Electronic Theses and Dissertations, 2004-2019

2010

Issues In Validity Generalization The Criterion Problem

Raquel Hodge
University of Central Florida

STARS Citation

Hodge, Raquel, "Issues In Validity Generalization The Criterion Problem" (2010). *Electronic Theses and Dissertations, 2004-2019*. 1528.
https://stars.library.ucf.edu/etd/1528

ISSUES IN VALIDITY GENERALIZATION: THE CRITERION PROBLEM

By

RAQUEL HODGE
University of Central Florida

A thesis submitted in partial fulfillment of requirements
for the degree of Master of Science
in the department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2010

ABSTRACT

Schmidt and Hunter's validity generalization model poses seven sources of error variance affecting validation studies. Of the seven sources of error variance, only four sources have been tested. This study looks at an additional source of error variance, the difference between studies in the amount and kind of criterion contamination and deficiency, as proposed by Schmidt and Hunter. The current study proposes a method of evaluating criterion contamination and deficiency in criterion measures in order to minimize their effects on the relationship between criterion and predictor measures. Two unique criteria are used including a traditional subjective measure of current performance and a non-traditional subjective measure of expandability (future performance). Data from 378 employees from a large international financial institution were used to test the proposed method.

Results do not support the hypotheses. Single criteria predicted the same or better than the combined criteria, suggesting that the criterion problem was not addressed. Possible reasons for these findings are discussed. An unexpected finding supports the utility of personality measures compared to cognitive ability measures. The study concludes with a discussion of the implications and limitations of the study as well as directions for future research.

This thesis is dedicated to my parents; for their constant encouragement and faith in me.

ACKNOWLEDGMENTS

I would like to offer a special thank you to my committee members Dr. Robert Pritchard and Dr. Nancy Reed.

Dr. Pritchard, I am so grateful that you accepted to be a part of my committee. I am especially grateful to you for not taking it easy on me and evaluating me to your highest standard. You expected no less from me than you would from your doctoral students and as a result, I also expect nothing less from myself.

Dr. Reed, your dedication to your students is something that is invaluable. Having your voice in my committee offered a perspective that I greatly respect. Thank you for sharing your many experiences with me and being a source of stress relief throughout this process by always keeping your candy jar full.

I would also like to offer a very special thank you to my committee chairperson, Dr. William Wooten. You have offered so much more help than I could ask for. You have made this whole process go smooth and easy for me. Thank you so much for taking the time to work with me and make this thesis something that I can be proud of.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Much like academic researchers hope their conclusions can generalize to different populations, the applied researcher is concerned with generalizing the results of validation studies across jobs and organizations. Since the initial publication of the validity generalization model by Schmidt and Hunter (1977) the topic of validity generalization has received much attention. Though the Schmidt and Hunter group have provided evidence in support of their model (Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman, & Shane, 1979; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, & Caplan, 1981; Schmidt, Hunter, Pearlman, & Rothstein-Hirsh, 1985) only four of the seven artifactual sources of error proposed by Schmidt and Hunter have been tested. This paper focuses on sources of error variance in validation research collectively known as the “criterion problem.”

The criterion has been a widely studied topic for I-O psychologists since the late 1800s/early 1900s (Austin & Villanova, 1992). Criteria have been defined in many different ways. Stone and Kendall (1956) defined a criterion as “a standard which can be used as a yardstick for measuring employees’ success or failure” (p. 271) (as quoted in Dunnette, 1963). Krug (1961) defines it as “an index against which other indexes may be compared and evaluated… a measure of job success…the desired end product of selection” (p. 107) (as quoted in Dunnette, 1963). Finally, Austin and Villanova (1992) define it as “a sample of performance, measured directly or indirectly, perceived to be of value to organizational constituencies for facilitating decisions about predictors or programs” (p. 838). Organizations can use a variety of information as criteria. Some categories include objective production data (e.g. number of goods produced, dollar sales), personnel data (e.g. absenteeism, turnover, and promotions), judgmental data (e.g. supervisor ratings), job or work sample data, and training proficiency data (Gatewood, Feild, & Barrick, 2008). Though efforts have been met with opposition by some (Dunnette, 1963), research in personnel psychology has pushed the use of criteria in selection validation research. Dunnette (1963) states the search for *the* criterion has consumed researchers for years. His position is that a search for an all encompassing measure of job success is unlikely to be found because the nature of work and job success is very complex and multi-faceted. Dunnette (1963) calls for researchers to abandon the fruitless search for *the* criterion and begin to look at success on the job as something that has many different facets. Dunnette (1963) makes a valid point about the complex nature of criteria. Well developed criteria can be a challenge for researchers because of the "difficulties involved in the process of conceptualizing and measuring performance constructs that are multidimensional and appropriate for different purposes" (Austin & Villanova, 1992 p. 836). Criteria are heavily theoretical and operational in nature.

A criterion is a performance measure. In validation research, performance measures are compared to test scores or other pre-employment information to determine the type of relationship between the criterion and the test scores (i.e. predictors). For businesses that employ a selection procedure, it is a primary concern that test scores are able to predict success on the job. Valid selection procedures promote individual, group, and organizational effectiveness and are legally defensible (Van Iddekinge & Ployhart, 2008). Organizations such as the Society for Industrial Organizational Psychologists (SIOP), American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), and The Equal Employment Opportunity Commission (EEOC) have developed documents that outline the requirements for selection procedures. Businesses are required to maintain selection procedures that are in agreement with the rules outlined in the documents.

According to the *Principles for Validation Use of Personnel Selection Procedures*, (SIOP, 2003), criteria are developed through job analysis and should reflect the work performed, the setting, and the organization's goals. Predictors and criteria should be chosen with an understanding of the "objectives for test use, job information, and existing knowledge regarding test validity" (p. 13). Predictors are used by the organization to predict future performance (i.e. criteria). Predictors are more likely to be valid if a strong relationship exists between the test and the behavior it is predicting. Thus, a company must choose a predictor that through job analysis data is related to one or more performance criteria. Criteria and predictors are related to each other in that the criteria chosen will be used to evaluate the predictors. Put another way, criterion measures are the criteria used to evaluate selection procedures and determine if they are valid (Gatewood et al., 2008). However, job success is composed of many independent components, which may be related to very different types of predictors. Attendance and error rates are examples and underscore the complex nature of criteria and prediction.

When a company employs a selection procedure, it is their hope that the outcome will be a test that is predictive of good (and bad) job performance. It is also their hope that they can use this predictor across various jobs in their organization. Generalizability becomes an issue of importance to managers because of the time and effort involved in conducting validation studies. There are two schools of thought on the issue of validity generalization which are expanded upon in the sections that follow.

Situational Specificity of Predictor Validity

Traditionally, validation studies have been carried out for each situation in which a test is used. For example, a test that has been validated for the position of firefighters in Birmingham, AL, will need to be re-evaluated for the position of firefighters in Tampa, FL. These two jobs may very well have the exact same job description, but their locations constitute a new situation. Schmidt and Hunter (1977) state that the explanation for this is that advocates of situational specificity believe that the human observer or analyst is not adequate enough to receive or process the subtle differences that occur from job situation to job situation. James, Demaree & Mulaik (1986) believe that individual validation studies are necessary because "true validities vary as a function of validation situation" (p. 440). In other words, situational specificity assumes that even minute differences in an organization's culture, applicant pool, core values, or structure of work can make huge differences in the results of validation studies (as expressed by validity coefficients). Even for similar jobs, a test that has been shown to be valid in one situation may not be valid in another situation.

This theory makes for a difficult situation for managers. For them it means that they have to invest company time, money, and other resources to conduct a validation study for every job in their organization. Even if two jobs appear to be identical, something like different locations will require a separate validation study. Robert Guion, author of *Personnel Testing,* argued that in order to progress the science, researchers need to show that generalizability is indeed possible across jobs (as cited in Schmidt & Hunter, 1977).

In addition, there are theoretical reasons why the validity coefficients of studies conducted at the local level are unstable. In their analysis of 406 published validity studies, Lent, Aurbach, & Levin (1971) found a mean sample size of 68. Schmidt, Hunter and colleagues argue that the sample size of unpublished studies may be even lower. Such small sample sizes causes problems regarding the study's power. When a study has low power, confidence intervals are wider in range making effect size estimates less precise (Shadish, Cook, & Campbell, 2002). A study with a sample size of 68 or below may not be sensitive enough to identify significant effects, particularly if the effects are small. Schmidt, Hunter and colleagues also argue that the presence of an outlier can change the observed validity from positive to negative making local empirical studies about validity uncertain. They suggest that in order to provide a sound conclusion about validity, validity coefficients from a number of local studies should be combined into a distribution and subjected to meta-analysis (Schmidt, Hunter, Pearlman, & Hirsh, 1985). Empirical studies at the local level should still be conducted, but conclusions about validity are recommended only after combining the studies in meta-analysis.

