. When reviewing the amount of case law that has accumulated over the years, it becomes apparent that Judges tend to side in favor of the defendant when selection practices are based on thorough job analysis information (Guion 1998, p. 177).

Guion (1998) commented that the sole use of VG to support test use in this litigious environment is probably premature. At a minimum, efforts carried out by job analysis should include site visits, multiple interviews with incumbents and supervisors, the breakdown of jobs into major tasks or worker behaviors, and documentation of the findings in a technical report. The next few sections briefly introduce the topic of job analysis and describe how a particular method of job analysis meets and in some areas exceed the criteria set forth by the Guidelines.

Job Analysis

Levine (1983) named job analysis as the cornerstone to all human resource strategic planning and decision-making practices. Conducting job analyses prior to the design or implementation of a selection system is not only sound practice, but the legal ramifications of avoiding potential lawsuits tend to make the utility of job analysis that much more beneficial.

The term job analysis refers to a number of methods that are aimed at breaking jobs down into specific components, tasks, duties, activities, and other units of work (Levine 1983). In determining which job analysis method is best for a given situation, it is necessary to identify the relevant goals one wishes to attain and which methods will help facilitate the process.

One of the relevant goals in the present study is to help aid practitioners in supporting current selection systems without having to conduct local validation. This would fall under what Levine (1983) calls the human resource planning stage.

Human resource planning involves organizations trying to peer out into the future to see not only where they need to go, but also what it is going to take to get them there. What future demands will the external market place on them, and do they or will they have a workforce with the job requirements necessary to meet such environmental demands? Factors such as these must be taken into account when selecting the most appropriate job analytic method. Using the right tool for the right job will make the process easier and the results more applicable.

Position Analysis Questionnaire

The Position Analysis Questionnaire (PAQ) developed by Ernest McCormick and his associates at Purdue University, is a 187-item questionnaire that can be used to analyze virtually any job. The PAQ is a deductive approach which evaluates qualitative entities such as worker behaviors and measures them within a quantitative methodology (McCormick, Jeanneret, & Mecham, 1972).

Six divisions of the PAQ (Information Input, Mental Processes, Work Output, Relationships with Other Persons, Job Context, and Other Job Characteristics) provide the framework needed for the job analyst to capture every aspect of a particular job (McPhail, Jeanneret, McCormick, & Mecham, 1998). The PAQ is a worker-oriented approach which looks at the information received by the worker, the mental processes involved in responding to that stimuli, and the response or work output that is the final product. The environmental context of the job is also considered, recognizing that work does not exist in a vacuum and that outside forces will have an effect on the overall outcomes related to work.

Within the focus of the present study, the PAQ has been used to identify tests that would most likely be used for selecting employees for particular jobs, and to predict mean test scores and validity coefficients for those jobs. The theoretical assumption the PAQ makes is even though jobs may vary considerably in regards to the tasks and technological aspects when compared with one another, the general human behaviors needed to perform those jobs may be the same or highly similar (McCormick et al., 1972).

Jobs sharing general human behaviors as rated by the PAQ can be placed on a common metric. Comparing the job dimension scores of one job to other jobs in the PAQ data base enables the job analyst to place a particular job among others sharing similar characteristics. The communalities that jobs share on similar human behaviors support the assumption that the same constructs may predict performance across jobs. If jobs can be placed on a common metric and directly compared to one another, then a practitioner could conduct a job analysis (using the PAQ) and infer predicted mean test scores and validity coefficients based on similar jobs existing in the PAQ database (Jeanneret, 1992).

According to the standards set forth by the Guidelines for job analyses, the PAQ exceeds most of the requirements. Within the technical standards for construct validity studies, the Guidelines state that carrying out a construct valid approach is an extensive and arduous effort which usually involves a series of research studies compiled from a number of criterion-related and content validity studies. To date, the PAQ is the most heavily researched job analysis tool containing over 30,000 jobs in its database (Guion 1998, p. 82).

The PAQ meets the demands set forth by the Guidelines with respect to identifying the constructs believed to underlie successful performance on the job. Each construct is named and defined as the Guidelines suggest distinguishing them from among one another. The Guidelines also suggest if groups of jobs are being studied, analysis at the group level identifying similar work behaviors at varying levels of complexity needs to be conducted. PAQ's statistical software was specifically designed to perform such operations.

Mentioned earlier, the PAQ has been used to identify tests that would most likely be used to select employees for particular jobs as well as predict mean test scores and validity coefficients for those jobs. The original work was conducted on the General Aptitude Test Battery (Marquardt & McCormick, 1974; McCormick, Mecham, & Jeanneret, 1977; McCormick, Mecham, & Jeanneret, 1989; Mecham & McCormick, 1969) followed by later research conducted on construct equivalent commercially available tests (McCormick, DeNisi, & Shaw, 1979). The following section provides a detailed description regarding the development and factor structure of the General Aptitude Test Battery.

General Aptitude Test Battery

The United States Employment Services (USES) developed the General Aptitude Test Battery (GATB) and first put it to use in 1947 (Cohen, Swerdlik, & Phillips, 1996). It is used as a tool to identify aptitudes required for performance in a broad range of occupations. The GATB consists of 12 tests measuring nine aptitudes that can be further divided into three composite scores measuring cognitive, perceptual, and psychomotor aptitudes (see Appendix A for a visual description). The nine aptitudes are commonly seen in the literature as: G-General Learning Ability; V-Verbal Aptitude; N-Numerical Aptitude; S-Spatial Ability; P-Form Perception; Q-Clerical Perception; K-Motor Coordination; F-Finger Dexterity; M-Manual Dexterity. With respect to the present study, only the cognitive (G, V, N) and perceptual (S, P, Q) components will be discussed.

The cognitive component of the battery is comprised of G, V, and N. General Learning Ability (G) is measured by three tests: Three-dimensional space, vocabulary; and arithmetic reasoning. Verbal Aptitude (V) is measured with one test which is the same vocabulary test used to measure (G). Numerical Aptitude (N) is measured with two tests. The first is the same arithmetic reasoning test used for (G) while the other is a computation test.

The perceptual component of the battery is comprised of S, P, and Q. Spatial Aptitude is measured using the same three-dimensional space test used for (G). Form Perception (P) is measured using a test of tool matching and one of form matching. Finally, Clerical Perception (Q) is measured by only one test of name comparison.

It is obvious that in both the cognitive and perceptual composites, some of the same tests are used. In the cognitive composite, (G) is a combination of three-dimensional space, vocabulary, and arithmetic reasoning. These tests are found in the verbal, numerical, and spatial categories. In the perceptual composite, the same three-dimensional space test is used. Questions arise concerning how much overlap these tests have. That is, in regards to using certain tests to operationalize underlying constructs (in this case, cognitive and perceptual aptitude) classical test theory supports the claim that predictors should be as independent as possible in the sense that they contribute to the explanation of unique variance over and above other predictors. With identical tests serving as predictors for superficially discrete composites, it is probable that these dimensions are highly intercorrelated to the extent that they do not cover the construct domains as clearly as it may first appear.

Appendix B illustrates the high test intercorrelations among the GATB constructs (Hartigan & Wigdor, 1989). Referring to the second column labeled "G", notice the high intercorrelations among V, N, and S (.84, .86, and .74, respectively). These are markedly higher than other GATB constructs that do not share identical tests.

Nevertheless, the GATB has been used extensively for the purposes of developing test batteries. In fact, it laid the groundwork for the developers of the PAQ in establishing a database that would serve the basis for their job component validity (JCV) model. Such a model could then be used for validity generalization purposes in transporting test validities across situations.

Job Component Validity

The Job Component Validity model is inherent in the PAQ. It was an expansion of Lawshe's idea of synthetic validity where one could: "infer test-battery validity from predetermined validities of the tests for basic work components" (Mossholder & Arvey, 1984, p. 323).

Jeanneret (1992) compared several alternate methods of synthetic validity. Three characteristics present in all of them were: (a) the use of job analysis to discover and systematically document important work components, (b) establishing the relationship of the test with work components, and (c) forming test-batteries using component validity information from the jobs in question.

