

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

EXPANDING THE PREDICTOR AND CRITERION SPACE TO REDUCE
ADVERSE IMPACT IN A PUBLIC SECTOR ENVIRONMENT

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

JUAN ENRRIQUE BENAVIDEZ

Norman, Oklahoma

2005

UMI Number: 3163310



UMI Microform 3163310

Copyright 2005 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

EXPANDING THE PREDICTOR AND CRITERION SPACE TO REDUCE
ADVERSE IMPACT IN A PUBLIC SECTOR ENVIRONMENT

A Dissertation APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Shane Connelly PhD

Robert Terry PhD

Jorge Mendoza PhD

Michael Buckley PhD

Daniel Wren PhD

Acknowledgements

The writing of a dissertation can be a lonely and isolating experience, yet it is obviously not possible without the personal and practical support of numerous people. However, there is one person above all others who deserves my deepest thanks and respect for her continued support during the writing of this dissertation: my wife, Teri Dawn Benavidez. I could not have done it without her.

My children, Adaira Lanae, Garion Del, and Avery Ryland, have been my greatest source of inspiration. Being tired of missing them, their goodnight kisses, their wonderful stories, smiles, their little voices, being tired of not being able to fulfill their requests when I wanted to so much are the best motivations I know for finishing a dissertation. This work would not have been possible without them.

To my father I owe thanks for life, support, and the beliefs and values that he has instilled in me though patient, unconditional love has made everything in my life possible.

Graduate school is a trial of fire, and rightfully so, to ensure that those who become doctors are motivated, skilled professionals, however, there were times the fire became a little too hot. Professors Jorge Mendoza and Mike Buckley through kind words and witty ‘one-liners’ made a world of difference in finishing this journey with my head high. I owe you both tremendous thanks, because without those kind words and ‘one-liners,’ I don’t think I’d be here today.

Professor Wren, you are just an awesomely kind and wonderful man. I never knew either of my Grandfathers, but I would like to think they would be a lot like you. Knowing you would be at my defense provided a calming effect that was much appreciated. Also, I think you were right, “it was the cookies.”

To Professor Craig Russell, I owe a great deal. You taught me many things about being a researcher, an applied professional, and man. Thank you for sharing your wisdom, and giving me the courage to say, “I’m sorry. I’m confused, please help me understand.”

Professor Mike Mumford, I owe thanks for teaching me how to be the best researcher that I can be, to understand the many intricacies that are involved in all things, and how to work really, really, really hard.

To Joe Davenport and Natasha Riley I owe a tremendous amount of thanks for allowing me the chance to develop my craftsmanship skills in my chosen field of endeavor. I especially owe Natasha Riley additional appreciation for championing my cause and ensuring that this study actually became a reality, and her support and helpful advice.

My graduate studies would not have been the same without the social and academic challenges and diversions provided by all my student-colleagues in the Department of Psychology of the University of Oklahoma. I am particularly thankful to my friends Greg Manley, Tommie Mobbs, Helen Fung, Jamie Culp, and Whitney Fauth. We not only studied, relaxed, and traveled well together, but they were even willing to read some portions of this dissertation and thus provided some very useful input.

Finally, this dissertation would not have been possible without the expert guidance of my esteemed advisors, Professors Mary Shane Connelly and Robert Terry. Not only were they readily available for me, but also they always read and responded to the drafts of each stage of my work more quickly than I could have hoped. Their oral and written comments were always extremely perceptive, helpful, and appropriate. Furthermore, their proactive acceptance to be Co-Major Professors for my Dissertation will forever be an act of compassion and understanding that I hope to carry forward to others. Thank you.

Of course, despite all the assistance provided by Professors Connelly, Terry and others, I alone remain responsible for the content of the following, including any errors or omissions, which may unwittingly remain.

Table of Contents

I.	Introduction	1
	Personnel Selection and Fairness	2
	Public Sector Issues	6
	Purpose of this Study	8
II.	Literature Review	
	Job Knowledge	9
	Adverse Impact and Job Knowledge	11
	Cognitive Ability	11
	Adverse Impact and Cognitive Ability	12
	Interpersonal Skills	14
	Adverse Impact and Interpersonal Skills	15
	Ratings of Training and Experience	16
	Job Experience	18
	Adverse Impact and Job Experience Measures	21
	Education	22
	Education and Adverse Impact	23
III.	Managing the Adverse Impact/Criterion Validity Tradeoff	
	Alternative Predictor Models	24
	Expanded Job Performance Domain	27
IV.	Study Hypotheses	30
	Validity	31
	Sub-Group Differences	32
V.	Method	
	Sample	34
	Measures	35
	Analyses	38
VI.	Results	39
	Validity	41
	Subgroup Differences	44
VII.	Discussion	45
	Future Directions	48
VIII.	References	50
IX.	Tables	55
Appendix A	Sample Training and Experience Measure	66

Abstract

This study is an attempt to reduce sub-group differences, while maintaining validity, by using alternative selection devices commonly seen in the public sector and expanding the criterion space by incorporating service-oriented behaviors. Participants were correctional security officers hired between 1999 and 2000 in a Midwestern state. Logistic regression results indicated that it is possible to significantly reduce sub-group differences and have greater validity for a composite of alternative predictors (i.e., Job Knowledge, Interpersonal Skills, Training and Experience) than a composite made up of specific job-related cognitive skills (i.e., Basic Math, Reading Comprehension, Attention to Detail, Memory Recall, Logic, Best English Usage, and Report Writing). Overall, this study provides encouraging results and information for selection system decision-makers in the public sector.

Currently, personnel selection practitioners face the dilemma of choosing between the oft-competing goals of organizational productivity optimized by rank ordering job applicants' scores on valid selection procedures and equal subgroup representation (Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). Research has consistently shown cognitive ability measures have high criterion validity. However, cognitive ability measures show large mean standardized differences for protected sub-groups that will typically lead to adverse impact (Bobko, Roth, & Potosky, 1999). Therefore, researchers and practitioners have advocated various approaches to maintain reasonably high validity and lessen adverse impact. Two approaches are use of alternative predictors that show less or no adverse impact in conjunction with cognitive ability tests and expanding the definition of good performance to include more than task outcome.

To date, there have been few studies examining these strategies to balance the validity/equal representation tradeoff. Furthermore, there is little published empirical research of the criterion validity of many selection devices seen in public sector hiring (e.g., ratings of training and experience and job knowledge tests) and even less examining the adverse impact of these devices. Consequently, there is little empirical evidence to assist public sector personnel professionals in making decisions about which selection instruments to use in making hiring decisions. This study examines the criterion validity/adverse impact trade off for cognitive ability and alternative selection assessments, as well as expanding the criterion space to include service behaviors that might be particularly useful in the public sector environment.

Personnel Selection and Fairness

One of the primary purposes of employment selection procedures is to identify the person with the highest job qualifications to meet particular organizational needs. From the point of view of practical value, the most powerful indicator available to the personnel specialist is predictive validity (Schmidt & Hunter, 1998). The ability to predict future job performance, success in job-related training, and long-term developmental opportunities has enormous impact on the economic well being of every individual. Collectively speaking, it can have a major impact on every organization and the nation as a whole (Wagner, 1997).

With uncorrected criterion validity coefficients in the mid .30s, one of the most valid predictors of job and training performance is a measure of general cognitive ability (Hunter & Hunter, 1984; Ree & Earles, 1992; Schmidt & Hunter, 1998). Additionally, decision-makers can use the same cognitive ability measure across jobs due to the high generalizability of cognitive ability (Hunter & Hunter, 1984; Ree & Earles, 1992). Another benefit of cognitive ability tests is that they are easier to develop and administer than alternative selection tools with similar criterion validity; such as trainability tests, work sample tests, biographical measures, and assessment centers (Reilly & Warech, 1993). Furthermore, significant research has demonstrated that cognitive ability measures do not show differential prediction (Schmidt, Gast-Rosenberg, & Hunter, 1980; Hirsh, Northrup, & Schmidt, 1986). Differential prediction, or predictive bias, refers to a situation where using a test to predict a criterion of interest results in systematic over-prediction and under-prediction based upon group membership. When differential prediction occurs, it is

possible that applicants with significantly different predictor scores could perform equally well on the job. The presence of differential prediction is determined via multiple-regression techniques. If one regression line represents the relationship between predictor and criteria for all groups then the test does not display differential prediction (Cleary, 1968). Although there are many definitions of test fairness, this standard of test bias, referred to as the Cleary Model of Test Bias, is the most accepted (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). Because cognitive ability tests, in general, display one regression line for all groups, cognitive ability tests are unbiased.

Although, cognitive ability tests are unbiased in a statistical sense, there are serious concerns regarding the perceived fairness of cognitive ability tests by society. The question of test bias is a technical question that psychometricians determine empirically. Fairness, on the other hand, is a dynamic concept that reflects the current values and beliefs of society (e.g., changing stance towards affirmative action programs). Norms, customs, and ultimately the law communicate the values of society. The Civil Rights Act of 1964 prohibited unlawful discrimination in hiring practices based upon sex, race, or ethnic origin and created the Equal Employment Opportunity Commission (EEOC) to ensure compliance of the Act. In 1978, the EEOC, the Departments of Justice and Labor, and the Civil Service Commission jointly published the *Uniform Guidelines in Employee Selection (Guidelines)* to assist test makers in development of valid selection devices. Although the *Guidelines* are not law, courts give them “great deference” when determining guilt in unlawful

discrimination cases. Section 4D of the *Guidelines* specify that adverse impact exists when “the selection rate for any race, sex, or ethnic group is less than four-fifths (4/5 or eighty percent) of the rate for the group with the highest selection rate. Federal enforcement agencies will regard violation by organizations of this four-fifths rule as evidence of adverse impact.”

When discussing possible adverse impact, there are two indicators that are important, the standardized ethnic group mean score difference (the difference in means divided by the sample-weighted average of the group standard deviations) and the selection ratio. Of the two, the focus should be on reducing sub-group differences in the selection procedure because the selection ratio is determined by organizational need and situation specific factors. Therefore, this study will focus on reducing sub-group differences rather than adverse impact per se.

Cognitive ability tests have consistently shown a standardized mean score difference (d) between Blacks and Whites of 1.0 and a $d = .83$ between Hispanics and Whites (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Violation of the four-fifths rule can expose an organization to employment law litigation at potential great expense. Currently, the state of Alabama is spending between 20 and 25 million dollars per year to deal with adverse impact compliance issues (Personal Communication Maury Buster, November 2002). In Fiscal year 2004, the EEOC recovered more than \$251 million through pre litigation resolutions (conciliation, mediation and other administrative settlements), and \$168 million through agency lawsuits filed in federal district court for a combined total of \$420 million, the most monetary benefits ever obtained by EEOC in a single year (EEOC, 2005).

In addition to legal pressures to reduce adverse impact, there is also the question of the social impact of inferences made from cognitive ability test scores. “Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and *actions* based on test scores” (Messick, 1995, emphasis added). Because decision-makers rely on inferences from test scores to make high-stakes decisions, Messick (1989) advocated analysis of societal impact from inferences made from test scores be an integral part of test validation. Messick (1995) suggested reevaluation of test use and even abandonment of the test if significant negative societal consequences occur from use of said test. Cronbach (1988) shared Messick’s concerns, “...validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences” (cited in Messick, 1989). Even though cognitive ability tests are unbiased, if they were to become the selection device of choice, as some advocate (Schmidt & Hunter, 1998; Ree & Earles, 1992), their use over generations to determine who is allowed entrance into educational gifted programs, undergraduate and graduate colleges, and to make employment and promotional decisions may likely cause Asians to be the premier social economic group, with Whites slightly behind. Hispanics would become second-class citizens and Blacks a distant third. The long-term consequences of using cognitive ability tests as the premier test for access into the more lucrative levels of American society would be disastrous.