Generalization of Predictor Validity

As an alternative to the situational specificity hypothesis, Schmidt & Hunter (1977) propose the validity generalization model. The idea behind validity generalization is that validity studies conducted in one situation can be generalized to new situations. Validity generalization is of practical significance to businesses because of the costs involved in conducting a validation study for every job. In addition, a company of smaller size would have much difficulty producing an acceptable sample size as well as the time and resources involved in a validity study. After a job has been validated, companies can use job analysis to compare similar jobs to the validated jobs (Pearlman, Schmidt & Hunter, 1980). Not only would this avoid a high cost, but companies would not be susceptible to litigation while they wait for every job to be validated.

For supporters of the validity generalization model, statistical artifacts are responsible for the observed changes in validity across situations (Schmidt & Hunter, 1977; Pearlman et al., 1980). As mentioned above, local empirical studies often use too small a sample size to detect significant and reliable effects. In their argument for their validity generalization model, Schmidt, Hunter, and colleagues compare their meta-analysis procedure to a single validity study. Validity generalization studies conducted by the Schmidt and Hunter group have consisted of 100 to 200 validity coefficients whose total samples sizes range from 8,000 to 20,000 making their conclusions about validity much more stable than a single study which can often show nonsignificant results even if the predictor is valid (Schmidt et al., 1985).

Once the statistical artifacts are controlled for, validity can easily be generalized to different situations. Brodgen & Taylor (1950) suggest four sources of error variance including: small sample sizes, computational and typographical errors, differences between studies in criterion reliability, and differences between studies in amount and kind of criterion contamination and deficiency. Other sources added to this list include range restriction (Schmidt & Hunter, 1977), differences between studies in test reliability, and differences in factor structure between tests of a given type (Pearlman et al., 1980). Finally, one must also consider variance due to true situational specificity.

The typical local empirical study can have one or more types of the statistical artifacts mentioned above. The effects of sample size have been previously explained. Details about the other artifacts are explained using the previous example of our firefighter position. Suppose that the firefighters in Tampa, FL have been on the job for an average of 15 years, while the firefighters in Birmingham, AL have only been on the job for an average of 5 years. A study done using currently employed firefighters to validate the use of a selection test (concurrent validity) might show greater range restriction in criterion scores within the Tampa sample. This may result in a non-significant validity correlation. Differences between studies in criterion and test reliability occur when local validity studies employ tests and performance measures that differ in their degree of consistency and stability. This would occur if Birmingham found good reliability estimates for their criterion and predictor measures, but Tampa did not. Differences in factor structure between tests can occur if both cities gave their firefighters a technical knowledge test, but the test in Birmingham was paper-and-pencil (heavily dependent of reading skills), while the test in Tampa was oral (heavily dependent on oral presentation skills). Criterion contamination and deficiency would occur if Birmingham graded their firefighter’s technical knowledge test for grammatical and spelling errors (a requirement not necessary for the position). Finally, computational and typographical errors occur when the data is mishandled or input incorrectly. These differences suggest that a new validation study is needed for a new situation.

The basic premise of the validity generalization model is that if statistical artifacts are controlled for, predictors of a similar group of jobs will show consistent validity results across multiple studies, and argue for the assumption of validity in new situations. Conceptually, one would need to convert a distribution of validity coefficients to Fisher’s z then simply subtract the error variance attributed to the artifactual sources from the total variance. If the resulting variance is zero, or close to zero, then the hypothesis of situational specificity can be rejected

(Schmidt & Hunter, 1977; Pearlman et al., 1980). Using meta-analytic techniques, the Schmidt & Hunter group correct prior distributions for four of these seven identified sources of error including sample size, range restriction, and differences in criterion and test reliability. Prior distributions are simply distributions of empirical validity coefficients collected across a number of similar studies. Like other distributions, prior distributions have a mean (average validity coefficient) and a variance (variability of validity coefficients). For example, in one study, Schmidt and Hunter found the mean to be .39 and the range to be -.04 to +.82 (1977). These large variances are cited as support for the situational specificity hypothesis. Collectively, the Schmidt and Hunter corrections both reduce the variance of the distribution and increase the average coefficient. Some corrections, such as correction for sample size, narrow the variance, and others, such as correction for attenuation, boost the mean. Schmidt and Hunter set an arbitrary criterion of 75% as a marker of success. If the application of their corrections reduces the variance by 75% or more, the validity generalization hypothesis is supported. The resulting mean coefficient, or a deviate score based on this mean, would then be used to ascertain if the tests are generally valid or not. Pearlman et al. (1980) also maintain that even if the result of this variance correction equation is not zero, validity generalization may still be possible. They state that after range restriction and criterion unreliability is accounted for as well as correcting the standard deviation, "it may become apparent that a very large percentage, say 90%, of the values in the distribution lie above the minimum level of useful validity…One could conclude with 90% confidence that true validity would be at or above this minimum level in a new situation involving this test type and job without carrying out a validation study of any kind" (p. 376). Job analysis information would be sufficient enough to ensure that a job is similar to the validated job class. A subsequent study (Schmidt, Hunter, & Caplan, 1981) also showed promising results for validity generalization.

Schmidt, Hunter, and colleagues have published a number of studies in support of validity generalization. Schmidt & Hunter (1977) applied their model to four validity distributions reported in previous literature. For the four distributions, nearly half of the observed variability was accounted for by three statistical artifacts including sample size, range restriction, and criterion reliability. Validity generalization was supported for two of the distributions. Schmidt, Hunter, Pearlman & Shane (1979) improved upon the previous model and applied it to 11 validity distributions for clerical workers and three distributions for first line supervisor positions. In this study, Schmidt and his colleagues attempt to account for the error variance from four different sources of error including sample size, range restriction, criterion unreliability, and predictor unreliability. In each of their empirical studies, Schmidt, Hunter, and colleagues use their 75% decision rule to determine if the situational specificity hypothesis can be rejected. Results showed that on average 60% of the variance was accounted for by the four artifacts. Although this result does not allow them to reject the situational specificity hypothesis based on their 75% rule, they conclude that the rest of the error variance can be accounted for by the other three sources not tested; especially differences in studies in the amount and kind of criterion contamination and deficiency. Schmidt and colleagues provide evidence (discussed previously) that allows them to justify validity generalization for most of the samples regardless of the inability to reject the situational specificity hypothesis. Pearlman, Schmidt & Hunter (1980) apply the model to 32 validity distributions based on job proficiency criteria and 24 validity distributions based on training criteria. Results showed that in a total of 28 cases error variance accounted for met the 75% rule allowing for validity generalization without further analysis. Average percentage of variance accounted for was 75% and 70% for validity distributions based on proficiency and training criteria, respectively.

Schmidt and Hunter (1977) attribute differences in validity across studies to artifactual sources. In their series of studies Schmidt, Hunter and colleagues only attempt to correct for four sources of error variance including differences between studies in test and criterion reliability, range restriction, and sampling error. Some occupational areas just fall short of their 75% rule. In theory, if more than four of the identified sources of error could be operationalized, validity generalization might find broader applicability. It is interesting to note that Schmidt and Hunter limited their approach to four sources because, in part, they felt that the others would be too difficult to operationally implement (Schmidt & Hunter, 1977). The focus of this paper is to present an operational implementation model to address an additional source of error variance which plagues primary empirical validation studies. This source of variance is referred to as the criterion problem. If successful, the applicability of the validity generalization model would be broadened. Meaning, for those jobs in which correcting for four sources of error variance is not enough to support validity generalization, correcting for the criterion problem may eliminate the necessary error variance to provide support for this theory.

The Criterion Problem

The term ultimate criterion was introduced and defined by Thorndike (1949) as, "…the complete final goal of a particular type of selection or training" (as quoted in Austin & Villanova, 1992). The ultimate criterion implies longitudinal measure of success over one's career. Ideally, predictors are chosen to try and predict the ultimate criterion, which is a person's all-encompassing measure of success. Predicting the ultimate criterion is not possible. If the ultimate criterion is an all-encompassing measure of success on the job, a person would have to be done with their career and thus, predicting performance would be futile. Instead, organizations and researchers use actual criteria that they can see and measure. Actual criteria are generally gathered after the predictor measure has been administrated. Because actual criteria are the only available alternative the ultimate criterion, they are referred to as the criterion. Unlike predictors, the criterion measure is not submitted to a test of its adequacy. The criterion must simply be logically justifiable and clearly connected to job analysis data (Brogden & Taylor, 1950). However, because of its unempirical nature, bias is introduced into the construction of criteria specifically, contamination and deficiency.