The JCV model relies on PAQ dimension scores to predict mean test scores and validity coefficients for cognitive ability constructs measured by the General Aptitude Test Battery (GATB). Four early studies (Marquardt & McCormick, 1974; McCormick, Mecham, & Jeanneret, 1977; McCormick, Mecham, & Jeanneret, 1989; Mecham & McCormick, 1969) have examined the ability of the PAQ's dimension scores to predict mean GATB scores and validities for a wide range of jobs. In those studies, job dimension scores derived from the PAQ served as independent variables, and GATB mean test scores and validity coefficients served as dependent variables in multiple regression analyses.

Job dimensions were developed using principal components analysis carried out on individual PAQ items to identify underlying dimensions that characterize the structure of jobs (McCormick, DeNisi, & Shaw, 1979). Results produced 32 divisional job dimensions across the six PAQ job dimensions (information input, mental process, work output, relationships with other persons, job context, and other job characteristics).

The initial JCV study was conducted with data on 90 different jobs with sample sizes ranging from 90 to 460 (Mecham & McCormick, 1969). Job dimension scores were entered into a separate stepwise multiple-regression analyses that predicted the mean test scores and validity coefficients previously obtained by the U.S. Employment Service for the nine GATB tests. Across the four initial studies, mean test scores were better predicted (median $R = .69$) than validities (median $R = .26$) and cognitive aptitudes (.29-.41) were better predicted than perceptual abilities (.19-.38), followed by psychomotor abilities (.20-.33)(Jeanneret, 1992).

McCormick, DeNisi, and Shaw (1979) expanded the JCV model, applying the regression equations originally developed to predict mean GATB test scores, to predict performance on commercially available tests. They found that when they plugged in commercial test data, the end result was that it highly correlated with the predicted scores obtained from the GATB.

Sackett (1991) commented on the JCV model's inability to predict observed validity coefficients as well as it predicts mean test scores. Sackett concluded that further research was needed in examining JCV’s predicted and observed validities before researcher's could feel confident in relying on JCV as a useful validation method.

It is not surprising that the JCV model better predicts mean test scores than validity coefficients. As reported in Hoffman and McPhail (1998), means are more stable point estimates than correlation coefficients. Correlations are based on a bivariate, rather than a univariate distribution, and are subject to the well-recognized artifacts outlined in VG studies. In addition, initial formation of the regression equations used in the JCV model predate the VG literature. Sample sizes used fell closer to the 90 than the 460 range. As a result, sampling error would be expected to be quite large, resulting in a large range of predicted validities.

Although the JCV’s ability to accurately predict validity coefficients may be restricted by commonly encountered artifacts (e.g., range restriction, sampling error, attenuation), such underestimates could still give practitioners some confidence in knowing that their observed validities would most likely be larger. McPhail (1995) reported three criterion related validation studies where both published and custom developed predictors were used. Jobs were analyzed with the PAQ and JCV predictions were obtained and compared to observed validities (See Appendix C). McPhail's two major conclusions were: (a) "despite the regression equation being relatively weak, the resulting predictions are nonetheless quite consistent with empirical results" (p. 8) and (b) "it appears that in most cases, the JCV predictions for validity underestimate the empirically obtained results, especially when the empirical results are based on measures that focus on more specific construct components" (p. 8).

Holden (1992) produced results similar to McPhail's (1995). Using data from a concurrent validation study, three similar jobs were combined to increase sample size and predicted validities were compared to observed validities using supervisory ratings as the criterion. Out of the four observed validity coefficients (See Appendix D) only (G) was lower than the JCV estimate. When using job knowledge and job sample criteria, Holden saw the correlations rise as much as three times the size of the analogous JCV estimates (See Hoffman & McPhail 1998 for a more complete review).

The previous two studies tend to support using the JCV procedure in determining the validity of selection measures without conducting local validation. Further research in this area would benefit by exploring methods which would close the gap between the JCV predictions and observed validities.

Hoffman and McPhail (1998) conducted a study comparing predicted JCV estimates for 51 clerical jobs with results from Pearlman, Schmidt, and Hunter's (1980) meta-analysis reporting mean observed validity coefficients for five DOT clerical categories. Hoffman and McPhail wanted to see how closely their JCV predictions mirrored the findings of Pearlman et al. (1980) study. A high correlation would provide support for using the JCV model to establish selection procedures for clerical jobs without having to conduct local validation.

Their results showed substantial similarity to the mean observed, uncorrected criterion related validity coefficients produced in Pearlman et al. (1980) study (See Appendix E). The overall correlation between predicted and observed validity estimates for all jobs was .97. Hoffman and McPhail (1998) attributed such a high correlation, when compared to McPhail's (1995) and Holden's (1992) study, to using average JCV estimates from a relatively large number of jobs. Just as Guion (1998) suggested, the authors found that averaging across a large sample of jobs provided more stable estimates, thus minimizing the effects of statistical artifacts normally encountered among single studies.

Purpose of Current Study

In the present study, we attempt to take current test validation strategies to the next step. It is based on the premise of the changing nature of work and the need to develop new methods designed to satisfy such demands. To Murphy's comment, "As jobs become more fluid and ambiguous, the idea of tailoring selection systems to the specific content of the job will become less useful" (as cited in Church, 1998, p. 96). Based on the VG literature regarding low power and small sample sizes, local validation alone does not meet these demands.

With all fairness, VG does not currently meet these demands either. Although the theory behind it along with the vast amounts of research to support it can be regarded as compelling, its lack of recognition in the professional guidelines and case law limits its impact of taking validation practices to the desired next level.

The JCV model on the other hand meets these demands. Not limited by the lack of job analysis information, a criticism of VG, practitioners can feel comfortable in instituting validation practices based on the JCV model knowing that it meets the requirements set forth by the Uniform Guidelines.

Early research on the JCV model, and its inability to produce validity coefficients similar to those found in local studies has kept it from gaining the recognition it deserves. Although studies such as McPhail (1995) and Holden (1992) have shown that predicted validities derived from the JCV model are at the least a conservative estimate of the actual true validity, critiques such as Sackett's (1991) limits the models acceptability and implementation.

These initial studies suffered the same consequences experienced in local validation research. Although their predicted validity coefficients were based on data derived from large PAQ databases, they were still trying to predict observed validity coefficients derived from small sample, low power local validation studies. In their example provided earlier, Schmidt and Hunter (1990) demonstrated the large variation observed in validity coefficients caused by small-sample sizes.

Hoffman and McPhail (1998) overcame this weakness by following what the VG literature voiced: (a) Conduct large-sample single studies; or (b) Combine results across many small-sample studies. In their research, they compared JCV estimates from 51 clerical jobs (a relatively large sample) to observed validity coefficients reported in Pearlman et al's. (1980) meta-analysis. Using a much larger sample of data, Hoffman and McPhail (1998) were able to demonstrate the high accuracy the JCV model is capable of in predicting observed validity coefficients.

In the present study, we will extend Hoffman and McPhail's (1998) study to include a wider array of jobs. Validity coefficients portraying the relationship between commercially available aptitude tests and supervisory ratings of overall job performance will be collected. Based on Guion's recommendations, a bare-bones meta-analysis will be conducted where correlations from individual studies will be weighted by their sample size and averaged to compute mean validity coefficients for each GATB test construct (G, V, N, S, P, Q).

Unlike past JCV research, the present study will follow Schmidt and Hunter's (1990) recommendations and compute confidence intervals around each observed mean validity coefficient. The confidence interval, and not the single correlation value, will be used to measure the similarity between mean observed validity coefficients and averaged JCV estimates collected from PAQ Services on similar jobs. It is assumed if the averaged JCV estimates fall within the confidence intervals of the observed validity coefficients, they can be treated equally. Such a study will counter criticism of JCV's inability to accurately predict observed validity coefficients and enable practitioner's to rely on this method in lien of local validation to support test use.