Public Sector Issues

In addition to legal mandates and validity considerations, for public sector employers there is a tremendous social burden to ensure equitable treatment of its citizens. Often, equitable treatment includes representation in public sector jobs representative of the populace of the state. In the public sector, specific federal, state, and local laws merge with strong public expectations to require tremendous levels of openness, fairness, and accountability in all aspects of testing (Wiesen, Abrams, & McAttee, 1994). These expectations have produced a system that allows multiple opportunities for applicants to appeal the selection process. Typically, applicants are allowed to appeal decisions, scores, subject matter, and individual items of the selection process. Therefore, examinations have to possess a high degree of face validity for applicants to “buy into” the appropriateness of the subject matter. In fact, most states require, through statute, that civil service examinations test for knowledge, skills, and abilities that are required to perform the job for which they are applying and that can be practically and reliably measured. Because of this legal requirement and large numbers of jobs with few incumbents, test developers in the public sector have come to depend upon the content validation strategy of test development. This reliance on content validity has resulted in highly face valid tests that mirror KSA requirements listed in the job description.

Public sector employers also face an economic constraint on the number and type of selection devices used to make selection decisions (Wiesen, Abrams, & McAttee, 1994). It is not uncommon for public sector employers to receive hundreds of applications for a small number of available positions. Coupled with little money

due to budget constraints, public sector employers are dependent upon machine-scored, written, multiple-choice tests for most civil service examinations. Assessment devices such as empirically keyed biodata questionnaires (which require large sample sizes to develop) and assessment centers (which are quite expensive to develop and administer) are not feasible for a large number of jobs. Instead, one is more likely to encounter measures of experience and education (known as training and experience measures or T&Es), job knowledge, and specific cognitive abilities (e.g., English usage and math skills), or interpersonal skills as selection devices of choice.

Unfortunately, these constructs have been of little interest to researchers and therefore their relationship to other indicators and job performance outcomes is not well understood. For example, there have only been three published articles examining criterion validity of the T&E method. In addition to little empirical examination of the criterion validity of these constructs, there has been little published examination of the adverse impact of these measures.

Because citizens feel that they are legitimate owners of the government, there is an expectation of high customer service from public sector employees. In fact, oftentimes the citizen places a higher value upon being treated courteously than a favorable outcome. With private sector organizations becoming increasingly customer service conscience to ensure that they can compete globally, public sector customers have begun to demand more. The public sector has recognized this rise in public expectation and has begun to act accordingly. First, many public agencies now include contextual performance domains such as customer service, teamwork, and leadership in supervisor performance appraisals. Second, public sector agencies have

begun to experience competition, which used to not be the case. For instance, recently a large federal agency changed its policy to allow local support divisions to compete nationally with other local support divisions. Such competition will no longer allow the lackadaisical public sector employee to continue to hold service-behaviors in low esteem.

Purpose of this Study

There is a serious effort towards development of alternative strategies to select and rank applicants without inequity between hiring ratios. Schmitt, Rogers, Chan, Sheppard, and Jennings (1997), described the current dilemma for personnel professionals as the conflict between the goals of organizational productivity optimized by rank ordering job applicants' scores on valid selection procedures and the goal of equal subgroup representation. Many researchers (Aguinis, Cortina, & Goldberg, 1998; Chan & Schmitt, 1997; Hattrup, Rock, & Scalia, 1997; Kulik, Kulik, & Bangert, 1984) have examined various strategies to solve this dilemma. To date, researchers have modified individual's scores through within-group norming, examined eliminating biased items, provided coaching, used low test cutoff scores, and used multiple types of banding strategies to reduce adverse impact (Sackett & Ellingson, 1997; Sackett, Schmitt, Kabin, & Ellingson, 2001; Ryan, Ployhart, Greguras, & Schmit, 1998). However, these strategies have been of little use, are short-term fixes, and/or have been declared illegal. Two other strategies examined by researchers are the modification of the predictor-criterion relationship and modification of the format of the test, both of which have shown some success.

The rationale for the modification of the predictor-criterion relationship method is simple and straightforward. Although cognitive ability is important, it is not the only determining factor of good job performance. If job analyses results indicate that certain job knowledge, skills, and abilities (KSAs) are required for successful job performance, then measures of those KSAs should be utilized in the hiring process—a typical content validity strategy. Furthermore, today a significant proportion of the U. S. economy is service-orientated rather than manufacturing-oriented as it had been in the past (Department of Labor, 2002). In service-oriented organizations, customer service, teamwork, and other prosocial behaviors are a business necessity (Hough, Oswald, & Ployhart, 2001). In order to more accurately reflect the needs of modern organizations, researchers and personnel professionals have begun to redefine what constitutes good performance.

The current study is an attempt to maximize criterion validity while minimizing adverse impact by expanding the predictor space (e.g., using other measures in addition to cognitive ability) and criterion space (e.g., inclusion of service-oriented behaviors in performance appraisal) in a public sector environment. This study examines the expansion of the predictor-criterion relationship in a public sector correctional security officer position.

Job Knowledge

Job knowledge is the “culmination of facts, principles, concepts, and other pieces of information that are considered important in the performance of one’s job” (Dye, Reck, & McDaniel, 1993). By definition, job knowledge is job-related because the universe being measured is the knowledge structure communicated to test

developers from successful job incumbents and their supervisors. This strong theoretical relationship between test domain and test items leads to high predictive criterion validity coefficients with desired criteria (i.e., job performance). Because of the high content validity of job knowledge tests and the reasonableness of their use, there has been relatively little empirical research investigating the effectiveness of job knowledge tests.

Hunter and Hunter (1984) presented the results of two unpublished studies Hunter (1982) and Dunnette (1972). Hunter (1982) found job knowledge tests to have high criterion validity with supervisor ratings ($\bar{r} = .48$) and work-sample performance ($\bar{r} = .78$). Meta-analysis of Dunnette's (1972) data indicated that job knowledge has a criterion validity of .51. Likewise, Hunter (1983) indicated that the major impact of cognitive ability on job performance is indirectly through its influence on the acquisition of job knowledge ($r = .61$). He found that job knowledge significantly predicted job performance as measured by the work sample exercise ($r = .55$), and supervisor's ratings ($r = .34$).

Schmidt, Hunter, and Outerbridge (1986,) following suggestions by Guion (1983), expanded Hunter's 1983 model to include job experience. They found similar results to Hunter (1983), with job knowledge mediating the relationship between cognitive ability and job performance, as well as the relationship between job experience and job performance. Job experience was the best predictor of job knowledge ($r = .57$), followed by cognitive ability ($r = .46$). Job knowledge in turn predicted work sample performance ($r = .66$) and supervisor ratings quite well ($r =$

.34). It appears that Job Knowledge generally has equal to, or greater, validity, both content and criterion, than g.

Adverse Impact and Job Knowledge

A literature review gleaned only two studies examining the adverse impact of job knowledge measures. One, (Kesselman & Lopez, 1979) indicated that job knowledge tests do not show adverse impact, however, the sample sizes in this study were very small (24 Whites and 28 minorities). Even more problematic is the fact that the minority sample included 11 Asians, who typically perform better than Whites on cognitive ability tests. Therefore, it is quite possible that the Asians “boosted” the group mean in this experiment.

In the second study (Wing, 1981) showed that the group means between Whites and Blacks on a law enforcement promotional test were not significantly different. However, she states that the first Black on a top-down list of ranked applicants was in the tenth position. Therefore, given top-down selection, nine Whites had to be hired before the first Black could be hired.

Currently, there is not a reliable answer to the amount of adverse impact caused by job knowledge tests and there are no studies, that I am aware of, that list the standardized ethnic group mean difference on these types of tests.

Cognitive Ability

General Cognitive Ability is a specific scientific term used in place of what the general population would call intelligence. “Cognitive ability allows reasoning, planning, problem-solving, abstract thinking, the understanding of complex ideas, rapid learning and learning from experience. It is not simply encyclopedic

knowledge, a particular academic ability or expertise for doing tests; rather, it reflects a wider and more profound capacity for understanding one's environment – awareness, making sense of things, or imagining what should be done" (Maranon & Pueyo, 2000). Cognitive Ability predicts many beneficial outcomes such as successful job performance (Hunter & Hunter, 1984), successful scholastic performance (Cronbach & Snow, 1977), social status, total income, and many other correlates of a "successful" life (Jencks, 1979). To the general populace the statement that "smart people tend to do well" would likely receive an, "Isn't that obvious" response. After all, the concept of intelligence is very basic. However, it was only in the early 1980s that researchers had the evidence to back up that claim. Until then, the findings on the criterion validity of cognitive ability were inconsistent. Research over the last two decades has supported the generalizability of cognitive ability in predicting job and training performance (Schmidt & Hunter 1998, Gottfredson, 2002) across all jobs and types organizations. The most recent meta-analytic uncorrected estimate of criterion validity between cognitive ability and job performance is .30 (Bobko, Roth, & Potosky, 1999).

Adverse Impact and Cognitive Ability

Most researchers agree that the concept of cognitive ability, or intelligence, is a three level structure. At the apex is a common factor of human cognitive abilities that tends to account for a sizeable portion of variance in test scores. This "common factor" is what is referred to as general mental ability or "g." The second level is comprised of specific cognitive abilities (e.g., verbal ability, math ability) that load onto the general factor "g." However, there is still much debate on the specific types

and number of specific abilities. Finally, the third level is composed of specific mental processes (e.g., processing speed, short term recall). Roth, Bevier, Bobko, Switzer, and Tyler (2001) point out that selection systems operate at different levels of the cognitive ability hierarchy and therefore it is important to have a precise understanding of standardized ethnic group differences at each level. Table 1 (Hough, Oswald, & Ployhart, 2001) attests to the sentiments expressed by Roth et. al. (2001).

Insert Table 1 about here

Cognitive ability, g , has consistently shown a standardized subgroup mean difference (d) equal to 1 between Blacks and Whites, between Hispanics and Whites $d = .5$, and between East Asians and Whites $d = -.2$, in the East Asians favor. However, for specific cognitive abilities (verbal, quantitative), the d values are lower than the d value for g . Furthermore, the d values for mental processes are also lower than those for g . Therefore, one may reduce the amount of adverse impact by using a measure of specific abilities required to perform the job as dictated by the job analysis.

The current Black-White difference is an old phenomenon that has been observed for many years (Cleary, 1968; Guion, 1966). The effect of a one standard deviation difference is very dramatic. As Hunter and Hunter (1984) point out, “if a test is used to select at a level equivalent to the top half of white applicants, it will select only the top 16% of black applicants.” Obviously, such a dramatic difference in hiring rate will always lead to a prima facie case of adverse impact according to the

four-fifths rule described in the *Uniform Guidelines for Employee Selection Procedures* (EEOC, 1978).

Interpersonal Skills

There are very few jobs that operate within a vacuum. Employees work in an organization that requires them to communicate with others, coordinate their actions, and follow instructions (Van Scotter & Motowidlo, 1996). Because jobs require employees to interact with one another, interpersonal skills play an extremely important role in organizational cohesion and productivity. Therefore, it is quite reasonable to measure applicants' interpersonal ability in a selection system.