At the core of the criterion problem is the notion that success at work is a function of multiple and independent components. Success is not only how much a worker produces, but how well a worker gets along with coworkers, how well a worker adopts values and culture, or how well a worker navigates office politics. This multidimensional ultimate criterion will not be adequately reflected by any single actual criterion used by the organization. Because actual criteria are narrow by definition (they fall short of the ultimate criterion), they give rise to criterion deficiency. Sales volume, for example, is almost automatically deficient because it does not include other important elements of the sales job such as customer service and customer relations. Add in criterion contamination, and it becomes easier to understand why empirical results vary so much. Too high a reading level in cognitive tests such as the Wonderlic, for example, makes it contaminated when used with relatively low level workers, introducing a source of test variance not associated with the criterion, actual or ultimate. While deficiency and contamination are often defined as characteristics of the test, they can also be defined as characteristics of actual criteria. Compared to the theoretical ultimate criterion, sales volume, an actual criterion, is both deficient, but possibly contaminated (i.e., “A certain amount of kick back is expected.”) as well. The key to the criterion problem could lie in expanding the dimensionality of the actual criteria used. For example, adding a contextual performance indicator, such as “political awareness” to a traditional performance indicator such as sales volume could have the effect of increasing empirical coefficients. To explain a bit further, the criterion problem can be expressed as the amount of shared variance between a predictor and a criterion. The ratio between the shared variance and total variance represents an index that can be used to gauge or measure the extent of the problem.

Criterion Problem Ratio (CPR) = shared variance/(unshared deficiency + unshared contamination + shared variance)

In the equation above, the Criterion Problem Ratio is the ratio between criterion shared variance and total system variance. The criterion shared variance is the covariance between each predictor and each criterion. The total system variance represents all of the variance in the model. As CPR approaches 1, the proportion of shared variance increases and the observed validity coefficient is less affected by the criterion problem. As CPR approaches zero, the proportion of shared variance decreases and the observed validity coefficient is more affected by the criterion problem. Theoretically, adding more logically related but independent components to the actual criteria should result in higher CPR values. Higher CPR values should result in larger observed empirical validities. The effect size of implementing similarly constructed multidimensional criteria could then be estimated by comparing those studies using meta-analytic procedures.

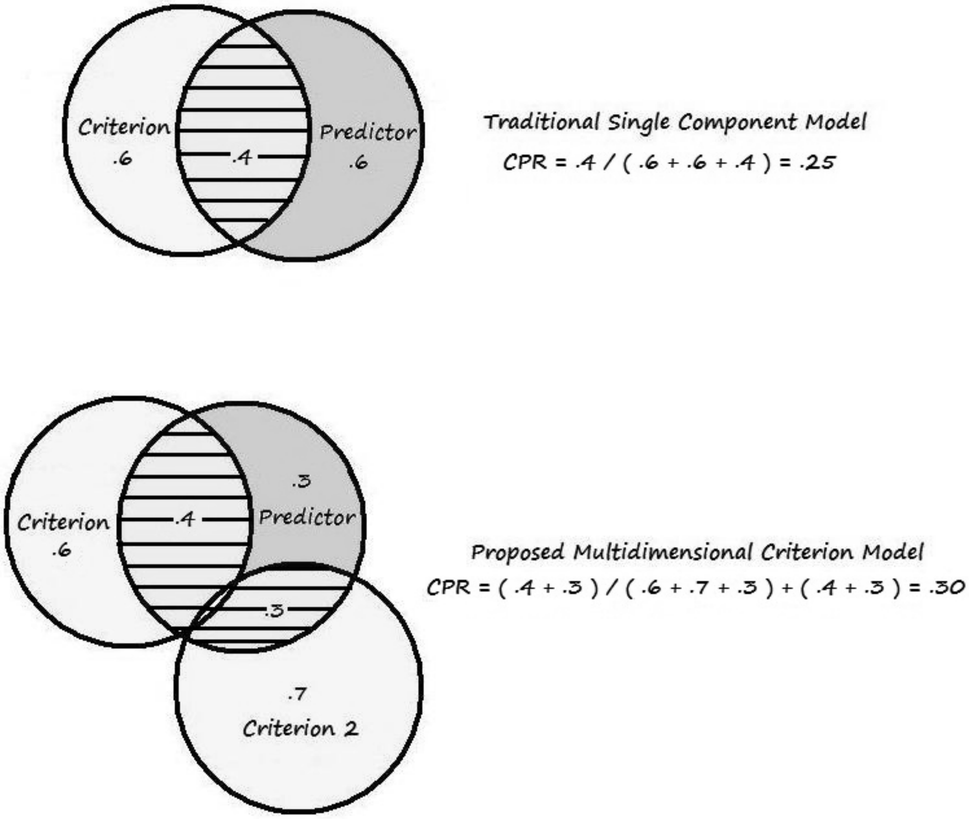
Criterion
.6
.4
Predictor
.6
Traditional Single Component Model
CPR = .4 / (.6 + .6 + .4) = .25
Criterion
.6
.4
.3
Predictor
.3
.7
Criterion 2
Proposed Multidimensional Criterion Model
CPR = (.4 + .3) / (.6 + .7 + .3) + (.4 + .3) = .30

Figure 1 Proposed Multidimensional Criterion Model

In the model pictured above, the striped area represents common variance between a predictor (e.g. cognitive ability test) and the criterion (traditional single component criterion such as sales volume). The light gray shaded area represents “criterion” deficiency, or variance not accounted for by the predictor. The dark gray shaded area represents criterion contamination. The ratio between these three sources of variance represents the “criterion problem.”

Theoretically, this ratio increases as additional, independent components are added to the criterion. While several obstacles lie in the path of fully implementing this model (i.e., theoretical nature of criterion bias, limitations of correlation coefficients, etc.), the current paper is an attempt to demonstrate its efficacy.

Criterion Bias

Brogden and Taylor (1950) define bias in a general sense by stating that a biasing factor is “any variable, except errors of measurement and sampling error, producing a deviation of obtained criterion scores from a hypothetical “true” criterion score” (p. 161). When utilizing data in a real situation, researchers will make their best attempt at removing or refraining from introducing error or bias in their analysis. However, it is not likely that perfection will be achieved in real world settings. In their study Schmidt & Hunter (1977) do not correct for criterion contamination and deficiency because they state, “it is difficult to estimate their frequency or magnitude” (p. 532). With this in mind, a researcher must be familiar with the different kinds of bias, their effects, how to eliminate them, and the resulting effects of bias which cannot be controlled. According to Brodgen and Taylor (1950) different biasing factors have different effects on data. Some bias will affect mostly the validity coefficients, others will affect mostly the criterion reliability, and others will affect both. For the purpose of this study, the focus will be on two types of criterion bias: criterion contamination, and criterion deficiency.

A criterion is considered contaminated when it includes extraneous elements (SIOP, 2003; Brodgen & Taylor 1950). Brodgen & Taylor (1950) state that contamination is most likely to occur when constructing scales for what is being measured. SIOP (2003) lists in the *Principles* several examples of contaminating factors such as differences in the quality of equipment, unequal sales territories, raters' knowledge of predictor scores, job tenure, shift, location of the job, and attitudes of the raters. They state that while it is extremely difficult for a researcher to identify and remove all sources of contamination, efforts should be made to minimize their effects. Standardizing the administration of the criterion measure, and measuring the contaminating variables in order to control for them are two suggestions offered in the *Principles* to minimize the impact of contamination.

The second type of criterion bias, criterion deficiency, occurs when a criterion excludes elements that are critical and relevant to the job in question (SIOP, 2003; Brogden & Taylor, 1950). Deficiency is most likely to happen during the analysis of the situation when determining which variables to include in the criterion measure. A thorough and systematic job analysis is the best way to avoid deficiency. A job analysis would minimize the possibility of overlooking important elements to be included in the criterion; it would give clues as to the most practical way of measuring those elements; it would show the relative importance of each element; and it facilitates other steps in criterion development as well as revealing which predictors would be the most valid (Brodgen & Taylor, 1950). Another reason for deficiency is using only one type of criterion measure. For example, one of the most common criterion measures, overall performance ratings, has less of a chance of being deficient because it takes into account a more complete picture of the job (given a thorough job analysis). However, subjective ratings such as overall performance ratings only have an advantage over objective methods given a thorough job analysis. Poor scale construction or poor conceptualization of performance can lead to deficiency in the measure. In sum, although subjective ratings (e.g. overall performance ratings), are more dimensionally complex, the concern for deficiency in the measure remains (Bommer, Johnson, Rich, Podsakoff, Mackenzie, 1995). Deficiency is also possible through the improper weighting of the factors that compose the overall rating score. There will be a difference in the amount and kind of criterion deficiency between raters, or for different ratings of the same rater because of the emphasis placed on some factors and the ignorance of others (Brodgen & Taylor, 1950). Dunnette (1963) points out that the search for a composite measure of job success has led research to ignore the many different facets and measures of employee success. Different techniques may be more useful for different criterion elements. By using only one type of measure, deficiency is easily introduced into the criterion.