We also extend JCV research into another dimension. Currently, the JCV method reports single, univariate validity estimates for each GATB test construct. In practice, it is unlikely that a practitioner would use only one predictor to select candidates for a specific job. Multiple predictors are often used in order to cover a larger portion of the overall selection criterion. Reported in Hoffman and McPhail's (1998) study, Ruch, Weiner, McKillip, and Dye (1985) produced higher observed validity coefficients when using a 4-test generic battery than using the best predictor alone. In addition, Murphy and Shiarella (1997) argued that multiple predictors combined into a battery are superior to the predictor in the battery with the highest validity. Such multivariate frameworks yield higher effect sizes than single predictors and as a result, produce higher statistical power while requiring smaller sample sizes.

The second part of this study will take the single, univariate GATB constructs predicted by the JCV model, and create estimated battery validity coefficients. Such a tool would provide practitioners with a more realistic estimate than relying on the highest, single univariate construct as a conservative estimate of the actual validity.

Mentioned earlier, the GATB uses identical tests to comprise individual test constructs. Such superficially discrete composites resulted in higher than expected test intercorrelations and limits the usefulness of adding another predictor to account for additional variance in the criterion. Because commercially available tests are not subject to the same limitations as the GATB, larger JCV battery coefficients are likely to result due to expected lower intercorrelations among test constructs. Therefore, an intercorrelation matrix based on commercially available test data will be created and used to compute JCV battery validity estimates. Such a tool would be useful to practitioners working in the field where they will be able to compute JCV battery validities for any combination of tests used in the targeted employment testing process.

CHAPTER TWO

METHODOLOGY

Selection of Studies

Bare Bones Meta-Analysis

The first part of this study required a large collection of observed validity coefficients from published and unpublished studies. The compilation of studies followed the procedures recommended in Pearlman et al. (1980) meta-analysis conducted on clerical jobs.

The goal was to compile a database of sufficient scope and size to permit a large-scale test of the current procedure. Two stages were undertaken: (a) the development of a classification and coding system that captured all of the potentially relevant data from published and unpublished validity reports; (b) an extensive search of published and unpublished validity studies and recording the information according to the coding system.

The search for published and unpublished studies looked to the following resources: (a) major commercial test manuals reporting validity information on tests; (b) contacting test publishers to obtain unpublished validity data; (c) contacting research groups and private consulting firms, and (d) tracing back primary studies used in other meta-analysis studies.

Job Component Validity Battery Validity

The second part of the study required the development of a commercially available test intercorrelation matrix. The matrix was created by identifying and averaging test intercorrelations from several major test publication manuals as well as raw employment testing data files collected by the author.

Decision Rules

Bare Bones Meta-Analysis

Several types of information for each validity study was coded and recorded into raw numeric form in a data set: (a) uncorrected correlation coefficient; (b) type of correlation coefficient; (c) sample size; (d) criterion measure used; (e) type of validation strategy employed; and (f) name of test used.

Data were collected only from studies that met the following requirements: (a) validity results in the form of a bivariate correlation coefficient (uncorrected for either attenuation or range restriction); (b) sufficient information that allowed the job to be appropriately classified by a Dictionary of Occupational Title code;

(c) sample size was reported; (d) there was sufficient information reported in order to classify the type of criterion measure used (supervisory ratings, production data, work samples).

Decision rules regarding what data to record when the validity for a particular study reports coefficients for two or more predictors in the same test type category, multiple criteria, and multiple subgroups followed Pearlman et al. (1980) recommendations. In studies where two or more predictors belong to the same test type category (e.g., several types of verbal tests), each coefficient was used. In studies using two or more criterion measures (e.g., supervisory ratings, training performance) each coefficient was used.

Job Component Validity Battery Validity

Individual tests were coded and assigned to one of the six GATB constructs used in the study (G, V, N, S, P, Q). Test intercorrelation values were then assigned to one of fifteen possible bivariate test combinations. Appendix F provides the raw values that were used to produce the commercially available test intercorrelation matrix.

Data Collection

Bare Bones Meta-Analysis

Participation letters were sent to 39 companies and test publishers. The letter identified the author, the people on the thesis committee, and the type of information being requested. A short summary of the study was attached for those interested in more detail (see Appendix G).

Eight companies/test publishers responded. Of those, six participated by sending data, however only four companies sent data that could be used in the current study.

Data were collected on 518 validity coefficients representing 89 unique job titles. The majority of data collected came from technical reports provided by test publishers (494 validity coefficients), the rest came from the individual participating organizations (24 validity coefficients).

The 89 job titles and corresponding DOT codes were sent to PAQ Services in Logan Utah to match up corresponding job titles and DOT codes existing in their databases. PAQ Services matched data on 54 of the 89 jobs. Appendix H shows the list of studies collected on observed validities while Appendix I shows the JCV estimates PAQ Services matched.

Data Analysis

Bare-Bones Meta-Analysis

Separate analyses were conducted for each GATB test construct (G, V, N, S, P, Q). Job titles containing one or more observed validity coefficient for a particular GATB test construct were sorted by DOT code. DOT codes were then used to match jobs to corresponding JCV estimates received from PAQ Services on similar jobs with identical DOT codes.

Job titles, sample sizes, and the observed correlation coefficients were entered into a program titled, "Meta-Win 16: Psychometric Meta-analysis Program". Standard output from the program included the mean observed validity coefficient weighted by sample size ($\bar{r}_w$), the number of studies that went into the analysis (k), and the total sample size (Σn). With this information, the program computes the total variance among the observed validity coefficients (s^2_r), the error variance (s^2_e), and the residual variance (s^2_p). This information was then used to compute the percent of total variance accounted for by sampling error (%Explained) and the 95% confidence interval for each GATB test construct (95% CI). Matching JCV estimates were simply averaged together to compute the predicted value to be compared to the observed 95% CI for each GATB test construct.

Job Component Validity Battery Validity

Tabachnick and Fidell, (1997, p. 141) provide matrix equations used to compute multiple R among several predictor variables (employment tests), and one criterion variable (ratings of overall job performance). Matrix calculations were performed using a statistical program called GANOVA. The first step in the process was to multiply the inverse of the test intercorrelation matrix to a column vector of corresponding Job Component Validity coefficients. Because multiplication by an inverse is the same as division, the column matrix of correlations between predictor and criterion variables is divided by the correlation matrix of predictor variables resulting in standardized regression coefficients. The standardized regression coefficients are then assembled into a column vector and multiplied by a row vector of corresponding Job Component Validity coefficients. The result is multiple R^2, when one takes the square root, this results in multiple R, or the JCV battery coefficient. Appendix J provides an example of how to create a JCV battery validity estimate using the matrix equations provided by Tabachnick and Fidell (1997, p. 141).

CHAPTER THREE

RESULTS

Bare Bones Meta-Analysis

Appendix K shows the results of the bare bones meta-analyses conducted on observed, commercially available aptitude tests across the six GATB test constructs.

The first analysis estimated the validity of general learning ability (G). The total sample size across 32 studies reporting observed correlations was 1,898. The proportion of variance explained due to sampling error was 75.97%. The average correlation weighted by sample size was .23 with a 95% CI ranging from .19 to .27. The averaged JCV estimate on matching jobs collected by PAQ Services (.29) fell outside the 95% CI of the observed validity coefficient.

The second analysis estimated the validity of verbal aptitude (V). The total sample size across 32 studies reporting observed correlations was 5,042. The proportion of variance explained due to sampling error was 83.01%. The average correlation weighted by sample size was .20 with a 95% CI ranging from .17 to .22. The averaged JCV estimate on matching jobs collected by PAQ Services (.22) fell within the 95% CI of the observed validity coefficient.

The third analysis estimated the validity of numerical aptitude (N). The total sample size across 72 studies reporting observed correlations was 6,780. The proportion of variance explained due to sampling error was 91.82%. The average correlation weighted by sample size was .24 with a 95% CI ranging from .22 to .26. The averaged JCV estimate on matching jobs collected by PAQ Services (.26) fell within the 95% CI of the observed validity coefficient.

The fourth analysis estimated the validity of spatial aptitude (S). The total sample size across 42 studies reporting observed correlations was 4,444. The proportion of variance explained due to sampling error was 88.88%. The average correlation weighted by sample size was .23 with a 95% CI ranging from .20 to .26. The averaged JCV estimate on matching jobs collected by PAQ Services (.20) fell within the 95% CI of the observed validity coefficient.