In a generalizability study of cognitive ability with law enforcement jobs (Hirsh, Northrop, & Schmidt, 1986), results indicated that cognitive ability is less effective at predicting job performance ($r = .13$) than usually seen with other occupations. The authors felt that human relations skills are very important in law enforcement occupations. This hypothesis is supported by the result that the highest correlate of job performance in this study was a model that included human relations skills ($r = .31$). Ferris, Witt, and Hochwarter (2001) found an even greater relationship between interpersonal skills (or social skills) and job performance in a study of 126 software engineers. Correlations between interpersonal skills and job performance dimensions ranged between .19 and .23. Goldstein, Yusko, and Nicolopoulos (2001), looking at Black-White differences in assessment centers found that an interpersonal competency (i.e., human relations) moderately correlated ($r = .16$) with overall job performance. It appears that interpersonal skills can be an effective member of a predictor set.

Adverse Impact and Interpersonal Skills

To date there has been very little research examining Black-White score differences in interpersonal skills measures. Much of the work that has been conducted has taken place in the assessment center literature. Research has shown that assessment center exercises tend to vary in cognitive load required to perform the exercise (Russell, 1987). Exercises requiring large amounts of information processing, decision making, and reading and writing (e.g., in-basket exercise) place a higher cognitive load upon the test-taker, and therefore, should lead to higher observed standardized differences. On the other hand, exercises requiring more interpersonal types of skills and behaviors (e.g., role playing exercises) place a lower cognitive demand upon the test-taker and should lead to lower standardized group differences. Studies by Goldstein and colleagues (Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Goldstein, Yusko, & Nicolopoulos, 2001) have found that exercises that are interpersonal do not display significant standardized Black-White mean differences ($d = .02$ and $.03$, respectively).

Because interpersonal skills are a large component of the structured interview (Huffcutt, Conway, Roth, & Stone, 2002), it seems reasonable to extrapolate information from that research stream into decisions regarding interpersonal skills Black-White differences. Huffcutt and Roth (1998) conducted a meta-analysis to assess the effect of race on employment interview evaluations across 31 studies. They found standardized mean Black-White and Hispanic-White differences of $.25$ and $.26$, respectively. Schmitt, Clause, and Pulakos (1996) also found Black-White and Hispanic-White standardized differences to be quite small ($ds = .15$ and $.19$,

respectively). However, Roth, Van Iddekinge, Huffcutt, Eidson, and Bobko (2002), found that correcting for range restriction of applicant-level data led to $d = .36$ and $.54$ with two different structured behavioral interviews. It appears that the d value for interpersonal skills may be underestimated and more research may be necessary.

Ratings of Training and Experience

Ratings of training and experience (T&Es) are a unique product of civil service test developers during the beginning of the 20th century (Abrams, Johnson, Sproule, & Ash, 2002). Developers intended to generate a less subjective process to analyze applicant's background information and generate a score similar to an exam score, thereby allowing screening of professional level applicants without hindering recruitment through a formal testing process.

There has been a paucity of criterion validity studies of the effectiveness of various T&E methods (Ash & Levine, 1985; McDaniel, Schmidt, & Hunter, 1988). McDaniel, et al. (1988) comment that Mosel's (1952) study is the earliest published study found with criterion validity coefficients and that they were unable to locate any other published study with correlation coefficients. Ratings of training and experience have to be the most understudied personnel selection instrument. This is quite disturbing when one considers that public sector employers use them quite extensively. According to a recent study (IPMA, 2002), 70% of public sector employers surveyed used T&Es. Furthermore, 59% felt it was the most effective method to identify well-qualified applicants.

Mosel (1952), examined the criterion validity of T&E ratings for 13 Civil Service jobs (see table 2). Results of Mosel's (1952) study indicate that overall there

is very little or no relationship between rational ratings of training and experience and supervisory judgment. However, Mosel's sample size for each job analyzed was quite modest; sample sizes ranged from 51 to 125.

Insert Table 2 about here

McDaniel, Schmidt, and Hunter (1988) performed a meta-analysis on the one published T&E criterion validity study (Mosel, 1952) and several other non-published criterion validity studies. McDaniel, et al's, (1988) meta-analysis of 91 criterion validity coefficients obtained an uncorrected mean criterion validity coefficient of .06 for the point method of T&E ratings across a heterogeneous sample of data. However, McDaniel et al. (1988) found there was considerable unexplained variance in their analysis due to four sources, 1) heterogeneity of constructs measures, 2) heterogeneity of measurement process, 3) heterogeneity of occupational groups, and 4) low number of analyzed criterion validities, which led to second-order sampling error. Second-order sampling error occurs when there are a small number of studies from which population estimates are derived in meta-analytic procedures. Schmidt, Hunter, and Outerbridge (1986) argue that relative differences in mean job experience lead to individual differences in job performance. Therefore, as the sample mean level of experience increases, the criterion validity of job experience is expected to decay. Because point method scoring strategies give substantial weight to the amount of job experience, it is possible that criterion validity of point method evaluations vary with mean level of job experience in the study. To test this hypothesis, McDaniel, et al.

(1988) reanalyzed data from the Molyneaux (1953) study. Results from the moderation test indicate that higher grade levels, which correspond to more job experience, have lower obtained criterion validity coefficients. McDaniel, et al. (1988) concluded that the point method yields low criterion validity, but may show useful levels of criterion validity for samples with low mean experience levels.

Some possible reasons why T&Es are still used so extensively is that they 1) are easy to develop, 2) have high content validity when developed correctly, 3) are perceived to be fair, and 4) are less likely to “drive-away” applicants who are not willing to take tests (Abrams, Johnson, Sproule, & Ash, 2002). Furthermore, one has to assume that there is some contribution made to the decision-making process or their use would have discontinued a long time ago. The consistent finding of low to moderate levels of criterion validity seems odd when one considers the fact that a significant portion of the T&E score is based upon job experience and education both of which have displayed high and moderate criterion validities respectively.

Job Experience

Researchers have shown that job experience is highly correlated with practical job criteria such as job knowledge ($r = .83$), work samples ($r = .47$), and supervisor ratings ($r = .60$) (Schmidt, Hunter, Outerbridge, & Goff, 1988; Vance, Coover, MacCallum, & Hedge, 1989). In fact, Vance, Coover, MacCallum, and Hedge (1989) found “Experience played a larger role than aptitude in determining task proficiency levels” (pg. 453). However, the relationship between job experience and job performance is not a linear relationship. Instead, it is a monotonically increasing negatively accelerated curve (Avolio, Waldman, & McDaniel, 1990; Schmidt,

Hunter, Outerbridge, & Goff, 1988), such that, in general, the predictive criterion validity of job experience sharply declines as time progresses. There is disagreement as to how long experience is a valid predictor, with the low end being five years (Schmidt, Hunter, Outerbridge, & Goff, 1988) and the upper end being 20 years for more complex jobs (Avolio, Waldman, & McDaniel, 1990).

Many of the studies that have concluded that experience does not predict performance past five years have used incumbents and their current level of job experience in a concurrent criterion validity studies. Concurrent criterion validity procedures as a method are reasonable to investigate theoretical relationships when a purely predictive study is not feasible. However, for personnel selection the question of concern is “Is an applicant with more job-experience more likely to be successful in the job than an applicant without job-experience?” In practical terms it comes down to having employees that require little to no training and are able to “hit the ground running.”

Furthermore, there are many possible reasons why longer tenure would lead to decreasing performance for incumbents within the same organization. One possible reason is that employees become complacent. They first start the job with enthusiastic vigor to meet probationary requirements and to fit-in. However, as they learn how to “game” the system there is a decrease in effort put forth such that performance decreases to a meets standards level. Second, sometimes the new employee does not decrease performance, but rather maintains that high level of vigor required to maintain an “exceed standards” level. Those types of employees tend to be promoted up and out of the current job in which they reside. Subsequently, their performance is

not included in criterion validity studies (Avolio, Waldman, & McDaniel, 1990).

Third, it is possible that poor compensation and reward systems that do not recognize effort, or even worse reward poor performers based on other non-job related criteria, could lead longer tenured employees to decrease their efforts (Cosier & Dalton, 1983). A fourth possible reason is ceiling effects. Performance appraisals have a ceiling, often an “exceeds standards” or “excellent” rating. It is possible that incumbents with more experience reach this ceiling quickly and cannot go any higher in terms of performance ratings. One would expect those types of folks to be promoted or to continue to receive the highest ratings allowed in the performance management system, hence the asymptotic effect that is often seen. Therefore, these types of studies may contaminate the predictor-criterion relationship.

In addition to job experience having a non-linear relationship with job performance measures, experience is a multidimensional construct (Quinones, Ford, & Teachout, 1994). Consolidating all types of experience and examining the relationship between experience and job performance has resulted in loss of information and led to conservative estimates of the criterion validity of experience. Quinones, et. al. (1994) hypothesized that measurement mode and level of specificity would moderate the relationship between work experience and job performance. Specifically, amount measures of experience have the highest correlation with performance because they focus on what a person is actually doing rather than simply how long they have been doing it. Level of specificity refers to the level of analysis from which experience measures were harvested; levels are organizational, job, and task.

According to results of their meta-analysis, across all studies, experience is a good predictor of performance ($\bar{r} = .22$). Additionally, there is a direct relationship between level of analysis and criterion validity of the predictor-criterion relationship. Specifically, criterion validity increases as one increases specificity of analysis from the organization level down to task level ($\bar{r} = .16, .22, \& .34$), with population parameter estimates showing larger “true” relationships. Furthermore, as predicted, the relationship between experience and job performance increases as the measurement of experience progresses from type, to time, to amount ($\bar{r} = .19, .22, \& .36$). Quinones, et al. (1994) effectively showed that experience measures can predict performance quite well, especially if the measure of experience focuses on amount of tasks performed and hard performance data is the criterion.

Adverse Impact and Job Experience Measures

A search of the relevant literature domains did not reveal a single study examining the possible adverse impact effects of job experience measures. However, given the unequal opportunity that Blacks used to encounter in access to employment opportunities, it may be possible that there is some form of adverse impact for experience measures. One would expect this effect to decrease and hopefully disappear in the next 10 to 20 years, due to the strident diversity initiatives that American businesses, governments, and educational systems have initiated voluntarily and involuntarily.

Education

Criterion validity studies examining the relationship between education and job performance are scarce. Arvey (1979) suggested that it is difficult for most organizations to conduct empirical criterion validity studies with educational requirements because educational degree requirements are part of established selection systems making it difficult to find individuals without the required education. Use of educational requirements is based on the assumption that attainment of an educational degree indicates one's ability to succeed in training programs and/or acquisition of required job knowledge. Additionally, attainment of certain degrees (e.g., engineering, chemistry, medicine) indicates applicant exposure, and competence in, job required skills and knowledge.

Meritt-Haston and Wexley (1983) conducted a meta-analysis of the relationship between educational requirements and job performance. Data for their study came from 12 studies obtained from professional journals, as well as members of the Academy of Management and the American Psychological Association. The occupational groups examined in the studies were sheriffs, secretaries, refinery technicians, and salespersons. Criteria were supervisor ratings, organizational tenure, and number of promotions. Meritt-Haston and Wexley (1983) computed an average correlation with sample-size weightings for each criterion validity coefficient; however, they made no corrections for range restriction, sampling error, or measurement error. Table 3 shows the results of their analysis. Generally, educational requirements predicted job performance measures quite well. Hard measures of performance were better predicted by educational requirements than soft measures.

The lower correlation for ratings is most likely due to unreliability and subjectivity typically inherent in performance ratings.