As mentioned above, Brodgen & Taylor (1950) believe that a thorough job analysis (step one in criterion construction) is necessary in order to reduce the amount of deficiency as well as avoid other bias in the subsequent steps. They comment that researchers primarily focus on searching for available criterion measures. Using what is easily available neglects systematic analysis that can ensure all important elements of the job are included. As it follows, method of measurement and the combination of sub-criterion variables (steps two and three of criterion construction) will also be subject to availability and convenience. For example, when deciding on a criterion measure, it may be most convenient to use information on absenteeism because it is readily available. However a systematic job analysis may reveal more practical and efficient methods of measuring job success.

Contamination and deficiency will also undoubtedly have an effect on the predictor validities and partial regression weights. Brodgen & Taylor (1950) continue to say that this bias may "even result in the inclusion of tests in the battery that predict only bias and have no relationship to the ‘true’ criterion” (p. 163). The kinds of criteria organizations use come from various sources and all are susceptible to contamination or deficiency. Judgmental data consist of performance appraisals or ratings. Supervisors will make judgments of the subordinate’s behaviors important to job success. However, because of their subjectivity, these ratings can also be based on other factors not related to job success. Objective production data and personnel data (e.g. dollar sales and accident rates, respectively) are objective criteria, but are narrow in their scope and do not include a holistic picture of the job. Objective criteria can also be contaminated. For example, absenteeism can be both voluntary and un-voluntary. If a worker has an accident and is not physically able to come to work, his performance has not changed, but using absenteeism as a criterion may show otherwise. Work sample data are a miniature representation of the job, but may not be appropriate for more complex jobs. Finally, training proficiency data when available and used on relatively new incumbents can be contaminated by errors due to subjective trainability ratings, or can also include narrow objective measures.

Cognitive Ability Tests

Cognitive ability tests have long been used in personnel selection. Meta analysis studies on the correlation between cognitive ability tests and job performance show a mean validity coefficient of .30, which after correction for attenuation is reported at .50 (Outtz, 2002). The usefulness of cognitive ability tests in a selection setting is not a point of debate for researchers. Cognitive ability tests are some of the most valid selection instruments (Gatewood et al., 2008; Hunter & Hunter, 1984). What seems to cause controversy is the adverse impact caused by the use of cognitive ability testing.

There is a substantial mean difference between the scores of Whites and minority groups on ability testing (Hunter & Hunter, 1984; Outtz, 2002). According to one researcher, this difference in means on test scores is not reflected in the groups' job performance. Outtz (2002) reports that cognitive ability testing produces "differences that can be 10 times larger than racial group differences on measures of job performance." He goes on to comment that organizations typically have multiple goals: they want to hire the best people, but they also want a diverse workforce that will broaden their business to a diverse customer base. The problem with these separate goals is that a worker's value to the organization is normally determined by his productivity. Selection tools such as cognitive ability tests will not be enough to help employers meet their separate goals. Adverse impact can be an indication of a cognitive ability test's contamination. The test may include factors outside of job performance that result in adverse impact.

Cognitive ability tests are valid for a wide variety of jobs, more specifically for jobs with "greater information processing and problem solving demands" (Gatewood et al., 2008). The explanation that Gatewood et al. (2008) offers is that cognitive ability and job knowledge are highly correlated and in turn, job knowledge and job performance are highly correlated. This means that cognitive ability and job performance are also related. Gatewood and his colleagues (2008) note that cognitive ability tests are one of the best tools for selection in complex jobs. To avoid the consequences of adverse impact, they suggest increasing the effort to recruit minority applicants thereby increasing the number of minority applicants that pass a cut off score on the test; this was original intent of affirmative action programs.

Regardless of their wide spread use and validity data, cognitive ability tests alone cannot account for all types of success on the job. Mental ability will be important for the technical portions and task requirements of a job, but will not be as important for things like organizational/cultural fit. Cognitive ability tests are deficient in this aspect. They do not provide a way to predict future success on contextual factors. In addition, work often involves a codependency with coworkers and teams. As individual performance becomes more and more blurred, cognitive ability may be less sufficient for measuring success on the job.

Personality Tests

The use of personality tests in selection has become a common practice, but is not without controversy. In 1991 two meta-analyses on the validity of personality tests for personnel selection were published (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Up until these articles were published, personality inventories were seen as unfit to use in a selection setting. Since that time, the research on personality tests in selection has increased considerably. Researchers such as Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmidt (2007a) feel that a second look is needed on the use of personality tests in selection. In their article they criticize personality tests as having low criterion-related validity as well as being easily manipulated or faked by motivated applicants. However, research has shown that faking does not have an effect on the test's predictive validity and that distortion may in fact be related to real differences in personality (Gatewood et al. 2008). In their article Morgeson asks his colleagues (each an editor for *Personnel Psychology* or *Journal of Applied Psychology*) to comment on their review of the literature on personality tests in a selection context. The authors conclude that personality measures should be avoided in a selection context. Some researchers disagree with the opinion that personality tests should be taken out of selection (Tett & Christiansen, 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007). Morgeson et al. (2007b) however, believe that the data support the conclusion that “personality measures do not seem to have much value as predictors of job performance” (p. 1035).

The problem with dismissing personality characteristics is that managers intuitively believe that they matter (Gatewood et al., 2008). Personality traits such as conscientiousness and emotional stability are appealing in incumbent employees because managers want hardworking, persistent and achievement oriented people who can handle the stress that comes with the work. As modest as the relationship between personality and work performance is, managers regard personality to be as important as mental ability. This reliance on personality characteristics may not entirely be misplaced. Personality traits are useful predictors of a variety of outcomes that matter to managers including: avoiding counterproductive work behavior, reducing turnover and absenteeism, exhibiting teamwork and leadership, effective customer service, exhibiting citizenship behavior, increased job satisfaction and commitment, and enhanced safety (Gatewood et al., 2008). Tett et al. (1991) suggest that personality traits that are chosen on the basis of job analysis - clearly making a connection between the trait and the job functions - will show a trait-criterion relationship. In addition, an interesting finding in a study by Judge, Higgins, Thoresen, and Barrick (1999) showed that the effects of personality can compound over one’s career. In their study they report a composite validity of the Big Five traits at over .60 when predicting occupational success 30 to 50 years after assessing personality. Another reason managers like personality tests is because unlike cognitive ability measures, personality tests have never been shown to have adverse impact on any demographic group (Gatewood et al. 2008).

The work context or situation may call for a particular type of personality characteristic. For example, in a sales position, the worker is required to interact with and engage strangers in order to make a sale. Extraversion will be a trait that results in high sales performance for this position (Gatewood et al., 2008). In their study, Day & Silverman (1989) found that job related personality measures were significantly related to job performance ratings that were not predicted by cognitive ability measures alone. Mount and Barrick (1995), found Conscientiousness to be more related to motivational factors than ability factors in performance. Thus, the relationship between personality and performance becomes one of criterion dimensionality. In other words, using personality measures to predict traditional task related performance may not be appropriate. Perhaps what personality is predicting is not task performance, but contextual performance. If criteria do not reflect the appropriate performance dimension, personality measures may not be the most predictive. However, a unique dimension of performance that is not completely related to mental ability may be better predicted by personality measures.

Summary of Cognitive Ability and Personality

Cognitive ability tests are good predictors of job performance as determined by criterion measures that focus on task requirements. However, job performance can also be determined by "people requirements" (Day & Silverman, 1989). People requirements include things such as how a person works with others, how a person fits in with the organization's culture, and other contextual factors related to performance. Cognitive ability tests may not be the best predictors where people requirements are concerned. Instead, personality measures that directly relate to those characteristics which are important to the organization may be a better choice. Previous meta-analyses on the subject of personality and job performance missed this because they only looked at how personality relates to either subjective or objective measures. Because personality measures do not correlate highly with measures of task performance, perhaps a criterion measure that deals with contextual performance would show a higher relationship between personality and the criterion measure.