The fifth analysis estimated the validity of form perception (P). The total sample size across 7 studies reporting observed correlations was 703. The proportion of variance explained due to sampling error was 95.33%. The average correlation weighted by sample size was .27 with a 95% CI ranging from .20 to .34. The averaged JCV estimate on matching jobs collected by PAQ Services (.20) fell within the 95% CI of the observed validity coefficient.

The sixth analysis estimated the validity of clerical perception (Q). The total sample size across 28 studies reporting observed correlations was 2,145. The proportion of variance explained due to sampling error was 88.34%. The average correlation weighted by sample size was .24 with a 95% CI ranging from .20 to .28. The averaged JCV estimate on matching jobs collected by PAQ Services (.21) fell within the 95% CI of the observed validity coefficient.

Job Component Validity
Battery Validity

Appendix L shows the averaged test intercorrelation matrix based on commercially available tests. Not enough data was available to compute either an SQ or a PQ test intercorrelation. In order to compute JCV battery validity estimates using these test combinations, the corresponding test intercorrelations from the GATB (Hartigan & Wigdor, 1989) served as substitutes.

Appendix M shows the matrix computations worked out for a cognitive (G, V, N) and a perceptual (S, P, Q) JCV battery.

The inverse of the JCV cognitive battery was computed using the corresponding commercially available test intercorrelations (.34, .38, and .46), resulting in a 3 X 3 matrix and multiplied by a 3 X 1 column vector of averaged JCV estimates for G, V, and N (.29, .22, .26). This resulted in a 3 X 1 column vector of standardized regression coefficients which were multiplied by a 1 X 3 row vector of the corresponding averaged JCV estimates. This resulted in an R^2 of .12; the square root of this value produces a multiple R of .34.

The inverse of the JCV perceptual battery was computed using the corresponding commercially available test intercorrelations (.37, .39, and .65), resulting in a 3 X 3 matrix and multiplied by a 3 X 1 column vector of averaged JCV estimates for S, P, and Q (.20, .20, .21). This resulted in a 3 X 1 column vector of standardized regression coefficients which were multiplied by a 1 X 3 row vector of the corresponding averaged JCV estimates. This resulted in an R^2 of .06; the square root of this value produces a multiple R of .25.

CHAPTER FOUR

DISCUSSION

One of the concerns the present study addressed was the degree of accuracy with which JCV could be used to predict observed validity coefficients between commercially available aptitude tests and supervisory ratings of overall job performance for a wide range of jobs. Appendix K shows that for five of the six GATB test constructs (V, N, S, P, Q) averaged JCV estimates fell within the 95% CI of observed validities. Since at least 75% of the variation between studies could be attributed to sampling error, the averaged observed validities found in Appendix K can be regarded as accurate and stable estimates of the true population values.

Within the past few years, leaders in the field of personnel selection have stressed the importance of future test validation efforts, emphasizing that researchers should focus less on the specific content in individual jobs, and focus more heavily on the overarching constructs required to perform these jobs (Church, 1998). Earlier JCV research used PAQ dimension scores to predict mean GATB test scores and observed validity coefficients for a wide variety of jobs (Marquardt & McCormick, 1974; McCormick, Mecham, & Jeanneret, 1977; McCormick, Mecham, & Jeanneret, 1989; Mecham & McCormick, 1969). Later research expanded the JCV method to predict mean test scores for commercially available tests (McComick, DeNisi, & Shaw, 1979). Hoffman and McPhail (1998) extended the JCV research still further to predict observed validity coefficients for clerical jobs (using commercially available test data as the predictor variable).

The present study adds to the JCV research literature in several ways. First, based on Schmidt and Hunter's (1990) recommendations and the success of Hoffman and McPhail's (1998) research on clerical jobs, analyses were conducted using a relatively large number of individual studies based on a variety of jobs. Second, unlike prior JCV research, the present study relied on confidence intervals to assess the degree of similarity between observed and predicted JCV validity coefficients. By conducting a large sample study, and relying on confidence intervals that center around the observed values, this research overcame many of the limitations experienced by earlier small sample JCV (Holden, 1992; McPhail, 1995).

The present study adds another dimension to current JCV research. The Uniform Guidelines (1978) state that when designing a selection system, the greater the magnitude of the relationship between performance on a selection procedure and one or more criteria of performance on the job, the more likely the predictor will be appropriate for a given employment situation. Hoffman and McPhail (1998) commented that since most employment testing practices rely on multiple tests to screen candidates, relying on the single highest JCV predicted value is likely to be a conservative estimate of the overall battery validity.

Using the matrix equations provided by Tabachnick and Fidell (1997, p. 141) and the test intercorrelation matrix shown in Appendix J, JCV battery validity coefficients can be easily computed for any possible combination of commercially available tests. In the present study, a JCV cognitive battery (G, V, N) and perceptual battery (S, P, Q) resulted in multiple R's of .34 and .25, respectively. This resulted in an increase in validity when compared to the single highest JCV estimate of 15% and 16% ([.34 - .29]/.34 = .15%, and [.25 - .21]/.25 = 16%). Computing JCV battery estimates will result in higher effect sizes and add to the defensibility as well as the utility of the selection procedure in question.

Using the commercial test intercorrelations has other advantages as well. Computing JCV-battery estimates using

G, V, and N from commercially available tests (see Appendix L) compared to the GATB intercorrelations for the same constructs (see Appendix B) results in an increase in validity of 12% ([.34 - .30]/.34).

In practice, a researcher could compute his/her own intercorrelation matrix based on the actual tests used in the study, or rely on test publisher norms if appropriate. Then, the researcher could compute the matrix equations using the single JCV estimates from the PAQ job analysis to obtain the JCV-battery estimates on a job-by-job basis.

Limitations and Recommendations for Future Research

As with any meta-analysis, it is always a challenge to gather enough primary research to conduct a feasible study. The majority of data used in the present study came from technical manuals provided by test publishers. Although only 4 of the 39 companies that were contacted and asked to participate sent viable data, it is believed many more would have sent data if it were available.

Two of the six GATB test constructs in the present study (G and P) had far fewer studies and markedly smaller sample sizes compared to the other four (V, N, S, P). This was unfortunate, but may reflect what is actually being practiced in the field. Very few commercially available tests are available that are designed to solely measure "G". In most instances, such tests are usually a combination of verbal, numerical, and spatial components. Commercially available tests designed to measure form perception (P) are even rarer. In fact, in one of the several job analysis reports provided by the PAQ, it provides commercially available equivalent tests for all but the "P" GATB test constructs.

Nevertheless, the bare bones meta-analysis conducted on the G and P test constructs accounted for most of the variance between observed validities across studies (75.97% and 95.33%, respectively). The larger range of values shown by the 95% confidence intervals accurately reflect the effects of the smaller samples.

At first, the small number of available studies for "G" was thought to be responsible for the predicted JCV estimate of .29 to fall outside of the 95% confidence interval of the observed mean validity (.19 to .27). However, a closer look at the factor structure of the GATB, previously describe in detail, leads one to believe that the JCV estimate for "G" is not a single, univariate estimate, but a JCV battery estimate comprised of V, N, and S.

There is some evidence that this might be the case. In all of their studies (Hoffman & McPhail, 1998; Holden, 1992; McPhail, 1995) the JCV estimates for "G" were always higher than observed validities, while in most other cases, JCV estimates for the other GATB test constructs resulted in lower than observed validities.

Some concern to the degree of generalizability of the test intercorrelation matrix described in Appendix L should also be addressed. It was computed based largely on a convenience sample of available test publication manuals and employment testing data from the author's workplace. However, a test intercorrelation matrix based on commercially available tests will most likely result in lower bivariate correlations between test constructs because of the factor structure problems associated with "G" in the GATB.

Another possible limitation in the current study was the absence of multiple raters used to classify studies during the bare-bones meta-analysis. Best practice would suggest using a consensus process to ensure accurate classification. However, because only one rater was used, any study where there was confusion regarding the proper classification was thrown out.