Insert Table 3 about here

Hunter and Hunter (1984) reported findings from several unpublished studies, summarized in table 4. The results displayed in Hunter and Hunter (1984) give further evidence that educational requirements do indeed predict job performance at a moderate level, with the exception of the Dunnette (1972) results.

Insert Table 4 about here

Education and Adverse Impact

A literature search did not uncover any studies examining the adverse impact of educational requirements. However, Meritt-Haston and Wexley (1983) discussed several court cases regarding educational requirements. Therefore, adverse impact has occurred in the past, the question becomes is it still present today? Given almost 40 years of effort into reducing official barriers into educational opportunities, I would like to believe that the *d*-values for education are very small to non-existent. Additionally, given another 25 years, the differences should be zero. This believe was echoed in the recent Supreme Court decision (*Grutter v Bollinger*) where the court stated that it, "...expects that 25 years from now, the use of racial preferences will no

longer be necessary to further the interest approved today.” Unfortunately, removing formal barriers does not guarantee that informal barriers are not still present.

Given the high criterion validity seen in job experience measures and moderate, but practical, levels of criterion validity seen in educational measures, it seems odd that the T&E point method yields such low criterion validity coefficients.

Managing the Adverse Impact/Criterion Validity Tradeoff

Alternative Predictor Models. As early as 1984, Hunter & Hunter offered the following suggestion to reduce adverse impact, “the most likely successful approach to reduced adverse impact is not through the discovery of substitutes for ability tests—there is no known test that approaches cognitive tests in terms of criterion validity for most jobs—but through the discovery of how best to use alternative measures in conjunction with cognitive ability tests.”

In one of the first studies examining the subgroup differences and criterion validity issue, Pulakos and Schmitt (1996) used an alternative predictor-model composed of biodata, a structured interview, and a situational judgment test to reduce subgroup differences. Pulakos and Schmitt (1996) collected data as part of a concurrent criterion validity study conducted to validate entry-level selection measures for a federal investigative job. There were 464 participants, of which 259 were White, 100 Black, and 97 Hispanic. Predictors were a verbal ability measure (verbal analogies, work knowledge, and reading comprehension), a 29-item biodata measure, a 33-item situational judgment test (problems involving the need to plan, organize, get along with others, adapt, and make judgments), and a structured experienced-based interview. The criterion for their study was a supervisor-rating

questionnaire developed to measure core investigative proficiency and effort and professionalism. Core investigative proficiency explained 54% of the variance and represents the technical “can do” aspects of the job. The second factor, effort and professionalism, explained 11% of the variance and represents the motivational “will do” aspects of the job. Results indicated that cognitive ability significantly predicted core investigative ratings ($r = .19$), but did not significantly predict effort and professionalism ratings ($r = .04$). Addition of the non-cognitive alternatives significantly increased criterion validity for both criteria, $R = .44$ for core investigative ratings and $R = .41$ for effort and professionalism ratings. Furthermore, the verbal ability measure had a Black-White $d = 1.03$ and a Hispanic-White $d = .78$. Whereas, a composite of verbal ability and the three predictors had a Black-White $d = .63$ and a Hispanic-White $d = .48$. Interestingly, the model consisting of only the three non-cognitive predictors had a Black-White $d = .23$ and a Hispanic-White $d = .16$. Therefore, addition of cognitive ability did not increase criterion validity significantly, but did increase subgroup difference markedly.

Pulakos and Schmitt (1996) effectively demonstrated that alternative predictor models could achieve criterion validity higher than cognitive ability, while at the same time reducing adverse impact to minority subgroups.

With a Monte Carlo simulation based on meta-analytic evidence, Schmitt, Rogers, Chan, Sheppard, and Jennings (1997) examined the effects of four factors, 1) number of predictors, 2) predictor intercorrelations, 3) criterion validity of the predictors, and 4) level of predictor subgroup differences, on the level of subgroup differences associated with the cognitive predictor model and the model of alternate

predictors (i.e., predictor model without cognitive ability). Predictor models were composed of the three non-cognitive low-adverse impact predictors (structured interview, conscientiousness, and biodata) coupled with and without cognitive ability.

Schmitt, et al's. (1997) results indicated that the smallest subgroup differences are observed when there are three alternate predictors each with zero subgroup differences, high criterion validity ($r = .30$), and high levels of intercorrelations (.50). However, even in this instance, d is not zero, but rather .26. Additionally, criterion validity in this particular scenario was .38 with cognitive ability included and .37 without. The highest criterion validity for the three-predictor model occurred when the predictor intercorrelations equaled .00, predictor criterion validity equaled .30, and there was zero subgroup difference. In this scenario, uncorrected criterion validity equaled .60 with cognitive ability and .52 without it.

Schmitt, et al. (1997), made the following comments to summarize their findings. First, the addition of three alternate predictors to cognitive ability will substantially reduce subgroup effect sizes, but will by no means remove the possibility of violating the four-fifths rule. Second, subgroup differences remain high over a broad range of reasonable levels of number of predictors, predictor intercorrelations, predictor validities, and levels of alternate predictor subgroup differences. Third, addition of alternate predictors whose subgroup differences are near zero and whose validities and intercorrelations are high will result in the lowest adverse impact levels. Fourth, manipulation of predictor intercorrelations causes the greatest concern for the trade-off between equal subgroup representation and high predictive power. With high intercorrelation (.50), the effect size is minimized but the

multiple correlation is much lower than it would have been with low intercorrelation (.00). Last, when the job analysis suggests that constructs other than cognitive ability are important to job performance, it would almost never be advisable to use only cognitive ability measures as the basis for selection decisions.

Sackett and Ellingson (1997) suggested researchers and practitioners following this strategy should be cognizant of some important issues for models where predictors are equally weighted. First, addition of a predictor uncorrelated with cognitive ability and free of adverse impact ($d = 0$) to a cognitive ability predictor, which shows high adverse impact ($d = 1$) will not result in a $d = .5$. Instead, for the model a $d = .71$ will result. Second, addition of a second measure that correlates moderately with cognitive ability may result in a larger d than either measure alone if the second measure has a $d > .6$. Third, the largest decrease in d occurs with the addition of the first measure and then rapidly decreases for each additional measure where each additional measure has $d = 0$.

Expanded Job Performance Domain

Many (Borman & Motowidlo, 1993, 1997; Campbell, McHenry, & Wise, 1990; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Pulakos & Schmitt, 1996) have advocated the necessity of expanding the definition of job performance to include factors other than task competency. This concept is intuitively appealing as organizations value more than “what” employees do, organizations also value “how” employees do their jobs; especially, in a service-based economy where customer service and teamwork behaviors are vitally important to organizational success. As

Hough, Oswald, and Ployhart (2002) point out “predictors derive their importance from their ability to predict criteria of interest to the organization.”

Campbell, McHenry, and Wise (1990) tested two hypotheses regarding differential relationships between predictors and criteria. First, model scores measuring cognitive abilities would be most useful for predicting scores on two task performance factors, core technical proficiency, and general soldiering proficiency. Second, a temperament/interest/reward model would be the best predictor of non-task outcome performance (e.g., personal discipline and physical fitness/military bearing).

The data confirmed both hypotheses; with the cognitive ability model better predicting core technical proficiency ($R = .65$ vs. $.44$) and general soldiering proficiency ($R = .69$ vs. $.44$). Furthermore, the temperament/interest/reward model was the better predictor of personal discipline ($R = .35$ vs. $.17$) and physical fitness/military bearing ($R = .38$ vs. $.23$). Furthermore, the non-cognitive model better predicted effort and leadership than the cognitive model ($R = .38$ vs. $.32$, respectively).

Motowidlo and Van Scotter (1994) tested the usefulness of the distinction between task performance and contextual performance by examining these domains' relationship with overall contribution to the organization and individual differences in ability, personality, experience, and training performance. The authors hypothesized that the results should show that task performance and contextual performance contribute independently to judgments of overall organizational contribution and that they are correlated with ability, personality, experience, or training performance in different ways. Table 5 displays the study results.

Insert Table 5 about here

Strangely, cognitive ability was an ineffective predictor of task performance ($r = -.01$). All personal characteristic significantly predicted contextual performance, however, some personality variables predicted contextual performance more effectively ($r_{\text{work orientation}} = .36$, $r_{\text{dependability}} = .31$, $r_{\text{cooperativeness}} = .22$, $r_{\text{internal control}} = .26$). Task performance was predicted well by years of experience ($r = .34$) and moderately by work orientation and dependability ($r = .23$ and $.18$). Hierarchical regression analysis indicated that task performance explained 13% of the variance in overall performance beyond that explained by contextual performance and that contextual performance explained 11% of the variance beyond that explained by task performance. These results support the hypothesis that task and contextual performance contribute independently to overall performance. The authors comment that the most straightforward implication of their study is that job performance as measured by supervisor ratings is not a unidimensional construct. There are at least two dimensions, task and contextual performance, that have different antecedents and contribute independently to supervisors' judgments about subordinate's performance.

From the studies reviewed, it is apparent that job performance encompasses more than task proficiency. Therefore, it is reasonable, if not required, that criteria definitions incorporate contextual performance.

Study Hypotheses

This study examines the strategy of expanding the predictor set and the criterion space in an attempt to maintain acceptable criterion-validity while simultaneously minimizing adverse impact. The study is conducted in the public sector and therefore faces the constrictions normally imposed in this type of environment (e.g., extreme openness and fairness, high demand placed on face validity, large budget constraints).

When developing a predictor model for a public sector job, it is important to select test domains and items that have high face validity, such that the applicant can see commonsensically why he or she is being tested on the material. For instance, applicants can comprehend why a tax enforcement specialist has to be able to add, subtract, multiply, and divide or why a customer service representative has to possess interpersonal skills. Public sector applicants have a harder time understanding the relationship between responses on sequential digit items to performance as a child welfare specialist. Therefore, from the point of view of reducing applicant grievances, use of only job-related specific cognitive ability components is preferred to overall esoteric measures of general cognitive ability.

Additionally, to ensure adequate coverage of required KSAs, test developers should measure all important KSAs that add unique variance to the selection equation. There are many jobs within both the private and public sectors that require specific KSAs (e.g., specific knowledge domains, personal characteristics, etc.), it is therefore quite reasonable, if not required, to hire based upon possession of these KSAs.

In this study, there are three predictor models of interest. The Full Model—composed of specific cognitive abilities, job knowledge, interpersonal skills, previous job experience, and previous job training. The g Model—composed of the specific cognitive ability measures math skills, memorization skills, reading comprehension, English usage, report writing, attention to detail, logic, and reading comprehension. The Alternative Model—composed of job knowledge, interpersonal skills, previous job experience, and previous job training. There are two dimensions of interest in this study; criterion validity and standardized ethnic group mean differences resulting from use of each model. Hypotheses are discussed within each dimension.

Validity

Full Model vs. g Model and Alternative Model. Although cognitive ability is important to successful job performance, other constructs are also related to job performance. Research has shown that these other constructs have significant levels of criterion validity and are not perfectly correlated with cognitive ability. Therefore, criterion validity for a predictor model containing specific measures of cognitive ability, job knowledge, previous training, previous experience, and interpersonal skills (Full Model) will be significantly higher than the criterion validity for a model composed of only specific cognitive abilities (g Model) or a model composed of only non-cognitive abilities (Alternative Model). Because R^2 will always increase with the addition of more uniquely contributing variables to a regression equation, the statement that Full Model will have higher criterion validity than g and Alternative Models is not profound. It is the comparison between the g Model and the Alternative Model that is of interest in this study.

g Model vs. Alternative Model. There is high correlation between cognitive ability and job knowledge ($r = .61$, Hunter, 1983). Furthermore, job knowledge often predicts job performance better than cognitive ability (Ree, Carretta, & Teachout, 1995; Schmidt, Hunter, & Outerbridge, 1996). Therefore, if one were to use job knowledge in a predictor model along with other non-cognitive predictors that display moderate (.15) to good (.25 and higher) validities, one would expect the job knowledge model (Alternative Model) to have a greater validity than a model composed of only specific cognitive-ability measures (g Model).