In comparison to cognitive ability tests, personality inventories have less validity, and less feasibility. However, personality inventories have no adverse impact (Gatewood et al. 2008), which is appealing to most managers, and they can predict those necessary people requirements for working in an organization. Also, the lack of correlation between personality and general mental ability means that personality traits may add to the validity of cognitive ability predictors of job success (Gatewood et al., 2008). There are problems associated with using either measure. Both kinds of tests show deficiency in their prediction of job success. However, the combination of the two is a way to reduce for the adverse impact caused by cognitive ability measures and improve the low validity of personality measures as well as include multiple facets of job success. No one measure can capture every aspect of success on the job. Organizations often use a combination of measures in order to tap into the various facets of job success. Using multiple measures accounts for deficiency and contamination found in a single measure and should result in higher empirical validity coefficients.

Performance Measurement

In his book, Thorndike (1949) said, “The most fundamental and most difficult problem in any selection research program is to obtain satisfactory criterion measures of performance on the job against which to validate selection procedures” (as quoted in Hoffman, Nathan, Holden, 1991, p. 601). There are many types of performance measures that can be used to rate an employee’s performance on the job. One distinction among performance measures is whether they are objective or subjective (Hoffman et al., 1991).

Objective measures such as sales numbers, number of goods produced, absenteeism, voluntary turnover, accident rates, salary history, promotions, and awards (Gatewood et al., 2008) are sometimes used as criteria for job success. However, these measures are deficient because they focus on such a narrow part of the position. Even a combination of various objective measures may still leave out important aspects of the job. These measures can also be contaminated. For example absence measures do not apply to some jobs, can be inaccurate, vary in their causes, and do not correlate with each other. Similarly, when using turnover as criteria, it is difficult to determine what turnover is voluntary and what turnover is involuntary (Murphy & Cleaveland, 1995). In addition, objective performance measures are often unavailable for many jobs. The many issues involved with using objective measures such as those mentioned have pushed the popularity of subjective measures.

Subjective measures, or performance ratings, are based on supervisory perception of an employee’s behavior and contribution to organizational goals. As previously mentioned above, overall performance ratings are one of the least susceptible measures to criterion deficiency (Brodgen & Taylor, 1950) because of their consideration of the entire job by and large. However, the issue faced when using multiple performance ratings is that the criterion can become deficient with the improper weighting of performance dimensions. For example, if you had 10 dimensions that an employee was rated on, one method of computing an overall score would be to average the raw scores. However, this approach would not accommodate real differences in the criticality of the different dimensions. What we would be doing in this instance is not weighting the dimensions equally, but weighting by the size of the standard deviation of each dimension; the larger the variability, the greater the weight of that dimension in the composite.

Subjective ratings are also susceptible to other errors such as leniency, halo and other errors which result in criterion contamination. Strauss, Barrick, and Connerley (2001) found that ratings were higher if a rater perceived the ratee to be similar in personality characteristics. Ferris, Yates, Gilmore, and Rowland (1985) found a rating bias against nurses between the ages of 40-61 in which older nurses where rated lower on performance than their younger counterparts.

Yet another reason for error in ratings may not be due to mistakes, but to the “rater’s rational pursuit of sensible goals” (Murphy & Cleveland, 1995). Murphy & Cleveland (1995) explain that leniency may not be a type of error but a type of behavior exhibited by raters that allows them to obtain rewards and avoid punishments. Several reasons why raters are reluctant to give low ratings include consequences for both the rater and ratee, avoidance or negative reactions, and maintaining the organization’s image. For example, in organizations where performance appraisals are used to provide a salary increase, a rater may purposely give higher ratings in order to be able to give a raise to a subordinate and avoid conflict. In addition, organizations rarely have systems in place that reward good raters and punish bad raters (Murphy & Cleveland, 1995).

Despite the risks, subjective performance measures remain popular because the errors can be minimized through rater training, use of multiple raters, and the use of multiple measures. Many studies have also found support for the use of subjective ratings (Hoffman, Nathan, & Holden, 1991; Conway & Huffcutt, 1997; Wall, Michie, Patterson, Wood, Sheehan, Clegg, & West, 2004). Hoffman, Nathan, & Holden (1991) found supervisor ratings to have significant validity. Conway & Huffcutt (1997) found supervisor rating mean reliability at .50 for single raters. The authors suggest use of the Spearman-Brown formula to determine how many raters would be necessary for higher reliability. Wall et al. (2004) found that judged against objective ratings, subjective measures showed a 95% success rate.

HYPOTHESES

The debate on situational specificity and validity generalization is one that is yet to be completely resolved. The goal of this study is to contribute to the literature by evaluating an approach to correct for criterion contamination and deficiency in criterion measures. Schmidt and Hunter propose that if 75% of the variability in validity coefficients could be accounted for by artifactual variables (i.e. sample size, range restriction, etc.) then the validity generalization hypothesis is supported. The rationale for the 75% figure is that it "corresponds to a correlation of .87…[this] would leave very little room for situational moderators to operate" (Pearlman, Schmidt, & Hunter, 1980, p. 383). However, Schmidt, Hunter and colleagues have not been altogether successful, accounting for roughly 60% of the variability in validity coefficients, across various types of jobs. If a method could be devised to include the criterion problem as one of the corrected artifacts, possibly the 75% figure could attained across a greater number of jobs.

The current study proposes a method of evaluating the impact of criterion contamination and deficiency in criterion measures. The study will employ a traditional subjective rating of current performance and a nontraditional index of promotability referred to as expandability, also subjective. This second criterion represents an assessment of fit in the organization and presents a different view of job success. The traditional subjective rating represents supervisory ratings of current performance. It is expected that by combining these two different aspects of performance, the resulting composite criterion will be less susceptible to contamination and deficiency issues and increase overall relevancy.

In the study, both cognitive ability measures and personality measures will used as predictors, and will be correlated with the criterion measures described above. Cognitive ability tests have been shown to be among the best predictors of job performance (Hunter & Hunter, 1984; Gatewood et al., 2008). In agreement with previous literature, this study hypothesizes the following:

Hypothesis 1

> Individual and linearly combined cognitive ability measures will be more strongly associated with the traditional current performance criterion measure than individual and linearly combined personality measures.

Job performance is a complex construct that is comprised of more than just task requirements. Ratings can be based on task requirements as well as people requirements including contextual factors related to performance (Day & Silverman, 1989). Ability measures may be good predictors of a person's ability to meet the task requirements, but may not be related to contextual factors. Instead, personality measures may better capture the contextual factors of performance (Day & Silverman, 1989; Mount & Barrick, 1995; Gatewood et al., 2008). Consequently personality measures are found to be related to job performance only in specialized situations where job analysis makes a clear connection between tasks and traits (Tett et al., 1991; Gatewood et al., 2008). Therefore, the current study hypothesizes the following:

Hypothesis 2

Individual and linearly combined personality measures will be more strongly associated with the expandability criterion measure than individual and linearly combined cognitive ability measures.

The current study attempts to minimize bias by accounting for contamination and deficiency through the use of multiple criteria. Figure 1 illustrates this point. The logic here is that by using multiple criteria, more of the ultimate criterion is included in the construction of actual criteria. Using different types of criteria is also critical to this argument. If the criteria used do not show a high correlation coefficient, one can assume that there are two different types of performance being assessed thus justifying the use of more area of the ultimate criterion in the figure. The more unique criteria used, the more of the criterion problem is addressed.

Figure 2 is the application of the CPR formula using the measures specific to this study. Relevancy of the measures is shown in stripes, deficiency in light gray, and contamination in dark gray. For both measures, adding another criterion adds more relevant variance than bias.