With regard to future research conducted on JCV, it would benefit most from exploring ways to compute fully multivariate JCV batteries. Murphy and Shiarella (1997) suggest future research on personnel selection not only continue to use multiple test-predictors, but that the construct of job performance itself is a complex domain that can be defined by many levels. They provide a simple and straightforward set of calculations to compute validity coefficients using multiple predictor and criterion measures.

Cascio (1995) recommended that present efforts in constructing valid selection procedures move beyond the use of job-based predictors in order to keep up with the changing nature of work. In instances where time, money, resources, and small sample sizes limit the feasibility of a local validation study, JCV may be the best available alternative. The present study demonstrated the usefulness of the JCV method and its generalizability across a wide range of jobs and predictor constructs. To quote Hoffman and McPhail (1998, p. 999), "There will likely always be situations where some type of validation effort is needed. The Job Component Validity procedure simply adds another tool to the practitioner's toolbox."

APPENDIX A

APTITUDES MEASURED BY THE

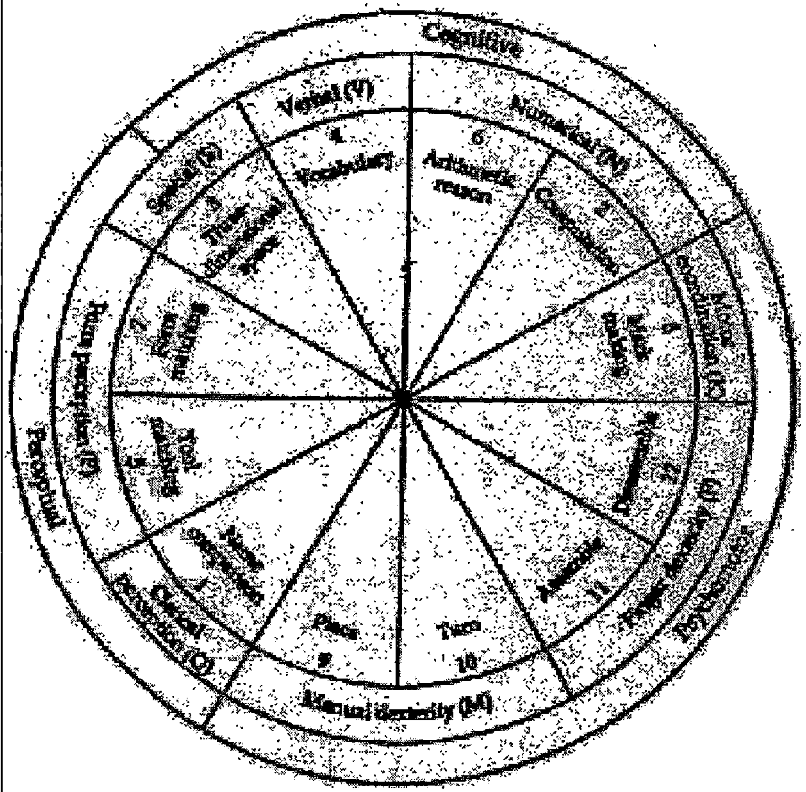
GENERAL APTITUDE TEST BATTERY

Aptitudes Measured by the Genralizated Aptitude Test Battery

General Learning Ability (also referred to as intelligence) (G)—"Catching on" and understanding instructions and principles as well as reasoning and judgment are tapped here. G is measured by Tests 3, 4, and 6 in the diagram.

Verbal Aptitude (V)—Understanding the meaning of words and relationships between them as well as using words effectively are some of the abilities tapped here. V is measured by Test 4.

Numerical Aptitude (N)—N is measured by tasks requiring the quick performance of arithmetic operations. It is measured by Tests 2 and 6.



Spatial Aptitude (S)—The ability to visualize and mentally manipulate geometric forms is taped here. S is measured by Test 3.

Form Perception (P)—Attention to detail, including the ability to discriminate slight differences in shapes, shading, lengths, and widths, as well as ability to perceive pertinent detail is measured. P is measured by Tests 5 and 7.

Clerical Perception (Q)—Attention to detail in written or tabular material as well as the ability to proffread words and numbers and to avoid perceptual errors in arithmetic computation is tapped here. Q is measured by Test 1.

APPENDIX B

GENERALIZED APTITUDE TEST

BATTERY TEST INTERCORRELATION

MATRIX

Generalized Aptitude Test Battery Test Intercorrelation Matrix (Hartigan et al. 1990)

	G	V	N	S	P	Q
G	1.00	0.84	0.86	0.74	0.61	0.64
V		1.00	0.67	0.46	0.47	0.62
N			1.00	0.51	0.58	0.66
S				1.00	0.59	0.39
P					1.00	0.65
Q						1.00

APPENDIX C

COMPARISON OF JOB COMPONENT

VALIDITY ESTIMATES WITH

OBSERVED VALIDATION RESULTS

Comparison of Job Component Validity Estimates with

Observed Validation Results (from McPhail, 1995)

Job	Construct	JCV prediction	Observed validity
Health physics technicians[a]	G	.38	.25
	N	.30	.37
	S	.10	.45[d]
	Q	.15	.30
Customer service representative[b]	G	.25	.16
	V	.13	.19
	V	.13	.42[e]
	N	.27	.25
Line repair workers[c]	N	.25	.64[f]

Note: All observed validity coefficients based on supervisor ratings criteria.
[a]nuclear power facility of an electric utility
[b]water products company
[c]electric utility
[d]Bennett Test of Mechanical Comprehension
[e]Following Oral Direction Test
[f]Content-specific proprietary test

APPENDIX D

COMPARISON OF JOB COMPONENT

VALIDITY ESTIMATES WITH

EMPIRICAL VALIDATION RESULTS

FOR DISTRIBUTION PLANNING JOBS

Comparison of Job Component Validity Estimates with Empirical Validation Results for Distribution Planning Jobs (from Holden, 1992)

	Predicted JCV by job[a]			Observed validity coefficients[b]		
GATB construct[c]	Assistant	Aide	Technician	Ratings	Job Know.	Job sample
G	.29	.28	.30	.20	.49	.49
V	.20	.19	.19	.19	.52	.29
N	.25	.23	.23	.28	.64	.47
S	.18	.16	.19	.32	.59	.46

[a]JCVs based on results of PAQ analysis for each job.
[b]Observed correlation between test score and supervisory ratings criterion, job knowledge criterion, or work sample criterion; combined N = 66 across three jobs.
[c]GATB constructs operationalized as follows: G (Adaptability); V (Industrial Reading Test); N (proprietary, custom-developed mathematics test); S (FIT patterns).

APPENDIX E

COMPARISON OF DESCRIPTIVE STATISTICS FOR JOB COMPONENT VALIDITY ESTIMATES AND OBSERVED VALIDITY COEFFICIENTS BY DOT CODE FOR CLERICAL OCCUPATIONS

Comparison of Descriptive Statistics for Job Component Validity Estimates and Observed Validity Coefficients by DOT Code for Clerical Occupations

(from Hoffman and McPhail, 1998)

	Mean job component validity estimate[a]					Mean observed validity coefficient[b]				
DOT Code	G	V	N	S	Q	G	V	N	S	Q
201-209 (A)	.24 (.05)	.18 (.03)	.24 (.04)	.12 (.03)	.20 (.03)	.24 (.18)	.19 (.16)	.23 (.14)	.09 (.11)	.22 (.17)
210-219 (B)	.26 (.02)	.19 (.02)	.25 (.03)	.14 (.02)	.20 (.02)	.23 (.17)	.20 (.17)	.25 (.15)	.20 (.15)	.24 (.15)
221-229 (C)	.25 (.01)	.18 (.01)	.24 (.02)	.14 (.02)	.20 (.01)	-	.18 (.13)	.30 (.17)	.23 (.16)	.22 (.13)
230-239 (D)	.24 (.03)	.18 (.03)	.25 (.03)	.11 (.02)	.20 (.02)	-	-	-	-	.19 (.16)
240-249 (E)	.27 (.02)	.20 (.02)	.25 (.02)	.14 (.02)	.21 (.01)	.21 (.12)	-	.21 (.08)	-	.18 (.14)
All clerical jobs (A-E)	.25 (.03)	.19 (.02)	.25 (.03)	.13 (.02)	.20 (.02)	.24 (.17)	.19 (.16)	.24 (.14)	.14 (.15)	.22 (.16)

Note: Standard deviations listed in parentheses. DOT groupings and letters (A,B, etc.) identical to Pearlman et al. (1980).