Hypothesis 1A: Alternative Model will have a larger adjusted R^2 value than g Model.

Hypothesis 1B: Adding Alternative components to a g predictor model will result in greater incremental validity compared to adding a g component to a predictor model containing Alternative predictors (i.e., Job Knowledge, Interpersonal Skills, and Training and Experience).

Task performance vs. Task and Service performance. Throughout this paper there is the assumption that increasing the criterion space by including job-related service behaviors will lead to increased validity. Therefore it is necessary to test this assumption.

Hypothesis 1C: For each predictor model, a model that contains only Task or Service performance will have a lower validity coefficient than one that contains both.

Sub-Group Differences

Alternative Model vs. Full Model and g Model. Given that the Alternative model does not contain cognitive ability domains, it is safe to articulate that, of the three models, the Alternative model will show the smallest subgroup effect.

Hypothesis 2A: Alternative Model will have a lower d -value than Full Model and g model

Full Model vs. g Model. It is difficult to derive hypotheses when very little is known about the standardized ethnic group differences of many of the members of the model. Currently, there is little to no information about standardized ethnic group difference effect sizes of job experience, job knowledge, or education. We do know that differences for specific cognitive abilities are less than that typically seen for generalized mental ability. Furthermore, we know that interpersonal skills, generally, display little to no group differences. For job knowledge, one could assume that there are large group differences because job knowledge is highly correlated with cognitive ability; however, this does not take into account job experience. Given time, employees who are lower in cognitive ability can become as proficient as those with higher cognitive ability. Therefore, cognitive ability may determine how quickly one gets up to speed, but with enough time, others can rise to the same level of proficiency. One cannot be sure how large the standardized group difference is for the job knowledge domain.

It is possible to estimate the standardized group difference for *g* Model using information from Sackett and Ellingson (1997). To do this, one needs to know the average intercorrelation among predictors, number of predictors, and sum of d . Given that there is a high intercorrelation among math skills, reading ability, and memorization (Wechsler, 1958) one could estimate the average correlation = .50. Using the results listed in Hough, Oswald, and Ployhart (2001), a sum of $d = 1.8$ for math skills, reading comprehension, and memorization ($d = .7, .6, \& .5$, respectively) is obtained. Using three for the number of predictors, a $d = .87$ to 1.15 is derived for the *g* Model.

For the Full Model this is a much more difficult task as the d -values for job knowledge, previous job experience, and education is unknown. Assuming that the d -value for the g Model is approximately equal to 1, a $d = .03$ for interpersonal skills, and zero intercorrelation between the two, an estimate of $d = .71$ is derived. As discussed in earlier sections, I anticipate job experience and education to have d – values close to zero. Given this reasoning, it is expected the d – value for the g model to be greater than the d – value for the Full model.

Hypothesis 2B: Full Model will have a lower d -value than g Model.

Method

Sample

The current study is a concurrent validation study involving 619 Correctional Security Officers (CSOs) hired in a Midwestern state correctional agency between January 1, 2000 and December 31, 2001. There were 463 Whites (75%), 58 Blacks (9%), 19 Hispanics (3%), and 79 Native Americans (13%). The sample contained 462 males (75%) and 157 females (25%). Demographic data were obtained from application records. The participant average age was 32.7 years. Previous experience for applicants was wide ranging, with some having previous law enforcement experience, many having previous military experience, and a significant number having previous truck driving experience, other than that there was no identifiable previous experience trends apparent. Forty-seven applicants had at least a baccalaureate degree at the time of application; the remaining applicants had at least a General Educational Development degree (GED).

Measures

Predictors

Paper-and-pencil Test. All participants were hired using the same selection process. Specifically, participants all had to meet minimum qualifications and take a paper-and-pencil test that was given in a structured setting by a trained test administrator. The job-analysis derived paper-and-pencil test is a 120-item test composed of three sub-tests measuring specific cognitive abilities (i.e., Basic Math, Reading Comprehension, Attention to Detail, Memory Recall, Logic, Best English Usage, and Report Writing), job knowledge (i.e., Correctional Principles, Legal Terminology, Safety, and Mechanics.), and interpersonal skills. There are 68 items measuring the various specific cognitive abilities ($\alpha = .83$), 43 items measuring job knowledge ($\alpha = .80$), and 9 items measuring interpersonal skills ($\alpha = .47$).

Rating of Training & Experience. With the assistance of subject matter experts, I developed a rating of training and experience (T&E) to evaluate the desirability of prior training and experience possessed by current incumbents. Ratings were made from the employee's original application blank. The T&E, a rationally scored weighted application blank, provides credit for previous education, certificates earned, and previous relevant job experience (See Appendix A for a copy).

The most points for education (120) are given for a Baccalaureate degree in a law enforcement related field (e.g., Criminal Justice, Police Science, etc.) or a behavioral science field (Sociology, Psychology). The Baccalaureate degree is the highest degree given credit because SMEs indicated that applicants generally do not have graduate level degrees. One hundred points are given to Baccalaureate degrees

in non-law enforcement or behavioral science fields and sixty points are given to applicants that have completed an Associates degree in a law enforcement related field from a community college or votechnical school.

The most valued certification (40 points) is a graduation certification from another state's correctional academy, which is followed by CLEET certification (30 points), which is followed by the Reserve CLEET certification (15 points). SMEs felt that up to fifteen years of experience was valuable because it takes that long to become fully trained in all of the odd atypical situations that occur in a correctional setting.

Experience was divided into "Same," "Similar," and "Related" experience, which is based on the degree of overlap between the duties of the current job and previous jobs. An additional category of "Other" was added because SMEs indicated that retired law enforcement applicants are frequent and are typically valued employees. Credit for the various categories are as follows; 2 points per month for Same experience, 1.5 points per month of Similar experience, .5 points for related experience, and 1 point per month for Other experience.

SMEs felt that education was important, but much less than experience; therefore experience is worth twice as much as education. Furthermore, education is valued three times as much as certification. The sum of Experience, Education, and Certification scores is the overall T&E score.

Criteria

Supervisor Performance Ratings. All CSOs are rated by their supervisor at least once a year. Supervisors use the Performance Management Process (PMP) to

rate CSOs. The PMP is made up of two sections, one concerning task performance and one concerning behaviors engaged in while conducting tasks and extra-role behaviors. The task section is made up of accountabilities with attached performance standards. Each correctional facility has developed a set of tasks that are encountered by CSOs in their facility to rate CSO job-performance in that facility. Many of the tasks are common to all facilities (e.g., conducting inmate counts, observing inmates, conducting searches, etc.); however, there are some unusual tasks that are facility specific (e.g., some facilities have gun-tower procedures, some have CSOs conduct mail search, etc.). The number of tasks per PMP ranges between four and eight tasks. The behaviors listed in the behavior section are the same for each facility and are as follows: Customer Service, Teamwork, Problem-Solving, and Observing Work/Leave Policies. Each task and behavior is rated from “Does Not Meet Standards” at the bottom end of the scale to “Exceeds Standards” at the top. Values were assigned to the ratings in the following manner, Does Not Meet Standards = 0, Needs Improvement = 1, Meets Standards = 2, and Exceeds Standards = 3.

Academy Grades. All newly hired Correctional Security Officers (CSOs) have to attend a six-week training academy, but not necessarily before beginning facility duty. While at the academy, CSOs take quizzes and tests that are graded and used to determine if the employee has mastered the information. The final grade given to each CSO is the measure of academy performance. Employees that quit the academy were given a score of zero. Average academy grade equaled 93% (SD = 3.03) and ranged between 82% and 99%.

Analyses

Typical descriptive statistics were calculated for each variable in the study.

Furthermore, correlations between each variable were also calculated.

Validity

Hierarchical regression analysis will be used to calculate the criterion validity of each predictor model (Hypothesis 1A). Next, to compare the performance of the g-model and the Alternative model, the incremental validity of each will be examined (Hypothesis 1B). R^2 is first calculated with the g-score entered into the analysis first and the Alternative components being added second. Then, R^2 will be calculated with the Alternative components added first followed by the g-score. Evaluation of the change in R^2 based on F and p-values will indicate which has greater incremental validity.

To test the effect of adding service-oriented behaviors to the criterion space (Hypothesis 1C), two regression analyses will be conducted for each model. The first analysis will incorporate only the task performance score as the criterion. Next, a second analysis that incorporates both task and service-oriented behavior in the criterion will be performed. Examination of the ΔR^2 , via F and p-values, will indicate if addition of service-oriented behaviors lead to increased predictive power.

Sub-group differences

Individual predictor d -values were computed by subtracting the minority test mean from the majority test mean (Whites), and dividing the difference by the pooled standard deviation weighted by sample size. Negative d -values indicated that Blacks,

Hispanics, or Native Americans scored higher than Whites, whereas positive effect sizes indicated the reverse.

The primary interest in this study is the psychometric properties (i.e., criterion validity and sub-group differences) of predictor composites or models; consequently, the typical calculation of *d*-values is insufficient because they are calculated with the mean of two individual measures. However, Sackett and Ellingson (1997) point out that the standardized sub-group difference (*d*-value) for a predictor is an algebraic transformation of the correlation between the predictor in question and a dichotomous variable reflecting group membership. Therefore, the formula for the correlation between a model and an external variable can be extended to reflect the *d*-value for multiple predictors combined to form a model. The formula used to calculate the *d*-

value for a model is:

$$d = \frac{\sum_{i=1}^k d_i}{\sqrt{k + k(k-1)r_{ii}}}$$

Where d_i is the *d*-value for each predictor in the model, k is the number of predictors in the model, and r_{ii} is the average intercorrelation among the predictors in the model.

Results

See Table 6 for descriptive statistics and univariate correlations between measures used in the study.

Predictors. Intercorrelations among written test component predictors (i.e., Interpersonal Skills, *g*, and Job Knowledge) were moderate to high with *g* and Job Knowledge being the most correlated ($r = .51, p < .01$) and *g* and Interpersonal Skills the least ($r = .38, p < .01$). Only Job Knowledge and Interpersonal Skills were significantly correlated with T&E score ($r = .23, p < .01$ & $r = .08, p < .05$). The

significant correlation for Job Knowledge was expected, as one would expect persons with more opportunity to learn on the job to have mastered job knowledge to a higher degree.

Insert Table 6 about here

Criteria. Academy Grades were significantly related to supervisor ratings of Service ($r = .30, p < .01$), Task ($r = .36, p < .01$), and Overall Performance ($r = .30, p < .01$). Unfortunately Service and Task scores are almost perfectly correlated ($r = .92, p < .001$), which is a strong indicator that supervisors may not differentiate between these two variables. Because of the non-differentiation of Service and Task performance, only Overall performance will be used in validity analyses.

Relationship between Predictors and Criteria. Interpersonal Skills, Job Knowledge, and g scores were significantly correlated with the four criterion measures at the $p < .01$ level. The T&E measure was not significantly related to any of the criteria. The correlation of g with Academy Grades ($r = .54$) is about twice that of Interpersonal Skills and Job Knowledge with Academy Grades ($r = .30$ & $.39$). This was not surprising because g predicts how well one learns new information (Schmidt & Hunter, 1998). Additionally, one would expect those who had mastered more relevant job knowledge to also perform better at the academy so a high validity coefficient is reasonable. However, the moderate coefficient obtained for Interpersonal Skills was not expected. Regarding supervisor ratings of employee

Service, Task, and Overall Performance categorization, there is very little difference between g score and Job Knowledge score (.28 vs. .27; .31 vs. .29, and .14 vs. .15).