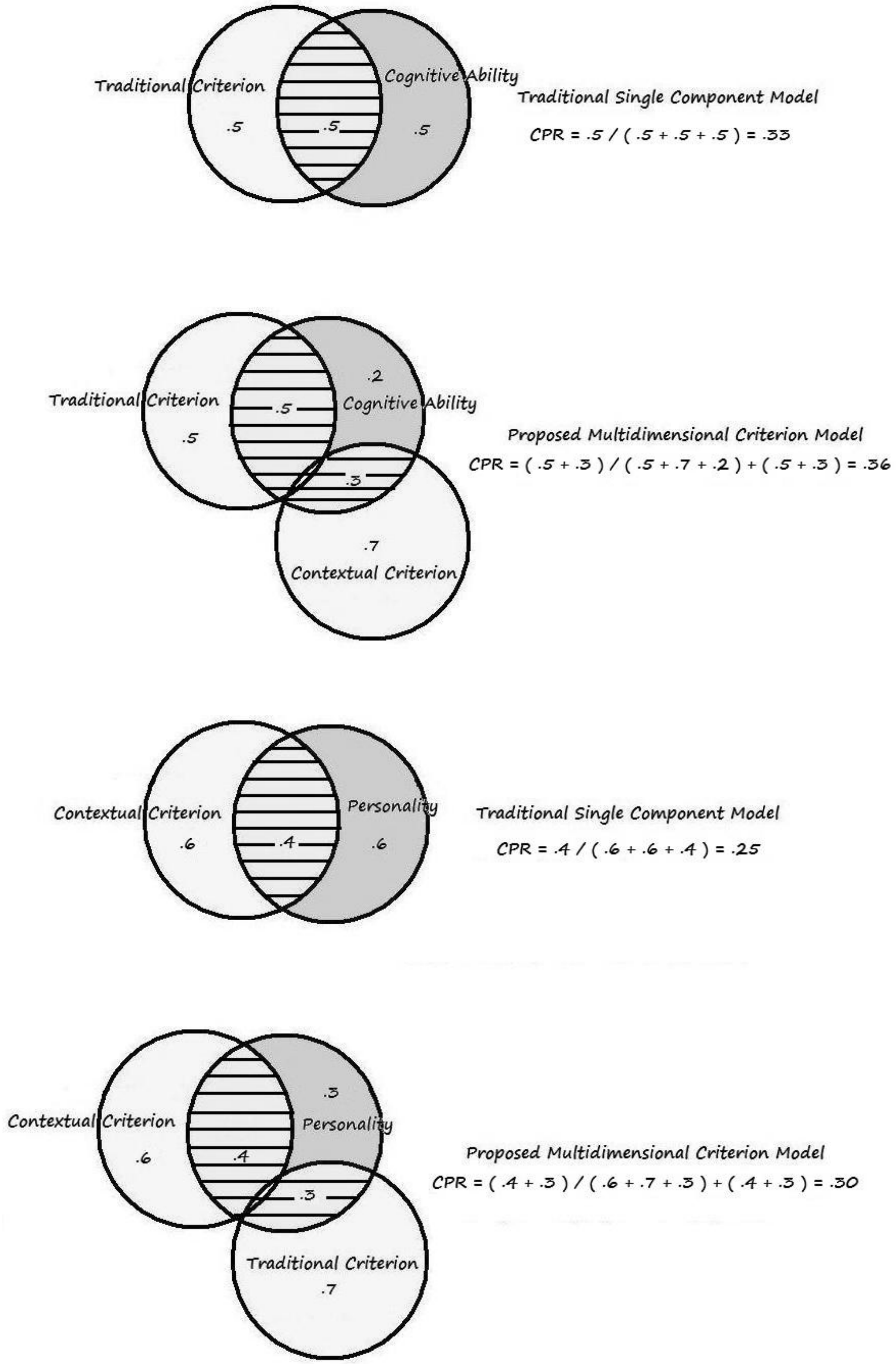
Traditional Criterion
Cognitive Ability
.5
.5
.5
Traditional Single Component Model
CPR = .5 / (.5 + .5 + .5) = .33
Traditional Criterion
.5
.5
.2
Cognitive Ability
.3
.7
Contextual Criterion
Proposed Multidimensional Criterion Model
CPR = (.5 + .3) / (.5 + .7 + .2) + (.5 + .3) = .36
Contextual Criterion
Personality
.6
.4
.6
Traditional Single Component Model
CPR = .4 / (.6 + .6 + .4) = .25
Contextual Criterion
.6
.4
.3
Personality
.3
Traditional Criterion
.7
Proposed Multidimensional Criterion Model
CPR = (.4 + .3) / (.6 + .7 + .3) + (.4 + .3) = .30

Figure 2 Multidimensional Criterion Models for Cognitive and Personality Measures

Dunnette (1963) argued that job success is a complex construct that cannot be measured using a single criterion. In accordance with this position, the current study will combine the two unique criteria, current performance ratings and expandability ratings, into one composite measure thus, taking into account different facets of job performance. The following is hypothesized:

Hypothesis 3

> Both individual and linearly combined cognitive ability measures and individual and linearly combined personality measures will be more strongly associated with the combined criterion than with either of the uncombined criteria.

The two criteria assess different facets of performance (Murphy & Shiarella, 1997). This implies that a composite measure of performance should yield a higher validity coefficient than the individual criteria.

In consideration of the methodology for this study, the Schmidt and Hunter meta-analysis procedure was not used. In order to replicate the meta-analysis and account for amount and kind of criterion contamination and deficiency, the studies included must all have used more than one predictor in validation, and have reported information on both predictors.

METHOD

Participants

Archival data was used in this study. Participants consisted of 378 employees from the Customer Care Division of a large international financial services institution with offices in the UK, Canada, and the US. All participants from this study were from the US and included positions ranging from entry level associates to senior management. Participants were hired on dates ranging from February, 2000 to September, 2005. Approximately 90% of the employees within the Customer Care Division participated in this study. Tenure ranged from less than one year to 5.67 years.

Procedure

Assessment data collected at the time of hire were correlated with performance data collected during a routine performance review of employees. These data were retrieved from organizational archives. Test score data and performance ratings correspond to the position currently held by each employee.

Instruments

Tests

Upon entry, participants were required to take a battery of tests administered by the organization. There are four cognitive ability measures and one personality measure with eight subscales.

The Gordon Personal Profile Inventory (GPP-I) test for 9 different traits including ascendency, responsibility, emotional stability, sociability, self confidence, cautiousness, original thinking, personal relations, and vigor. Respondents are presented with four statements and asked to indicate which statement is most characteristic of them and which statement is least characteristic of them. The self confidence scale is not used in this study for a total of eight subscales. Reliability estimates for the scales range from r = .88 to r = .89.

The Watson Glaser Critical Thinking Appraisal tests respondents' knowledge of five critical thinking skills including inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments. Reliability is reported at r = .73.

The Thurstone Test of Mental Alertness contains 126 items that test four job related areas including adjusting to new situations, learning new skills quickly, understanding complex or subtle relationships, and thinking flexibly. Respondents receive an L-score of linguistic or verbal ability, a Q-score of quantitative ability, and a T-score or total score composed of the sum of the L and Q-scores. Only the L score and Q score will be included in the analyses. Reliability estimates range from r = .92 to r = .95.

The Verbal Reasoning test is a cognitive ability instrument used to measure the capacity to reason logically based on verbal problems. This test is in paper and pencil format and consists of 12 problems ordered for difficulty. Test-retest reliability for this test is reported at r = .69.

Lastly, the Basic Skills Test is a tool used to measure fundamental (basic) skills such as reasoning, numerical ability, perceptual ability, and verbal ability for people in customer service, clerical and administrative positions. The complete measure consists of 15 tests that individually measure a specific skill.

Criteria

Supervisor ratings on performance were collected for each participant. Ratings were based on a 5 point Likert scale where 1 = “Needs Improvement”; 2 = “Sometimes Achieves”; 3 = “Consistently Achieves”; 4 = “Overachieves”; 5 = “Significantly Overachieves”. Performance ratings are a traditional measure of each employee’s current level of achievement in their position.

Supervisor ratings on expandability were also collected for each participant. Ratings for this criterion are based on a 4 point Likert scale where 1 = “Not Suitable”; 2 = “Well Placed”; 3 = “Expandable”; 4 = “Highly Expandable”. Expandability ratings are a non-traditional criterion measure that focuses on the employee’s fit with the cultural values of the organization. This financial institution regards themselves as young, fast, active, and more risk taking than their traditional and structured competitors. It is important for them that their employees can fit into this progressive culture that they believe is set apart from the rest. Ratings were given in consideration to whether the employee could easily move into other and higher areas of the organization.

A multidimensional criterion was created by combining the performance and expandability measures. In the absence of any other rationale for weighting, the two measures will be given a unit weight of one. To create the composite measure, scores for the single criteria will be converted to standard scores and averaged.

RESULTS

Table 1 presents the descriptive statistics for all variables included in the study. The four cognitive ability instruments were the Verbal Reasoning Test, Thurstone Mental Abilities Test, the Watson-Glaser Critical Thinking Appraisal and the Basic Skills Test. These four instruments yielded 5 scores, as noted in table 1. The trait measure used was the Gordon Personal Profile-Inventory, which yields 8 specific sub scores. These are also noted on table 1. Also included are the sample sizes (n), means, standard deviations, and minimum and maximum scores within this sample. A review of the means and standard deviations indicate that they are consistent with what would be expected with a sample of this level. For example, Corsini & Renck (1978) report a mean and standard deviation of 2.76 and 6.31, respectively, for executives and for middle managers, .23 and .40. The criterion measures are also included in this table.

Table 2 presents the intercorrelation matrix for all variables included in the study. Correlations significant at the .01 level are indicated by bold entries, those significant at the .05 level are indicated by italicized entries. "NA" entries indicate situations where correlations cannot be computed due to a lack of participants. Also, it should be noted that each correlation is based on a different number of participants, and the critical values change from cell to cell.