[a]Based on 51 jobs in utility company PAQ job evaluation database; DOT Occupational Groups 201-209—16 jobs; 210-219—15 jobs; 221-229—5 jobs; 230-239—5 jobs; 240-249—10 jobs.

[b]Based on mean observed validity coefficients compiled and reported by Pearlman, Schmidt, & Hunter (1980).

APPENDIX F

SOURCE OF STUDIES USED TO

COMPUTE COMMERCIALLY AVAILABLE

TEST INTERCORRELATION MATRIX

Source of Studies to Used to Compute Commercially

Available Test Intercorrelation Matrix

	GATB Test Combination														
Source	GV	GN	GS	GP	GQ	VS	VN	VP	VQ	NS	NP	NQ	SP	SQ	PQ
Company Data	0.31	0.60	0.42			0.22	0.43			0.42					
Company Data	0.52	0.51			0.60		0.50		0.46			0.48			
Company Data										0.57	0.43		0.39		
EAS Tech Manual	0.26		0.35	0.19		0.22	0.26	0.10					0.34		
EAS Tech Manual	0.40		0.30	0.16											
EAS Tech Manual	0.27		0.29	0.22											
WTMA Tech Manual			0.46			0.56				0.37					
WTMA Tech Manual						0.72				0.20					
WTMA Tech Manual										0.26					
IPI Tech Manual		0.62		0.52						0.29	0.54				
IRT Tech Manual						0.67			0.73						
IRT Tech Manual						0.37			0.16						
IRT Tech Manual						0.38			0.03						
IRT Tech Manual						0.16			0.28						
Adaptability Tech Manual			0.49												
Adaptability Tech Manual			0.18												
Adaptability Tech Manual			0.44												
Adaptability Tech Manual			0.46												
Adaptability Tech Manual	0.42														
Adaptability Tech Manual	0.36														
Adaptability Tech Manual		0.31													
Reading Index	0.20					0.50	0.52								
Reading Index							0.39								
Reading Index						0.37									
Reading Index						0.37									
Arithmetic Index		0.26					0.61			0.56					
Arithmetic Index							0.52			0.41					
Arithmetic Index										0.41					
FACT Tech Manual			0.30							0.08					
FACT Tech Manual			0.27												
FACT Tech Manual			0.27												
FACT Tech Manual		0.26	0.50												
FACT Tech Manual			0.27		0.19										
FACT Tech Manual										0.14					
FACT Tech Manual		0.13	0.21		0.14										

APPENDIX G

RESEARCH PARTICIPATION LETTER

To Whom It May Concern:

I am a graduate student enrolled in the MS I/O Psychology program at California State University, San Bernardino. I am conducting a thesis research study titled, "Comparing Job Component Validity to Observed Validity Across Jobs." Dr. Kenneth Shultz, CSUSB, is my thesis chair, and Dr. Cal Hoffman, Alliant University, and Dr. Matt Riggs, Loma Linda University, are on my committee. I am requesting your participation in the data collection phase of my study.

I am collecting observed validity coefficients from commercially developed tests used to predict job and/or training performance in a wide-range of jobs varying in complexity. Once collected, I plan to compare these observed validity coefficients to predicted validity coefficients using the Job Component Validity feature provided by the Position Analysis Questionnaire (PAQ). The goal of my study is to provide further evidence to support test-use without conducting local validation.

Below lists the type of data that I need:

Must Haves:	Nice To Haves:
▪ Uncorrected correlation coefficient ▪ Type of correlation coefficient ▪ Sample Size ▪ Criterion measure used ▪ Type of validation study strategy employed ▪ Name of type of specific tests used ▪ DOT code or enough information about the job to appropriately classify the job myself.	▪ Sample composition in terms of employment status, gender, and race ▪ Mean and standard deviations of the test scores used in the study ▪ Criterion reliability coefficients

If you decide to participate, all research findings will be made available to you when the study is completed. All information and data you provide will be kept strictly confidential and be returned to you immediately if requested. There are several options to send me your data. You can email it as an attachment, fax it, or mail it. If necessary, you may charge me for the mailing costs, however I urge you to send it the most inexpensive way as possible.

I have enclosed a short summary which explains the study in more detail if you are interested. I have also included a form that provides an example of the data I am requesting. If you would like to participate but have a question or concern, please contact me at the phone number or email address below. Thank you in advance for your thoughtful consideration.

Sincerely,

David Morris

Enclosure

Comparing Job Component Validity to Observed Validity Across Jobs

By the early 1980's, the need to conduct local validation research to support using cognitive ability tests to make personnel decisions seemed to be eliminated. Schmidt and Hunter's meta-analytic research (1981) found that statistical artifacts accounted for most, if not all the variance between validation studies performed on similar types of jobs. This led to the claim that, "Professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs". Pearlman, Schmidt, and Hunter (1980) recommended, "All that is needed to generalize validity is enough information to be able to compare the targeted job to similar jobs used in the initial validation study.

Ten years later, Guion (1991) concluded, "The sole use of VG is probably premature. At a minimum, a job analysis should be carried out and contain site visits, multiple interviews with incumbents and supervisors, as well as the breakdown of jobs into major tasks and behaviors and findings documented in a technical report." Around the same time VG was introduced into the research literature, another type of "synthetic validity" surfaced. This one derived directly from job analysis ratings. The "Job Component Validity" model, part of the normal output from the Position Analysis Questionnaire (PAQ), produces estimated validity coeffecients used to predict mean test scores and validity coefficients for cognitive ability constructs such as verbal, numerical, spatial, and general mental ability (Jeanneret, 1992).

Unfortunately, early studies showed disappointingly low correlations between predicted and observed validity coefficients using the JCV procedure (Marquardt & McCormick, 1974; McCormick, Mecham, & Jeanneret, 1977; and Mecham & McCormick, 1969). In 1991, Sackett remarked, "Its inability to predict observed validity coefficients calls for further research before one could feel confident relying on JCV." However, in a recent study by Hoffman and McPhail (1998) JCV estimates from 51 clerical jobs were compared to Pearlman, Schmidt, and Hunter's (1980) meta-analysis reporting mean observed validity coefficients for five DOT clerical categories. Their results showed substantial similarity to the mean observed, uncorrected criterion related validity coefficients produced in Pearlman et al's. (1980) study. Hoffman and McPhail discovered that averaging across a large sample of jobs provided more stable estimates, thus minimizing the effects of statistical artifacts normally encountered among single studies. Thus, it seems evident that early JCV studies suffered from the same weaknesses local validation studies suffer from: Small sample sizes.

The current study is designed to extend Hoffman and McPhail's (1998) research to a wider array of jobs ranging in degree of complexity. In addition, it seeks to construct "multivariate" JCV estimates, thus replacing the need to rely on the single, highest univariate JCV coefficient as the best estimate of a battery-validity. The final result will be another selection tool researchers can add to their toolbox enabling them to support test use without having to conduct local validation.

