Development and validation of three alternative forms of a published general mental ability test

Lori Jean Casper

Follow this and additional works at: http://scholarworks.lib.csusb.edu/etd-project

Part of the Psychology Commons

Recommended Citation
http://scholarworks.lib.csusb.edu/etd-project/1295

This Thesis is brought to you for free and open access by the John M. Pfau Library at CSUSB ScholarWorks. It has been accepted for inclusion in Theses Digitization Project by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
DEVELOPMENT AND VALIDATION OF THREE ALTERNATE FORMS
OF A PUBLISHED GENERAL MENTAL ABILITY TEST

A Thesis
Presented to the
Faculty of
California State University,
San Bernardino

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Industrial/Organizational Psychology

by
Lori Jean Casper

March 1997
DEVELOPMENT AND VALIDATION OF THREE ALTERNATE FORMS OF A PUBLISHED GENERAL MENTAL ABILITY TEST

A Thesis
Presented to the
Faculty of
California State University,
San Bernardino

by
Lori Jean Casper

March 1997

Approved by:

Matt L. Riggs, Ph.D., Chair, Psychology
Kenneth S. Shultz, Ph.D., Psychology
Elizabeth Gale, Ph.D.,
Southern California Gas Company

Date
3/21/97
ABSTRACT

The purpose of this study was to create and validate three alternate forms of a published mental ability test. One hundred and twenty-two items were created and piloted with 152 students from two large universities, one in the midwest and one in the southwest. Item analysis was completed to identify those items which best correlated with participants' total scores. Three alternate forms were constructed from the remaining test questions in the item pool. One hundred and eighty employees from a large southwestern utility company took both the original published test and one of the three alternate forms. Descriptive statistics, item-difficulty levels, reliability estimates, correlations and group norms were calculated for the original test and three alternate forms.
ACKNOWLEDGMENTS

Every normal man, woman and child is...
a genius at something, as well as an idiot at something.

Spearman, 1927, p. 221

I would like to thank Dr. Calvin Hoffman for allowing me to use employees' time and data for this study. I also would like to express my appreciation to my thesis committee, especially my Chair, Dr. Matt Riggs, for sharing their time and expertise.

Thank you to my family and friends for always believing that I would finish and giving me that extra push when I needed it most. Lastly, I want to acknowledge Peter Stone whose love and support was essential to the completion of my Master's Degree.
TABLE OF CONTENTS

ABSTRACT ........................................................................................................ iii

ACKNOWLEDGMENTS .................................................................................... iv

LIST OF TABLES ............................................................................................... vii

INTRODUCTION

Purpose of Study .............................................................................................. 1

Intelligence and Mental Abilities .................................................................. 1

Testing Intelligence ......................................................................................... 5

Testing Intelligence to Predict Job Success ................................................. 9

SRA’s Adaptability Test ................................................................................ 14

General Problem Solving Test ..................................................................... 18

Hypothesis ...................................................................................................... 20

METHODS

Phase I: Item Development ......................................................................... 20

Initial Pilot ..................................................................................................... 21

Item Analysis ................................................................................................ 22

Development of the GPST Alternate Forms .............................................. 24

Phase II: Piloting of the GPST A, GPST B and GPST C ......................... 24

Analyses ....................................................................................................... 25

RESULTS

Data Screening/Cleaning ............................................................................. 26

Descriptive Statistics .................................................................................. 27
TABLE OF CONTENTS (Continued)

DISCUSSION
   Hypothesis........................................................................................................32
   Future Research..................................................................................................37

APPENDIX A: Standardized Pilot Study Test Instructions.........................38

APPENDIX B: Item-Total Correlations
   Calculated on the Pilot Study Data..............................................................39

APPENDIX C: Standardized Validation Study Test Instructions.................41

APPENDIX D: Scatterplot of GPST and Adaptability Test
   Total Scores - Form A ..................................................................................42

APPENDIX E: Scatterplot of GPST and Adaptability Test
   Total Scores - Form B ..................................................................................43