Table 1: Descriptive Statistics For All Study Variables

	N	*Min*	*Max*	*Mean*	*stdev*
Verbal Reasoning (VR)	131	4	35	23.47	7.06
Thurstone Mental Abilities - Language (TL)	60	17	63	36.88	9.49
Thurstone Mental Abilities - Quantatitive (TQ)	60	19	51	29.27	6.70
Watson-Glaser Critical Thinking (WG)	153	36	76	63.43	7.40
Basic Skills Test (BST)	49	4	17	9.63	2.89
Gordon Personal Profile - Ascendancy (GA)	245	15	33	25.26	3.24
Gordon Personal Profile - Responsibility (GR)	245	22	36	30.85	2.94
Gordon Personal Profile - Emotional Stability (GE)	245	15	40	27.43	3.75
Gordon Personal Profile - Sociability (GS)	245	12	34	22.67	3.49
Gordon Personal Profile - Cautiousness (GC)	245	2	35	24.79	4.00
Gordon Personal Profile - Original Thinking (GO)	245	19	39	31.96	3.13
Gordon Personal Profile - Personal Relations (GP)	245	15	38	28.29	3.93
Gordon Personal Profile - Vigor (GV)	245	17	40	30.87	3.74
Traditional Performance Rating (Perform)	378	1	5	3.42	0.72
Expandability Rating (Expand)	365	1	4	2.33	0.61

Table 2: Intercorrelation Matrix For All Study Variables

Study Variable	***Cognitive Ability Measures***					***Trait Measures***								***Criterion Measures***	
	VR	***TL***	***TQ***	***WG***	***BS***	***GA***	***GR***	***GE***	***GS***	***GC***	***GO***	***GP***	***GV***	***Perform***	***Expand***
Verbal Reasoning (VR)	---	na	na	**0.23**	na	0.05	-0.04	0.10	0.11	0.01	*0.18*	0.03	-0.01	-0.01	0.13
Thurstone Mental Abilities - Language (TL)		---	**0.88**	**0.84**	0.53	0.10	-0.17	0.01	-0.06	-0.04	0.07	0.09	-0.09	-0.03	-0.01
Thurstone Mental Abilities - Quantatitive (TQ)			---	*0.69*	**0.77**	0.07	-0.15	0.03	-0.01	0.03	0.04	0.12	-0.24	0.00	0.00
Watson-Glaser Critical Thinking (WG)				---	na	0.11	*-0.19*	0.09	-0.02	-0.03	0.15	-0.04	0.00	0.12	0.04
Basic Skills (BS)					---	-0.05	0.20	0.08	-0.34	0.13	0.17	-0.03	-0.04	-0.03	0.20
Gordon Personal Profile - Ascendancy (GA)						---	**-0.58**	**-0.43**	**0.51**	**-0.35**	**0.29**	**-0.20**	**0.31**	0.04	0.02
Gordon Personal Profile - Responsibility (GR)							---	**0.31**	**-0.47**	**0.36**	-0.07	*0.14*	0.08	0.01	0.00
Gordon Personal Profile - Emotional Stability (GE)								---	**-0.65**	0.15	-0.01	**0.33**	*-0.13*	*-0.14*	0.01
Gordon Personal Profile - Sociability (GS)									---	**-0.26**	0.02	-0.04	*0.15*	0.08	-0.08
Gordon Personal Profile - Cautiousness (GC)										---	**-0.26**	0.09	**-0.18**	0.04	0.01
Gordon Personal Profile - Original Thinking (GO)											---	**-0.21**	0.06	0.07	*0.13*
Gordon Personal Profile - Personal Relations (GP)												---	**-0.23**	-0.07	-0.03
Gordon Personal Profile - Vigor (GV)													---	0.04	0.04
Traditional Performance Rating (Perform)														---	**0.55**
Expandability Rating (Expand)															---

na - Cannot be computed because at least one of the variables is constant.

Bold Entry - Correlation is significant at the 0.01 level (2-tailed).

Italicized Entry - Correlation is significant at the 0.05 level (2-tailed).

The simple correlations indicate the relationships existing among the cognitive, trait and criterion measures. A review of the matrix indicates that 5 of the six cognitive intercorrelations, 20 of the 28 trait intercorrelations, and 2 of the 40 cognitive-trait incorrelations are significant at or beyond the .05 level. Across the criterion measures, only 2 of the 26 validity coefficients are significant at or beyond the .05 level. The intercorrelation between the two criteria is significant at the .01 level.

The relative high number of significant correlations is not a concern for the purposes of this study. In practical situations the magnitude of the validity correlations would be of prime interest, and, in most primary validity studies they would be corrected. Similarly, in meta-analyses and validity generalization studies, validity corrections would be adjusted for any number of sources of "error variance." However, these correlation coefficients have not been adjusted or corrected, and were not corrected because the main focus of the study is to review the pattern among the relationships, rather than the magnitude. The validity coefficients appearing on the far right side of the table will be compared to the multiple R validity coefficients presented in subsequent tables in this study.

Table 3 present the simple and multiple correlations for the cognitive ability measures. The premise of the argument being put forward is that the addition of a second and different criterion would at the very least decrease criterion deficiency and increase criterion relevancy. If the additional criterion contamination could be kept at a minimum, the ratio of criterion relevancy to criterion error (contamination + deficiency) would tilt in favor of greater predictability. In a specific study, increases in the multiple R over either of the simple r's indicate a reduction in the criterion problem.

Table 3: Simple and Multiple Correlations For Cognitive Ability Measures

	Simple Correlations		***Multiple R***
	Performance	*Expandability*	
Verbal Reasoning	-0.01	0.13	**0.15**
Thurstone Mental Abilities – Language	-0.03	-0.01	**0.09**
Thurstone Mental Abilities – Quantitative	0.00	0.00	**0.06**
Watson-Glaser Critical Thinking	0.12	0.04	0.12
Basic Skills Test	-0.03	0.20	**0.26**

Reviewing this table presents several interesting results. For, for 4 of the 5 cognitive measures the multiple R exceeds the simple r for both criteria. Collectively, this supports the notion that the addition of the second criterion minimizes, to some extent, the criterion problem. Albeit the correlations are small, the pattern exists. Of the five measures, 2 favor the expandability criterion, basic skills and verbal reasoning respectively, and only 1 favor the performance criterion, the Watson-Glaser. Again, the differences are small, but these data do not lend support to hypothesis 1 which states that cognitive ability measures will be more strongly associated with the traditional current performance criterion measure than personality measures when using uncombined criteria. A test for the difference between dependent correlations revealed that none of the correlations above are significantly different from each other.

Table 4 presents similar information for the trait measures. Of the eight measures, 3 multiple R's exceed the simple r's for both criteria including emotional stability, sociability and responsibility, respectively. Collectively, this does not clearly support the conclusion that the addition of a second criterion minimizes the criterion problem. Five of the 8 trait measures favor the performance criterion while only 1 favors the expandability criterion. This finding does not lend support to hypothesis 2 which states that personality measures will be more strongly associated with the expandability criterion measure than cognitive ability measures when using uncombined criteria. A test for the difference between dependent correlations revealed that none of the correlations on table 4 are significantly different from each other.

Table 4: Simple and Multiple Correlations for Trait Measures

	Simple Correlations		***Multiple R***
	Performance	*Expandability*	
Gordon Personal Profile - Ascendancy	0.04	0.02	0.03
Gordon Personal Profile – Responsibility	0.01	0.00	**0.02**
Gordon Personal Profile - Emotional Stability	-0.14	0.01	**0.16**
Gordon Personal Profile – Sociability	0.08	-0.08	**0.15**
Gordon Personal Profile – Cautiousness	0.04	0.01	0.04
Gordon Personal Profile - Original Thinking	0.07	0.13	0.13
Gordon Personal Profile - Personal Relations	-0.07	-0.03	0.07
Gordon Personal Profile – Vigor	0.04	0.04	0.04

Table 5 presents the multiple R's for the various combinations of the predictor-criterion groups, which represent the major thrust of the research. The criteria were defined as the performance score, the expandability score, and a unit weighted combined of the two. Unit weighting was used in the absence of any a priori rationale for differential weighting. The predictor groups included the cognitive measures, trait measures, and both sets of measures. It should be noted that due to a high instance of missing data, not all of the cognitive measures could be included in the multiple regression analyses. In this instance, cognitive measures include the Thurstone Mental Abilities - Language, Thurstone Mental Abilities - Quantitative and the Watson-Glaser Critical Thinking Appraisal.