Validity

Academy Grades. We conducted a series of hierarchal linear regression analyses to determine which Model showed higher criterion validity and incremental validity; results are presented in Table 7.

Contrary to Hypothesis 1a, the g model had a larger adjusted R^2 value ($F_{(1, 470)} = 190.3, p < .01, R^2 = .29$) than the Alternative based model ($F_{(3, 468)} = 33.7, p < .01, R^2 = .18$). The g model also showed more incremental validity ($F_{(1, 467)} = 91.9, p < .01, \Delta R^2 = .135$) than the Experienced based model ($F_{(3, 467)} = 5.6, p < .01, \Delta R^2 = .025$). The g model explained 11% more variance than the Alternative model. For Academy grades, Hypothesis 1b was not supported.

Insert Table 7 about here

Supervisor Ratings. During initial stages of analyses, it was apparent supervisors made performance ratings in a categorical manner (i.e., those who did not meet standards, those that did, and those that were above standards). Specifically, 55% and 43% of the sample received a two or “Meets Standards” on their Service or Task rating. Due to the categorical nature of the dependent variable, we felt that Logistic Regression was most appropriate method to analyze the data. We sorted employees into categories based on the following scheme: employees with Service and Task scores less than two (5% & 11% respectively) were coded as a three for “Poor,” those

scoring a two (55% & 43% respectively) were coded as a two for “Good,” and those scoring more than a two (40% & 46% respectively) were coded as one for “Exceptional.”

Our initial intention for the aggregation of Service and Task scores into an Overall performance score was to sum the two scores together, however, due to the categorical nature and the differing levels of harshness and leniency between Service and Task scores this was not acceptable. Instead, we used a matrix categorization scheme, whereby we developed a frequency table of Service and Task scores. The mean of the interaction was then used to determine what category those members of that cell would fall into. For example, the intersection of Poor Service performance and Good Task performance would result in a mean of two-point-five ($3 \text{ (Poor)} + 2 \text{ (Good)} = 5/2 = 2.5$), which would result in a three coding.

Because there are multiple levels of the criterion variable and the categories are ordered an Ordinal Logistic Regression procedure seemed reasonable. However, preliminary analysis with Ordinal Logistic Regression indicated that the proportional odds assumption was not met (p-values ranging between .006 and .015). Therefore, we instead conducted a Nominal Logistic Regression analysis using SPSS’s NOMREG procedure. One hundred and thirty-nine cases were missing supervisor ratings, however of those, forty-nine had academy grades. Therefore, we regressed academy grades onto supervisor ratings to obtain regression weight estimates. Those regression weights were then used to calculate predicted supervisor ratings, which were then used in validity analyses. We ran logistic regression analyses with ($n = 529$), and without ($n = 480$) the substituted data and there were no practical

differences in obtained results, therefore, we used the larger sample with the regressed values substituted for missing data.

Nominal regression results are presented in Table 8. As expected, the Full model outperforms the g and Alternative models. However, what is of particular interest in this study is the comparison between the g model and the Alternative model. Results indicate the Alternative model outperformed the g model. The Alternative model reproduces the data more effectively than the g model as indicated by larger χ^2_{model} values (41.78 vs. 34.04), higher Naglekerke R^2 values (.088 vs. .073), and larger effects for the Alternative model given the greater distance from one for the Alternative model odds ratios than the g model odds ratios.

Insert Table 8 about here

Table 9 displays the likelihoods of being a “Poor,” “Good,” or “Excellent” employee for all three models for applicants scoring the minimum, mean, and maximum scores for each DV. The Alternative and Full models are more effective at identifying Poor and Exceptional employees as seen in the larger spread of the distribution for those scoring the minimum and maximum values on the predictors. Furthermore, Alternative model predictions are very similar to those made by the full model. This is important because it further illustrates that there is no practical difference between using the Full model and the Alternative model. For supervisor ratings, hypothesis 1A and 1B are supported.

Insert Table 9 about here

Figure 1 graphically demonstrates the relationships between scoring the minimum, mean, and maximum scores on each component of the Alternative model. As can be seen, those persons scoring the minimum are highly likely to be unsuccessful employees, whereas those scoring the mean, or above, will be satisfactory employees. Applicants scoring the maximum are very likely to be exceptional employees.

Insert Figure 1 about here

As mentioned earlier, because of the very high correlation between Service and Task Scores ($r = .92$), we concluded that supervisors generally did not differentiate between these two constructs when making ratings. As further evidence of this, logistic regression results indicated that there was no additional explained space, as indicated by larger χ^2 and Nagelkerke values when both Service and Task scores were examined jointly over when only one was examined. Therefore, Hypothesis 1C is not supported.

Subgroup Differences

Mean subgroup differences in standard deviation units predictor measures and the subsequent predictor models are presented in Table 10. The most obvious finding is Native Americans have less of an issue with subgroup differences than Blacks or

Hispanics. Second, obtained subgroup differences for the g model were substantially lower than the 1.0 standard deviation typically seen for cognitive ability measures, which is quite encouraging and consistent with Roth et. al. (2001).

Insert Table 10 about here

Third, job knowledge had a lower d than the g model for Blacks, but not for Hispanics or Native Americans. However, because the Hispanic sample size is so low ($n = 19$), it is quite likely that these estimates are far from stable and not indicative of the true relationship. Fourth, the T&E showed a large subgroup difference between Blacks and Whites that was in the Blacks favor. Last, and perhaps most important, the Alternative model shows a drastically lower subgroup difference between Blacks and Whites than the g model.

Regarding hypotheses 2A and 2B, the results are dependent on the subgroup being examined. For Blacks, both hypotheses were supported, with the Alternative model having the lowest d value, followed by the Full model, then the g model. For Hispanics and Native Americans, the g model had the lowest d value, which was followed by the Alternative model, and then the Full model. As mentioned previously, the small sample size for Hispanics may be contributing to the unpredicted results. For Native Americans, although the results were not as anticipated, the d values represent what is typically considered a small effect (Cohen, 1988). However, Sackett and Ellingson (1997) found that d values of .20 would lead

to violation of the “four-fifths rule” when selection ratios for the majority are 25% or less.

Discussion

To repeat, the primary purpose of this study was to examine if alternative selection devices typically seen in public sector hiring coupled with an expanded criterion space including service-oriented behaviors would lead to lower subgroup differences while maintaining comparable validity to cognitive ability. Secondly, this study provides information on the degree of subgroup differences for the Training and Experience measure, a very common selection device seen in public sector hiring that has, to date, received very little attention.

Results were mixed depending upon the criteria used in the analysis. For academy grades, the Alternative model did not perform as well as the g based model. However, this should come as no surprise as we know that g is very effective at predicting future learning ability (Schmidt & Hunter, 1998). For supervisor ratings, the Alternative model was a more effective predictor than the g based model. We feel that there are two reasons why this may be. First, persons more proficient in Interpersonal skills are more likely to reduce conflict among inmates, are more likely to communicate and cooperate with case workers assigned to inmates, and more effectively communicate and interact with their supervisor, which in turn leads to higher supervisor ratings. Second, persons that have a better understanding and knowledge of critical correctional principles are more likely to be able to answer questions of coworkers, as well as to be able to be left in charge during the supervisor’s absence. These types of behavior are not likely to be detected during

academy course quizzes and tests, which then, in turn, was not reflected in the analyses.

In this sample, expanding the criterion space did not lead to increased understanding of the relationship between the predictor sets and criteria. Given that there was such a high correlation between the two performance domains it appears that supervisors did not differentiate between the two. We believe if supervisors were to participate in frame-of-reference training this problem may be resolved.

Regarding subgroup differences, study results were quite encouraging, with an Alternative model displaying vastly lower Black-White difference than the g based model. A significant portion of the reduction is due to the impact of the T&E measure (removal results in a $d = .33$). Regrettably, the T&E did not possess criterion validity; either through correlation with the criterion or as part of a regression composite. Therefore, considerable thought should be used before including a T&E in the selection process. Fortunately, a content validation strategy could reasonably be used to justify use of previous training and experience. For example, persons with previous experience or who acquire necessary training and certification are more motivated to remain in Law Enforcement type jobs, and would be less likely to turnover. Turnover in correctional security jobs is extremely high, and therefore, a reasonable criterion to utilize in decision-making.

Oddly, subgroup differences for Hispanics and Native Americans were larger for the Alternative model than the g based model. As stated previously, the sample size for Hispanics was very small ($n = 19$) and may not be representative of the true relationship. For Native Americans, even though the d -value for the Alternative

model is larger than that for the g based model, it is indicative of a small effect size (Cohen, 1985).

Subgroup differences for specific job-related skills were lower than those seen with overall g measures, which is promising and consistent with previous research (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Therefore, when job analysis results indicate that specific cognitive based skills are required to satisfactorily perform the job, selection specialists can use those types of measures with confidence knowing that the skill is job-related and adverse impact is minimized. Furthermore, use of job-related skills for selection purposes is viewed favorably by the courts as indicated by outcomes of recent Supreme Court cases against the University of Michigan (Mumford, 2004).

This study provides much needed information on the degree of racial subgroup differences of some understudied areas of interest to public sector personnel decision makers, specifically the Training and Experience measure. Results indicated that, in this instance, the T&E measure actually favored Blacks ($d = -.21$) and was essentially zero for Hispanics and Native Americans ($d = .00$ & $.07$, respectively). Furthermore, the study provided information on racial subgroup differences in common tested domains for the understudied Native American culture. It appears that generally, there is little to no problem with White-Native American differences on commonly seen selection devices in the public sector.

Future Directions

Perhaps the largest limitation of this study is that it is a concurrent validity study conducted on persons selected for the job using a different selection tool.

Therefore there are range-restriction issues operating on both the predictor and criterion side of the equation. Furthermore, Roth, et. al. (2001) showed that you can obtain different subgroup differences between job applicants and job incumbents, which may be responsible for the low subgroup effect sizes rather than higher specificity of the g components. Therefore, further research needs to be conducted with job applicants where the official selection device is one using this strategy.

Additionally, it has been argued that it is unreasonable to measure previous experience or job knowledge for entry-level jobs that will train new hires in that knowledge. However, in this case, job knowledge was an effective predictor of future performance. It is not clear if this is due to 60% of the sample having previous job experience/training or if the job knowledge domain is acting as a tacit knowledge test measuring applicants ability to generalize previous knowledge structures in a practical manner in novel domains (Sternberg, 1985). An interesting question to ask is, “Do applicants feel that the test was unfair because they were being measured on material that they feel they shouldn’t be required to know before entering the job?” The answer is a resounding, “No.” Since 1993, the test’s inception, there has never been a complaint lodged by an applicant about the unreasonableness of being tested on job-related job knowledge domains. Regardless, more research is necessary to determine when it is appropriate and effective to use job knowledge for entry-level jobs and when it is not.

Schmitt and colleagues have made a case for using overall cognitive ability for all jobs due to its generalizable nature. However, strict use of overall cognitive ability measures will lead to grave societal problems. It seems more reasonable to use

alternative measures, such as job knowledge and interpersonal skills, in conjunction with job-related specific cognitive components. There would be greater validity, less adverse impact, and more “buy-in” from applicants due to the high face validity of the domains. Applicants prefer to see the link between testing material and their perception of what the job entails. Esoteric analogies and spatial orientation items tend to leave applicants confused and upset when they fail the test because of the weak link between tested material and the job domain.