APPENDIX H

OBSERVED VALIDITY STUDIES

MATCHED TO JOB COMPONENT

VALIDITY ESTIMATES

Observed Validity Studies Matched to JCV Estimates

DOT	JOB TITLE	n	G	V	N	S	P	Q
003.167-018	Designers	16						0.07
003.167-026	Customer Extension Planners	32		0.30				
003.281-010	Drafter	99			0.18			
007.161-018	Engineering Assistants	11						0.08
029.261-022	Chemical Technicians	25		0.20	0.64	0.18		
030.162-010	Computer Programmers	1229	0.28	0.28	0.29	0.08	0.10	0.34
160.167-054	Claims Auditor	379		0.36	0.33			
166.167-034	Labor Relations Professionals	76		0.27				
183.117-014	Managers	122		0.32	0.35	0.18		
209.367-054	Yard Clerk	390		0.14	0.28			
209.567-010	Meter Readers	224		0.20	0.27	0.39		0.37
213.362-010	Computer Operator	257	0.25		0.33			0.32
222.387-034	Materials Clerks	54		0.38	0.37			
235.462-010	Telephone Operators (Information and Toll)	236		0.27	0.23			0.23
235.662-026	Telephone Service Representative	93			0.22			
253.357-010	Sales Representatives	107		0.42	0.48			0.37
292.353-010	Salesperson-Driver/Routeperson	88		0.27	0.18			
373.364-010	Probationary Firefighters	119				0.19		
375.263-014	Police Officers	209	0.03					
558.685-062	Chemical Operator	55		0.06	0.26			
600.280-022	Machinist	264		0.20	0.22	0.22	0.20	
616.380-018	Machine Operator	65				0.22		
619.686-022	Production Workers	422		0.17	0.21	0.25		0.15
620.261-010	Mechanics	190		0.15	0.17	0.24	0.13	0.20
620.281-046	Maintenance Specialists & Field Technicians	160	0.19		0.20			
638.281-014	Maintenance Mechanics	551		0.12	0.22	0.27	0.17	0.11
726.261-018	Technicians	327			0.22	0.16	0.27	0.23
729.281-014	Test Personnel	36			0.33	0.13		
821.261-014	Journeyman Line Maintainers	344			0.07	0.23		
822.281-018	Equipment Mechanics	119			0.23		0.29	
822.381-010	Equipment Installers	122			0.19		0.32	
822.381-014	Installer-Repairers	91			0.22			
824.261-010	Electrician	216		0.26		0.28	0.22	
829.361-010	Cable Splicers	88			0.26		0.27	
859.683-010	Heavy Equipment Operator	11		0.26		0.19		
860.381-022	Carpenter	144		0.25				
862.381-030	Plumber	90			0.05	0.22	0.06	
899.261-014	Plant Technicians	371			0.30	0.35		
913.463.010	Bus Drivers	179		0.05				
920.687-134	Packer	89		0.18		0.14	0.14	
921.683-050	Power Truck Operators	44		0.02	0.18			0.11
922.687-058	Laborers	432		0.11	0.18	0.12		0.31
959.574-010	Service Representatives	83		0.26	0.25	0.31		0.21
973.381-018	Press Workers	17			0.32	0.12	0.40	0.44

APPENDIX I

JOB COMPONENT VALIDITY ESTIMATES MATCHED TO OBSERVED VALIDITY STUDIES

JCV Estimates Matched to Observed Validity Studies

DOT	PAQ TITLE	G	V	N	S	P	Q
003.167-018	Eng Ele Pwrsys	0.41	0.30	0.32	0.27	0.19	0.25
003.167-026	Eng Sys Develo	0.39	0.29	0.30	0.24	0.18	0.23
003.281-010	Drafter Ele	0.36	0.23	0.28	0.22	0.19	0.21
007.161-018	Eng Mech Asst	0.36	0.24	0.28	0.24	0.20	0.21
029.261-022	Test Petroleum	0.31	0.21	0.25	0.21	0.21	0.20
030.162-010	Progr Computer	0.33	0.24	0.29	0.18	0.19	0.21
160.167-054	Auditor	0.33	0.27	0.30	0.18	0.15	0.21
166.167-034	Mgr Labor Relata	0.32	0.29	0.28	0.17	0.15	0.22
183.117-014	Spt Production	0.34	0.27	0.29	0.22	0.17	0.22
209.367-054	Clk Yard RR	0.22	0.16	0.25	0.10	0.18	0.17
209.567-010	Meter Reader	0.26	0.19	0.25	0.12	0.19	0.20
213.362-010	Computer Op	0.25	0.20	0.26	0.13	0.19	0.20
221.367-070	Clk Svc Repair	0.28	0.23	0.31	0.14	0.21	0.23
222.387-0.34	Clk Material	0.26	0.19	0.27	0.14	0.21	0.21
235.462-010	Teleph Op Cent	0.25	0.22	0.30	0.11	0.19	0.21
235.662-026	Teleph Answer I	0.22	0.20	0.29	0.08	0.18	0.21
253.357-010	Sales Pub Util	0.34	0.29	0.28	0.19	0.11	0.19
292.353-010	Driver Sales R	0.19	0.15	0.18	0.11	0.20	0.19
373.364-010	Fighter Fighter	0.31	0.24	0.28	0.19	0.23	0.25
375.263-014	Police Ofcr 1	0.27	0.24	0.25	0.16	0.16	0.21
558.685-062	Chem Op 2	0.32	0.21	0.29	0.22	0.24	0.24
600.280-022	Machinist Gen	0.32	0.21	0.25	0.29	0.23	0.21
616.380-018	Mach Op 1	0.24	0.17	0.23	0.19	0.20	0.19
619.686-022	Metal Fab Hlp	0.25	0.17	0.22	0.17	0.22	0.21
620.261-010	Auto Mech	0.31	0.22	0.25	0.25	0.22	0.21
620.281-046	Maint Mech	0.32	0.23	0.24	0.26	0.19	0.18
638.281-014	Maint Mech Gen	0.33	0.22	0.28	0.28	0.23	0.22
726.261-018	Ele Tester Gen	0.32	0.22	0.26	0.22	0.20	0.21
729.281-014	Repair Ele Met	0.34	0.22	0.26	0.27	0.21	0.20
821.261-014	Line Maintaine	0.30	0.22	0.25	0.20	0.17	0.19
822.281-018	Maint Mech Tel	0.29	0.21	0.25	0.17	0.21	0.21
822.381-010	Equip Installe	0.26	0.21	0.24	0.12	0.17	0.18
822.381-014	Line Installer	0.29	0.20	0.22	0.20	0.21	0.21
824.261-010	Electrcn	0.34	0.23	0.29	0.25	0.21	0.21
829.361-010	Cable Splicer	0.30	0.22	0.27	0.21	0.19	0.19
859.683-010	Operating Eng	0.27	0.19	0.23	0.19	0.23	0.22
860.381-022	Carpenter	0.31	0.22	0.26	0.26	0.22	0.23
862.381-030	Plumber	0.31	0.22	0.25	0.23	0.21	0.21
899.261-014	Maint Repair I	0.33	0.23	0.28	0.28	0.23	0.21
913.463-010	Bus Driver	0.25	0.20	0.25	0.12	0.21	0.22
920.687-134	Packer Agri Pr	0.11	0.09	0.15	0.08	0.16	0.13
921.683-050	Indust Truck O	0.22	0.15	0.23	0.15	0.22	0.20
922.687-058	Laborer	0.21	0.15	0.22	0.12	0.21	0.19
959.574-010	SVS Rep Util	0.26	0.18	0.34	0.05	0.17	0.22
973.381-018	Job Printer	0.23	0.19	0.21	0.18	0.20	0.22

APPENDIX J

JOB COMPONENT VALIDITY BATTERY

MATRIX EQUATIONS

JCV Battery Matrix Equations (Tabachnick & Fidell, 1997)

Another way of looking at R^2 is in terms of the correlations between each of the predictor and criterion variables. The squared multiple correlation is the sum across all predictor variables of the product of the correlation between the criterion and predictor and the (standardized) regression coefficient for the predictor.

In matrix form:

$$R^2 = R_{yi}B_I$$

Where R_{yi} is the row matrix of correlation between the criterion and the *k* predictor variables, and B_I is a column matrix of standardized regression coefficients for the same *k* predictor variables.

The standardized regression coefficients can be found by inverting the matrix of correlations among predictor variables and multiplying that inverse by the matrix of correlations between the criterion and predictor variables.

$$B_I = R^{-1}{}_{ii}R_{iy}$$

B_I is the column matrix of standardized regression coefficients, $R^{-1}{}_{ii}R_{iy}$ is the inverse of the matrix of correlations among the predictors, and R_{iy} is the column matrix of correlations between the criterion and predictor.