Table 5: Multiple Correlations For All Predictor/Criterion Combinations

	Criterion Measure		
	Perform	*Expand*	*Combined*
Cognitive Measures	0.28	0.09	0.25
Trait Measures	0.19	0.18	0.18
Both Measures	0.26	0.18	0.23

In all three cases, the multiple R using the combined criterion failed to exceed the multiple R's using the other two criteria. Therefore, these data do not support hypothesis 3 which states that cognitive ability measures and personality measures will be more strongly associated with the combined criterion than either of the uncombined criteria. However, an interesting pattern emerged. The traditional performance criterion resulted in the strongest multiple R's across the cognitive, combined, and trait criteria, respectively. This represents strong support for hypothesis 1. Also, when looking at the expandability criterion, the multiple R for the trait measures exceeded the multiple R for the cognitive measures, suggesting support for hypothesis 2.

DISCUSSION

In regards to the hypotheses predicted, the outcomes were as follows:

Hypothesis 1 predicted that cognitive ability measures would be more strongly associated with the traditional current performance criterion measure when using uncombined criteria. This prediction was supported. Table 5 shows that cognitive ability measures had a multiple R of .28 when analyzed with the traditional performance measure, and .09 when analyzed with the expandability criterion. Also, the cognitive ability measures had a stronger relationship with the traditional performance criterion than the personality (trait) measures and traditional performance (R = 0.19).

Hypothesis 2 predicted that Personality measures would be more strongly associated with the expandability criterion measure when using uncombined criteria. This hypothesis was partially supported. Personality measures had a stronger relationship with the traditional measure than the expandability measure. However, as expected, personality measures were more strongly associated with the expandability criterion than the cognitive ability measures.

Finally, hypothesis 3 predicted that cognitive ability measures and personality measures would be more strongly associated with the combined criterion than either of the uncombined criteria. This hypothesis was not supported. Although differences were minimal, table 5 shows that the predictors have a stronger relationship to the single criteria than the combined criterion. This suggests that combined criteria do not appreciably reduce the criterion problem. Table 3 does provide somewhat partial support for hypothesis 3 at the individual test level. This however, is not the goal of the present study as the focus is on the battery of tests and the combined criteria.

This main goal of the current study was to find a way to address the criterion problem using multiple criteria. This study did not find the desired effect. There are several reasons why this result occurred. First, the expandability criterion brought as much error into the equation as it did relevance. This is observed in the lack of increase in multiple R suggesting that the second criterion did not add relevance, or had as much relevance as the first criterion. Secondly, it was questionable as to how the expandability relates to the job. It may be that expandability is not related to the job, but to different jobs within the organization. Hence, the expandability criterion may be linked to an organizational need, but not the job. Thirdly, the criteria in this study did not seem to predict success on the job and particularly for this study, the criteria did not reduce the criterion problem. In other words, there was no increase the shared variance to total variance ratio (CPR). Lastly, the data used in this study was archival data. There were many missing values resulting in a variety of sample sizes for the tests that were run. In particular, table 5 only included three of the five cognitive ability tests, the Thurstone Mental Abilities - Language, Thurstone Mental Abilities – Quantitative, which had the lowest sample sizes, and the Watson-Glaser Critical Thinking Appraisal which showed the lowest correlation to the criteria. Methodological changes for future studies include adding another criterion with a logical link to the job as well as a more complete data set in which all participants provide a score in all measures.

This study aims to advise researchers regarding what information should be included in future studies so that criterion problems (deficiency and contamination) can be codified. Studies using multiple criteria which do not exhibit an increase in the multiple R of combined criteria over single criteria are essentially “adding” more error variance than criterion covariance, which may keep the CPR ratio the same. Studies that show such an increase are either adding more criterion covariance, adding less error variance, or some combination of the two. Collecting information about such studies, and grouping such studies into categories, will allow us to assess the impact of multiple criteria on the criterion problem, and reported validity coefficients. To this end, researchers are requested to minimally include the following:

1. Use multiple criteria with rational relationships to overall job performance whenever possible.
2. Report results for criteria individually and in combination when possible.
3. Report simple correlation matrices of all study variables, including individual criteria. These data can then be used to “estimate” CPR improvements over single criteria.

While we have developed a conceptual model for the CPR ratio, development of a statistical model is needed. Variance components can easily be collected after the fact, but a statistical model is needed to predict CPR effects prior to conducting the study and while considering criteria. Again, other researchers are called upon to assist in this effort.

Using a similar process as that of correcting for sample size, range restriction, and criterion and predictor unreliability, criterion contamination and deficiency can be corrected for thereby making the conclusions of validity generalization more viable.

REFERENCES

Austin, J.T., & Villanova, P. (1992). The criterion problem. *Journal of Applied Psychology, 77*(6), 836-874.

Barrick, M.R., & Mount, M.K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., and Macknzie, S.B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Brodgen, H.E., & Taylor, E.K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement, 10*, 159-186.

Bruning, J.L., & Kintz, B.L. (1968). *Computational Handbook of Statistics*. Glenview, IL: Scott, Foresman, and Company.

Conway, J.M., Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta analysis of subordinate, supervisor, peer, and self ratings. *Human Performance, 10*(4), 331-360.

Corsini, R.J., and Renck, R. (1978). *Verbal Reasoning Interpretation and Research Manual*. Park Ridge, IL: London House Inc.

Day, D.V., & Silverman, S.B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25-36.

Dunnette, M. (1963). A note on the criterion. *Journal of Applied Psychology, 47*(4), 251-254.

Ferris, G.R., Yates, V.L., Gilmore, D.C., & Rowland, K.M. (1985). The influence of subordinate age on performance ratings and causal attributions. *Personnel Psychology, 38*, 545-557.

Gatewood, R.D., Feild, H.S., & Barrick, M. (2008). *Human resource selection* (6th ed.). Mason, OH: Thomson.

Gordon, L.V. (1993). *Gordon personality profile – inventory manual*. San Antonio: The Psychological Corporation.

Hoffman, C.C., Nathan, B.R., & Holden, L.M. (1991). A comparison of validation criteria: Objective versus subjective measures and self versus supervisor ratings. *Personnel Psychology, 44*, 601-619.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72-98.

Hurtz, G.M., & Donovan, J.J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85*(6), 869-879.

James, L.R., Demaree, R.G., & Mulaik, S.A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology, 71*(3), 440-450.

Judge, T.A., Higgins, C.A., Thoresen, C.J., & Barrick, M.R. (1999). *Personnel Psychology, 52*, 621-552.

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

Morgeson, F.P, Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K., & Schmidt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Morgeson, F.P, Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K., & Schmidt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029-1049.

Mount, M. K., & Barrick, M. R. (1995). The big five personality dimensions: Implications for research and practice in human resources management. In G. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 13, pp. 153-200). Greenwich, CT: JAI Press.

Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal based perspectives*. Thousand Oaks, CA: Sage.

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.

Ones, D.S., Dilchert, S., Viswesvaran, C., & Judge, T.A. (2007). In support of personality assessment in personnel settings. *Personnel Psychology, 60*, 995-1027.

Outtz, J.L. (2002). The role of cognitive ability tests in employment selection. *Human Performance, 15*(1/2), 161-171.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for test used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*(4), 373-406.

Ruch, W.W., Shub, A.N., Moinat, S.M., Dye, D.A. (1982). *Administrator's guide: PSI basic skills tests for business, industry, and government*. Los Angeles: Psychological Services, Inc.

Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*(5), 529-540.

Schmidt, F.L., Hunter, J.E., & Caplan, J.R. (1981). Validity generalization results for two groups in the petroleum industry. *Journal of Applied Psychology, 66*(3), 261-273.

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

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

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton Mifflin Co.

Society for Industrial and Organizational Psychology (2003). *Principles for validation use of personnel selection procedures.* (4th ed.). Bowling Green, OH: Author.

Strauss, J.P., Barrick, M.R., & Connerley, M.L. (2001). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology, 74*, 637-657.

Tett, R.P., & Christiansen, N.D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmidt. *Personnel Psychology, 60*, 967-993.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-752.

Thurstone, L.L., Thurstone, T.G. (1998). *Thurstone test of mental alertness: Examiner's manual*. Minneapolis: NCS Pearson.

Wall, T.D., Michie, J., Patterson, M., Wood, S.J., Sheehan, M., Clegg, C.W., West, M. (2004). On the validity of subjective measures of company performance. *Personnel Psychology, 57*, 95-118.

Watson, G., & Glaser, E. (1964). *Watson-Glaser critical thinking manual*. New York: Harcourt Brace Jovanovich, Inc.

Van Iddekinge, C.H., & Ployhart, R.E. (2008). Developments in the criterion-related validation of selection procedures: a critical review and recommendations for practice. *Personnel Psychology, 61*(4), 871-925.