Summary

In the current study, we found that a composite made up of the alternative predictors of Job Knowledge, Interpersonal Skills, and Training and Experience lead to higher validity than a g-based composite for supervisor ratings, but not academy grades. Furthermore, the Alternative composite displayed a smaller sub-group difference for Blacks, but not Hispanics or Native Americans. Additionally, this study provides much needed information on sub-group differences present in ratings of Training and Experience, a much-neglected topic in published literature.

References

- Abrams, N., Johnson, J., Sproule, Ch., & Ash, R. (2002). Training and Experience (T&E) Ratings: Training Manual. International Personnel Management Association Assessment Council.
- Aguinis, H., Cortina, J. M., & Goldberg, E. (1998). A New Procedure for Computing Equivalence Bands in Personnel Selection. *Human Performance*, 11, 351-365.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational Psychological Testing*. Washington, DC: American Educational Research Association.
- Arvey, R. (1979). *Fairness in Selecting Employees*. Reading, Mass. Adson-Wesler.
- Ash, R. A., & Levine, E. L. (1985). Job applicant training and work experience evaluation: An empirical comparison of four methods. *Journal of Applied Psychology*, Vol. 70, No. 3, 572-576.
- Avolio, B. J., Waldman, D. A., & McDaniel, M. A. (1990). Age and Work Performance in Nonmanagerial Jobs: The Effects of Experience and Occupational Type. *Academy of Management Journal*, Vol. 33(2), 407-422.
- Bobko, P., Roth, P. L. & Potosky, D. (1999). Derivation and Implications of a Meta-Analytic Matrix Incorporating Cognitive Ability, Alternative Predictors, and Job Performance. *Personnel Psychology*, Autumn 1999.
- Borman, W. C. & Motowidlo, S. J. (1993). Expanding the Criterion Domain to Include Elements of Contextual Performance. In N. Schmitt & W.C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 71-98.). San Francisco: Jossey-Bass.
- Borman, W. C. & Motowidlo, S. J. (1997). Task Performance and Contextual Performance: The Meaning of Personnel Selection Research. *Human Performance*, Vol. 10, 99-109.
- Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling Job Performance in a Population of Jobs. *Personnel Psychology*. Vol. 43. No. 2. 313-333.
- Chan, D. & Schmitt, N. (1997). Video-Based Versus Paper-and-Pencil Method of Assessment in Situational Judgment Tests: Subgroup Differences in Test Performance and Face Validity Perceptions. *Journal of Applied Psychology*, Vol. 82, No. 1, 143-159.
- Cleary, T. A. (1968). Test Bias: Prediction of Grades for Negro and White Students in Integrated Colleges. *Journal of Educational Measurement*, Vol. 5, 115-124.
- Cosier, R. A., & Dalton, D. R. (1983). Equity Theory and Time: A Reformulation. *Academy of Management Review*. 8, 311-319.
- Cronbach, L. J. & Snow, R. E. (1977). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*.

- Department of Labor. U.S. Labor Market in 2001: Economy Enters a Recession. Monthly Labor Review, February 2002.
- Dunnette, M. (1972). Validity study results for jobs relevant to the petroleum refining industry. Washington, DC: American Petroleum Institute.
- Dye, D. A., Reck, M., & McDaniel, M. (1993). The Validity of Job Knowledge Measures. *International Journal of Selection and Assessment*, Vol. 1, 153-162.
- Grinberg, D. & Ryan, J. (2005). EEOC Releases Fiscal 2004 Year-End Data: Highlights Include Record Monetary Relief, Expanded Mediation, and Outreach. U.S. Equal Employment Opportunity Commission, February 15, 2005. (<http://www.eeoc.gov/press/2-15-05.html>).
- Ferris, g. R., Witt, L. A., & Hochwarter, W. A. (2002). Interaction of Social Skill and General Mental Ability on Job Performance and Salary. *Journal of Applied Psychology*. Vol. 86. No. 6, 1075-1082.
- Goldstein, H. W., Yusko, K. P., & Nicolopoulos, V. (2001). *Personnel Psychology*. Vol. 54. No. 4. 783-807.
- Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, B. D., & Chung, B. (1998). The Role of Cognitive Ability in the Subgroup Differences and Incremental Validity of Assessment Center Exercises. *Personnel Psychology*. Vol. 51. No. 2. 357-374.
- Gottfredson, L. S. (2002). Where and Why *g* Matters: Not a Mystery. *Human Performance*, 1/2, 25-46.
- Grutter v. Bollinger (CA6 2002) 288 F.3d 732.
- Guion, R. (1966). Employment Tests and Discriminatory Hiring. *Industrial Relations*, 5, 20-37.
- Guion, R. (1983). Comments on Hunter's Model. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance Measurement and Theory*. Hillsdale, NJ: Erlbaum.
- Hattrup, K., Rock, J., & Scalia, C. (1997). The Effects of Varying Conceptualizations of Job Performance on Adverse Impact, Minority Hiring, and Predicted Performance. *Journal of Applied Psychology*, Vol. 82, No. 5, 656-664.
- Hirsh, H. R., Northrup, L., & Schmidt, F. (1986). Validity Generalization Results For Law Enforcement. *Personnel Psychology*, Vol. 39, 399-420.
- Hough, L. M., Oswald, F. L. & Ployhart, R. E. (2001). Determinants, Detection, and Amelioration of Adverse Impact in Personnel Selection Procedures: Issues, Evidence, and Lessons Learned. *International Journal of Selection and Assessment*, Vol. 9, No. 1/2, 152-194.
- Huffcutt, A., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and Meta-Analytic Assessment of Psychological Constructs Measured in Employment Interviews. Vol. 86, No. 5, 897-913.
- Huffcutt, A. I., & Roth, P. L. (1998). Racial Group Differences in Employment Interview Evaluations. *Journal of Applied Psychology*. Vol. 83. No. 2. 179-189.

- Hunter, J. E. (1982, May). The Validity of Content Valid Tests and the Basis for Ranking. Paper presented at the International Personnel Management Association Conference, Minneapolis, MN.
- Hunter, J. E. (1983). A Causal Analysis of Cognitive Ability, Job Knowledge, Job Performance, and Supervisory Ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance Measurement and Theory*. Hillsdale, NJ: Erlbaum.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and Utility of Alternative Predictors Across Studies. *Psychological Bulletin*, 96, 72-98.
- International Personnel Management Association (2001). 2000/2001 Benchmarking Report of the International Personnel Management Association (IPMA) and the National Association of State Personnel Executives (NASPE).
- Jencks, C. (1979). *Who gets ahead? The determinants of economic success in America*. New York: Basic Books.
- Kesselman, g. A., Lopez, F. E. (1979). The Impact of Job Analysis on Employment Test Validation for Minority and Nonminority Accounting Personnel. *Personnel Psychology*. Vol. 32(1), 91-108.
- Kulik, J. A., Kulik, C. C. and Bangert, R. L. (1984). Effects of Practice on Aptitude and Achievement Test Scores. *American Educational Research Journal*, Vol. 21. 435-447.
- Maranon, R. C., & Pueyo. (2000). The Study of Human Intelligence: A Review at the Turn of the Millennium. *Psychology in Spain*, Vol. 4. No 1, 167-182.
- Mathis, R. L. & Jackson. (2000). *Human Resource Management. 10th ed.* Mason, OH: Thomson South-Western.
- McHenry, J. J., Hough, L. M., Toquam, J. L., & Hanson, M. A. (1990). Project A Validity Results: The Relationship Between Predictor and Criterion Domains. *Personnel Psychology*. Vol. 43. No. 2. 335-354.
- Meritt-Haston, R. & Wexley, K. N. (1983). Educational Requirements: Legality and Validity. *Personnel Psychology*, 36, 743-753.
- Messick, S. (1989). Validity. Linn, Robert L.; Ed; *Educational measurement (3rd ed.)*; p. 13-103.
- Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, Winter, 1995, 5-8.
- McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A Meta-Analysis of the Validity of Methods for Rating Training and Experience in Personnel Selection. *Personnel Psychology*, Vol. 41, No. 2, 283-314.
- Molyneaux, J. W. (1953). *An Evaluation of Unassembled Examinations*. Unpublished master's thesis, The George Washington University, Washington, DC.
- Mosel, J. N. (1952). The Validity of Rational Ratings on Experience and Training. *Personnel Psychology*, 5, 1-10.

- Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence That Task Performance Should Be Distinguished From Contextual Performance. *Journal of Applied Psychology*. Vol. 79, No. 4, 475-480.
- Mumford, M.D. (April, 2004). The University of Michigan Court Cases: Promises and Problems. Presented at the Society for Industrial Organizational Psychologists Conference, Chicago, IL.
- Pulakos, E. D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion validity. *Human Performance*, Vol 9(3). Special Issue: Public sector assessment. pp. 241-258.
- Quinones, M. A., Ford, J. K., & Teachout, M. S. (1995). The Relationship Between Work Experience and Job Performance: A Conceptual and Meta-Analytic Review. *Personnel Psychology*, 48(4), 887-910.
- Ree, M. J. & Earles, J. A. (1992). Intelligence is the Best Predictor of Job Performance. *Current Directions in Psychological Science*, Vol. 1, 86-89.
- Reilly, R. R., & Chao, g. R. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology* Vol. 35(1), Spring 1982, 1-62.
- Reilly, R. R., & Warech, M. A. (1993). The Validity and Fairness of Alternatives to Cognitive Ability Tests. In Linda Wind and Bernard R. Gifford (Eds.). *Policy Issues in Employment Testing*. Norwell, MA: Kluwer.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S. & Tyler, P. (2001). Ethnic Group Difference in Cognitive Ability in Employment and Educational Settings: A Meta-Analysis. *Personnel Psychology*, Vol. 54, No. 2, 297-331.
- Roth, P. L., Van Iddekinge, C. H., Huffcutt, A., Eidson, C. E., & Bobko, P. (2002). Corrections for Range Restriction in Structured Interview Ethnic Group Differences: The Values May be Larger Than Researchers Thought. *Journal of Applied Psychology*. Vol. 87, No. 2, 369-376.
- Russell, C. J. (1987). Person Characteristic Versus Role congruency Explanations for Assessment Center Ratings. *Academy of Management Journal*. Vol. 30. No. 4. 817-826.
- Ryan, A. M., Ployhart, R. R., Greguras, G.J. & Schmit, M.J. (1998). Test Preparation Programs in Selection Contexts: Self-Selection and Program Effectiveness. *Personnel Psychology*, Vol. 51, 599-621.
- Sackett, P. R., & Ellingson, J. E. (1997). The Effects of Forming Multi-Predictor Models on Group Differences and Adverse Impact. *Personnel Psychology*, Vol. 50, No. 3, 707-722.
- Sackett, P. R., Schmitt, N., Kabin, M. B. & Ellingson, J. E. (2001). High-Stakes Testing In Employment, Credentialing, and Higher Education: Prospects in a Post-Affirmative-Action World. *American Psychologist*, Vol. 56, No. 4, 302-318.
- Schmidt, F. L., Gast-Rosenberg, I, & Hunter, J. E. (1980). Validity Generalization for Computer Programmers. *Journal of Applied Psychology*. Vol. 65, 643-661.