Because multiplication by an inverse is the same as division, the column matrix of correlations between the predictors and the criterion is divided by the correlation matrix of predictor variables.

See example below:

$$B_I = \begin{bmatrix} 1.203 & -.317 & -.204 \\ -.317 & 2.671 & -1.973 \\ -.204 & -1.973 & 2.622 \end{bmatrix} \begin{bmatrix} .57 \\ .73 \\ .75 \end{bmatrix} = \begin{bmatrix} .319 \\ .291 \\ .402 \end{bmatrix}$$

$$R^2 = \begin{bmatrix} .59 & .73 & .75 \end{bmatrix} \begin{bmatrix} .319 \\ .291 \\ .402 \end{bmatrix} = .702$$

R = .84

APPENDIX K

RESULTS OF BARE BONES

META-ANALYSIS

Results of Bare Bones Meta-analysis

Predictor	K	N	$\bar{r}$	s^2_r	s^2_e	s^2_p	%Explained	95% CI	JCV
G	32	1898	.23	.2022	.0154	.0049	75.97	.19-.27	.29
V	52	5042	.20	.0116	.0096	.0020	83.01	.17-.22	.22
N	72	6780	.24	.0104	.0095	.0009	91.82	.22-.26	.26
S	42	4444	.23	.0097	.0086	.0011	88.88	.20-.26	.20
P	7	703	.27	.0091	.0087	.0004	95.33	.20-.34	.20
Q	28	2145	.24	.0132	.0117	.0015	88.34	.20-.28	.21

APPENDIX L

TEST INTERCORRELATION MATRIX

FOR COMMERCIALLY AVAILABLE

TESTS

Test Intercorrelation Matrix for Commercially Available Tests

	G	V	N	S	P	Q
G	1.00	0.34	0.38	0.35	0.27	0.31
V		1.00	0.46	0.41	0.10	0.46
N			1.00	0.33	0.49	0.48
S				1.00	0.37	NA
P					1.00	NA
Q						1.00

APPENDIX M

JOB COMPONENT VALIDITY BATTERY

COMPUTATIONS FOR COGNITIVE AND

PERCEPTUAL TEST COMBINATIONS

JCV Battery Computations for Cognitive and Perceptual Test Combinations

Commercial Test Intercorrelation Matrices

	G	V	N
G	1.00	.34	.38
V		1.00	.46
N			1.00

	S	P	Q
S	1.00	.37	.39
P		1.00	.65
Q			1.00

<u>Cognitive Example</u>:

$$B_I = \begin{bmatrix} 1.218 & -.255 & -.345 \\ -.255 & 1.322 & -.511 \\ -.345 & -.511 & 1.366 \end{bmatrix} \begin{bmatrix} .29 \\ .22 \\ .26 \end{bmatrix} = \begin{bmatrix} .20742 \\ .08403 \\ .14269 \end{bmatrix}$$

$$R^2 = \begin{bmatrix} .29 & .22 & .26 \end{bmatrix} \begin{bmatrix} .20742 \\ .08403 \\ .14269 \end{bmatrix} = .11574$$

R = .34

<u>Perceptual Example</u>:

$$B_I = \begin{bmatrix} 1.213 & -.245 & -.314 \\ -.245 & 1.781 & -1.062 \\ -.314 & -1.062 & 1.813 \end{bmatrix} \begin{bmatrix} .20 \\ .20 \\ .21 \end{bmatrix} = \begin{bmatrix} .12766 \\ .08418 \\ .10553 \end{bmatrix}$$

$$R^2 = \begin{bmatrix} .20 & .20 & .21 \end{bmatrix} \begin{bmatrix} .12766 \\ .08418 \\ .10553 \end{bmatrix} = .06453$$

R = .25

REFERENCES

Cascio, W. F. (1995). Whither industrial and organizational psychology in a changing world of work? American Psychologist, 50, 928-939.

Church, A. H. (1998) From both sides now: A look to the future. The Industrial-Organizational Pyschologist, 35(2), 91-104.

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996) Psychological Testing and Assessment: An Introduction to Tests and Measurement (pp. 644-650). Mountain View, CA: Mayfield Publishing Company.

Ghiselli, E. E. (1966). The generalization of validity. Personnel Psychology, 19, 397-403.

Guion, R. M. (1998). Assessment, measurement, and prediction for personnel decisions. Mahwah, NJ: Lawrence Erlbaum Associates.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, D. C.: National Academy Press.

Hoffman, C. C., & McPhail, S. M. (1998). Exploring options for supporting test use in situations precluding local validation. Personnel Psychology, 51, 987-1003.

Holden, L. M. (1992). Job analysis and validity study for the distribution planning office technical progression. Unpublished technical report, Southern California Gas Company, Los Angeles.

Jeanneret, P. J. (1992). Applications of job component/synthetic validity to construct validity. Human Factors, 5(1&2), 81-96.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. American Psychologist, 41, 1183-1192.

Lawshe, C. H. (1952). What can industrial psychology do for small businesses? Personnel Psychology, 5, 31-34.

Levine, E. L. (1983). Everything you always wanted to know about job analysis. Tampa, FL: Mariner Publishing, Inc.

Marquardt, L. D., McCormick, E. J. (1974). The job dimensions underlying the job elements of the Position Analysis Questionnaire (PAQ Form B; Tech. Rep. No. 4). West Lafayette, IN: Purdue University, Psychological Sciences Department.

McCormick, E. J., Mecham, R. C., & Jeanneret, P. R. (1977). Technical manual for the Position Analysis Questionnaire (PAQ). Logan, UT: PAQ Services.

McCormick, e. J., Mecham, R. C., & Jeanneret, P. R. (1989). Technical manual for the Position Analysis Questionnaire (PAQ). Logan, UT: PAQ Services.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the position analysis questionnaire (PAQ). Journal of Applied Psychology Monograph, 56,(4), 347-368.

McCormick, E. J., DeNisi, A. S., & Shaw, J. B. (1979). Use of the position analysis questionnaire for establishing the job component validity of tests. Journal of Applied Psychology, 64,(1), 51-56.

McPhail, S. M. (1995). Job component validity predictions compared to empirical validities. In C. C. Hoffman (Chair), Current innovations in PAQ-based research and practice. Symposium conducted at the annual conference of the society for Industrial and Organizational Psychology, Orlando, FL.

McPhail, S. M., Jeanneret, P. R., McCormick, E. J., & Mecham, R. C. (1998). Technical manual for the Position Analysis Questionnaire (PAQ). Logan, UT: PAQ Services. (Available through Consulting Psychologists Press Inc., Palo Alto, CA.)

Mecham, R. C., & McCormick, E. J. (1969). The use of data based of job elements and job dimensions based on the Position Analysis Questionnaire. Occupational Research Center, Purdue University, June 1969. (Prepared for office of naval Research under contract Nonr-1100 (28), Report No. 3.) (b)

Mossholder, K. W., & Arvey, R. D. (1984). Synthetic validity: A conceptual and comparative review. Journal of Applied Psychology, 69, (2), 322-333.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. Personnel Psychology, 50, 823-854.

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

Ruch, W. W., Weiner, J. A., McKillip, R. H., & Dye, D. A. (1985). Technical manual for the PSI basic skills tests for business, industry, and government. Glendale, CA: Psychological Services, Inc.

Sackett, P. R. (1991). Exploring strategies for clustering military occupations. In Wigdor AK, Green BF (Eds.), Performance Assessment for the Workplace, Vol. II, (pp. 305-332), Washington, DC: National Academy Press.

Schmidt, F. L., & Hunter, J. E. (1983). A within setting empirical test of the situational specificity hypothesis in personnel selection. Personnel Psychology, 37, 317-326.

Schmidt, F. L., & Hunter, J. E. (1990). Methods of Meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage Publications.

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

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. Journal of Applied Psychology, 78, (1), 3-12.

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

Tabachnick, B. G., & Fidell, L. S. (1997). Using multivariate statistics (3rd ed.). Northridge, CA: Harper Collins College Publishers.

Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg. 38290-38315 (1978).