- Schmidt, F. L. & Hunter, J. E. (1998). The Validity and Utility of Selection Methods in Personnel Psychology Practical and Theoretical Implications of 85 Years of Research Findings. *Psychological Bulletin*, Vol. 124, No. 2, 298-307.
- Schmidt, F. L., Hunter, J. E. & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology* Vol. 71(3), Aug 1986, 432-439.
- Schmidt, F. L., Hunter, J. E., Outerbridge, A. N. & Goff, S. (1988). Joint Relation of Experience and Ability with Job Performance: Test of Three Hypotheses. *Journal of Applied Psychology*, Vol. 73, No. 1, 46-57.
- Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup Differences Associated with Different Measures of Some Common Job-Relevant Constructs. *International Review of Industrial and Organizational Psychology*. Vol. 11, 115-139.
- Schmitt, N., Rogers, W., Chan, D., Sheppard, L. & Jennings, D. (1997). Adverse Impact and Predictive Efficiency of Various Predictor Combinations. *Journal of Applied Psychology*, Vol. 82, No. 5, 719-730.
- Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. New York: Cambridge University Press.
- Vance, R. J., Coover, M. D., MacCallum, R. C. & Hedge J. W. (1989). Construct Models of Task Performance. *Journal of Applied Psychology*, Vol. 74, No. 3, 447-455.
- Van Scotter, J. R. & Motowidlo, S. J. (1996). Interpersonal Facilitation and Job Dedication as Separate Facets of Contextual Performance. *Journal of Applied Psychology*. Vol. 81. No. 5. 525 – 531.
- Wagner, R. K. (1997). Intelligence, Training, and Employment. *American Psychologist*, Vol. 52, No. 10, 1059-1069.
- Wechsler, D. (1958). *The Measurement and Appraisal of Adult Intelligence*. 4th ed. Baltimore: Williams & Wilkins.
- Wiesen, J. P., Abrams, N., & McAttee, S. A. (1994). Employment Testing: A Public Sector Viewpoint. In L. C. Wing & B. Gifford (Eds.), *Policy Issues in Employment Testing*, Boston, MA: Kluwer Academic Publishers.
- Wing, H. (1981). Estimation of the Adverse Impact of a Police Promotion Examination. *Personnel Psychology*, Vol. 34 No. 3, 503-510.

Table 1: Summary of Standardized Mean-Score Differences on Cognitive Ability, Constructs for Ethnic, Gender, and Age Groups (Hough, Oswald, & Ployhart, 2001)

Comparison Groups	Cog Ability	Verbal Ability	Quantitative Ability	Science Achieve	Spatial Ability	Memory	Mental Process Speed
Black/White	1.0	.6	.7	1.0	.7	.5	.3
Hispanic/White	.5	.4	.3	.6	--	--	.4
East Asian/White	-.2	--	--	--	--	--	--

Table 2: Criterion validity of T&Es with Supervisor's Ratings of Job Performance (Mosel, 1952)

Job	Sample Size	r
Carpenter	51	.19
Equipment Repairman	40	.16
Machinist I	76	.00
Machinist II	100	.04
Machine Operator	108	.14
Ordinanceman-Torpedo	125	.09
Radio Mechanic	116	-.17
Aviation Metalsmith	98	.12
Highlift Forklift Operator	116	-.04
Auto Mechanic	99	.27**
Painter	75	.26*
Ordinanceman	100	-.09
Hand Compositor	116	.16

* p < .05. ** p < .01.

Table 3: Criterion Validities for Educational Requirements by Criterion Type (Meritt-Haston & Wexley, 1983)

Criterion	Sample Size	- r
Overall	71,374	.26
Tenure	69,633	.27
Ratings	1,541	.15
Promotions	200	.23

Table 4: Criterion validity of Educational Requirements from Several Unpublished Studies (Hunter and Hunter, 1984)

Study	Criterion	Number of Correlations	r
Dunnette, 1972	Not Specified	15	.00
Vineberg and Joyner, 1982 ^a	Global Rating of Performance	25	.14
	Suitability—(e.g. Potential Ability)	10	.36
	Overall	35	.25
Hunter, 1980	Training	90	.20
	Proficiency	425	.10
	Overall	515	.15

^aCorrected for error of measurement

Table 5: Correlations between Task Performance, Contextual Performance, and Personal Characteristics (Motowidlo & Van Scotter, 1994)

Personal Characteristic	Sample Size	Task Performance	Contextual Performance
Experience	269	.34**	.16**
Cognitive Ability	174	-.01	.15*
Training Performance	192	.14*	.18**
Work Orientation	252	.23**	.36**
Dominance	254	.04	.11*
Dependability	253	.18**	.31**
Adjustment	254	.08	.14*
Cooperativeness	255	.03	.22**
Internal Control	256	.08	.23**

* p < .05. ** p < .01.

Table 6

Means, Standard Deviations, and Intercorrelations Among Variables for Total Sample

Variable	n	M	SD	1	2	3	4	5	6	7	8
1. g	619	48.2	5.92	1.0							
2. JK	619	32.2	4.01	.51**	1.0						
3. IP	619	8.7	1.18	.38**	.41**	1.0					
4. TE	619	40.4	75.30	.07	.23**	.08*	1.0				
5. ACAD	472	93.5	3.14	.54**	.39**	.30**	.06	1.0			
6. SERV	529	2.0	.29	.28**	.27**	.20**	.06	.30**	1.0		
7. TASK	529	2.0	.27	.31**	.29**	.22**	.07	.36**	.92**	1.0	
8. OVER	529	1.9	.67	.14**	.15**	.12**	.03	.30**	.57**	.62**	1.0

*p < .05, **p < .01 Note: g = Cognitive Ability, JK = Job Knowledge, IP = Interpersonal Skills, TE = Training and Experience, ACAD = Academy Grades, SERV = Service Score, TASK = Task Score, OVER = Overall Performance Categorization.

Table 7:
Hierarchal Regression of Cognitive and Alternative Models with Academy Performance

Model	R ²	R ² _{adj}	ΔR ²	Beta
g only	.288*	.287*	---	.537**
g + Experience	.313*	.307*	.025*	---
g	.288*	.287*	.288*	.437**
JK	.308*	.305*	.020*	.144**
IP	.313	.309	.005	.080
TE	.313	.307	.000	-.008
Experience only	.178*	.173*	---	---
JK	.150*	.148*	.150*	.330**
IP	.177*	.173*	.027*	.176**
TE	.178*	.173*	.001	-.032
Experience + g	.313*	.307*	.135*	---

* p < .01 Note: g = Cognitive Ability, JK = Job Knowledge, IP = Interpersonal Skills, TE = Training and Experience

Table 8:
Nominal Regression of Overall Performance on g Based Model, Alternative Model, and Full Model

Model	X^2_{MODEL}	Incremental X^2 of the Model	Naglekerke R^2	Odds Ratio Excellent vs. Poor	Odds Ratio Good vs. Poor
Overall Performance Score					
g	34.04**	9.64**	.073	1.137**	1.120**
Alternative	41.78**	17.38**	.088	--	--
JK	--	--	--	1.137**	1.147**
IP	--	--	--	1.423**	1.160
TE	--	--	--	1.001	.998
Full	51.42**	--	.108	--	--
g	--	--	--	1.083**	1.081**
JK	--	--	--	1.081	1.067
IP	--	--	--	1.307*	1.067
TE	--	--	--	1.001	.999

* $p < .05$, ** $p < .01$ Note: g = Cognitive Ability, JK = Job Knowledge, IP = Interpersonal Skills, TE = Training and Experience

Table 9:
The Likelihood of Being in the Poor, Good, or Exceptional Category based on Minimum, Mean, and Maximum Scores

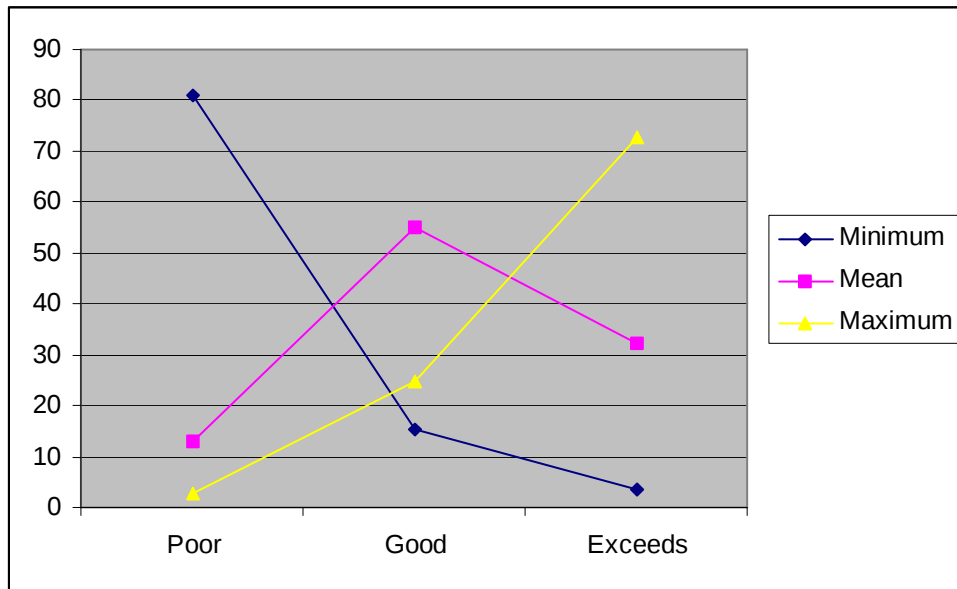
		MINIMUM			MEAN			MAXIMUM		
Model		Poor	Good	Exc	Poor	Good	Exc	Poor	Good	Exc
Overall	g	73.56	18.69	7.75	12.54	55.08	32.39	2.62	56.77	40.61
	Alt	81.08	15.26	3.66	12.90	54.83	32.27	2.74	24.57	72.7
	Full	89.14	8.79	2.07	12.45	56.3	31.25	1.51	34.56	63.93

Table 10:
Mean Subgroup Differences for Predictor Measures

	Black Vs White	Hispanic Vs White	Native American Vs White
g Model	.60	.36	.09
Job Knowledge Score	.44	.51	.20
Interpersonal Score	.12	.41	.18
T&E Score	-.21	.00	.07
Alternative Model	.17	.44	.21
Full Model	.35	.47	.20

Note: Sample sizes: White = 463, Black = 58, Hispanic = 19, Native American = 79

Figure 1:
Graphical Representation of Success Categorization Using the Alternative Model for an Applicant Scoring Minimum, Mean, and Maximum Scores



Appendix A: Sample Training & Experience Measure

EDUCATION POINTS (Maximum)

120

Bachelors Degree in Law Enforcement Related or Behavioral Science field (i.e., Criminal Justice, Police Science, Psychology, Sociology)	120 Points
Bachelors Degree in any other field	100 Points
Associates Degree in Law Enforcement or Behavioral Science field (i.e., Criminal Justice, Police Science, Psychology, Sociology)	60 Points

CERTIFICATION POINTS (Maximum)

40

Graduation from another state's Correctional Academy	40 Points
Council on Law Enforcement Education and Training (CLEET)	30 Points
Reserve CLEET	15 Points

EXPERIENCE POINTS (Maximum)

360

Type of Experience	Experience acquired within the past 10 years
Same: Experience working in a Federal, State, or private adult long-term correctional facility regardless of security level.	3 pts./mo.
Similar: Experience working in adult city or county short-term detention centers or jails.	2.5 pts./mo.
Related: Experience working in juvenile detention centers as a security officer.	1.5 pt./mo.
Other: Experience working as a certified law enforcement officer.	2 pt./mo.