APPENDIX F: Scatterplot of GPST and Adaptability Test
   Total Scores - Form C ..................................................................................44

APPENDIX G: Range Restriction Formula.......................................................45

BIBLIOGRAPHY....................................................................................................46
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.</td>
<td>Validation Coefficients for the Adaptability Test</td>
<td>16</td>
</tr>
<tr>
<td>Table 2.</td>
<td>Means and Standard Deviations for the Adaptability Test and Each General Problem Solving Test Alternate Form</td>
<td>28</td>
</tr>
<tr>
<td>Table 3.</td>
<td>Difficulty Levels</td>
<td>29</td>
</tr>
<tr>
<td>Table 4.</td>
<td>Correlations and Effect Sizes of General Problem Solving Test Total Scores and Adaptability Test Total Scores</td>
<td>31</td>
</tr>
<tr>
<td>Table 5.</td>
<td>Norm Comparisons</td>
<td>33</td>
</tr>
<tr>
<td>Table 6.</td>
<td>Summary of Analyses Comparing the Adaptability Test and the General Problem Solving Test</td>
<td>36</td>
</tr>
</tbody>
</table>
INTRODUCTION

Purpose of Study

The purpose of this study is to create three alternate forms of Science Research Associates' (SRA) Adaptability Test for use at a large western utility company ("The Company"). The Adaptability Test is a general mental ability measure used by The Company for approximately ten years to help select candidates for cognitively demanding non-exempt field and clerical positions.

Intelligence and Mental Abilities

What is intelligence? The concept of intelligence is a well researched topic that historically has been difficult to define precisely. During the first forty years of this century, the idea of intelligence or general mental ability was actively researched by Psychologists. Many of the most talented Psychologists devoted their lives to its study. Consequently, early volumes of the British Journal of Psychology and many American journals contained a high proportion of articles on the subject (Butcher, 1968).

The concept of intelligence was systematized by a 19th century mind. Francis Galton defined the notion of intelligence as "a fundamental ability, super-ordinate to and distinct from special abilities, which is responsible for the human superiority in the evolutionary struggle" (Deese, 1993, p.107). However, it was Galton's half-cousin, Charles Darwin, who turned this notion into a
functional theory. With the widespread acceptance of Social Darwinism, the race was on to determine the abilities which lead to superiority and inferiority which in turn helped further define intelligence.

Early in this century, powerful support for this fundamental ability came from two main sources. In 1904, Charles Spearman published a groundbreaking article in the *American Journal of Psychology* describing his now famous \( g \) theory. Spearman theorized and produced extensive statistical evidence supporting intelligence as a single construct which he called “general ability” or “\( g \)”. He even went so far as to deny the importance of specific abilities, or “\( s \)”. His ideas have been widely accepted for many years.

At about the same time, Alfred Binet of France developed the first standardized scale for assessing differences in intelligence. His impetus was the study of mental retardation. Binet was assigned to lead a commission to determine if mentally challenged children could benefit from the ordinary curriculum of the public schools. It was his responsibility to develop diagnostic procedures to identify the degree of retardation of French children and to help design an appropriate instructional program (Gould, 1981). By 1909, Sir Cyril Burt, Psychologist to the London Council, was also using standardized tests to demonstrate that many children certified as “mentally deficient” were really within the normal range of intelligence and were just slow learners (Burt, 1955).
In the succeeding forty years, the work of these pioneers was extended and refined, but few ground-breaking developments occurred. Since World War II, intelligence research has focused mostly on the debate between the existence of $g$ versus specific abilities.

Thurstone (1947) was the first to challenge Spearman’s concept of intelligence as a single entity. He proposed a small collection of entities which formed a composite and a “second order general factor,” which the primary entities had in common. This challenge opened the floodgates. Perhaps the most exhaustive work negating the notion of a $g$ factor was published by J. P. Guilford and his associates in 1967. They used factor analysis to establish 120 possible different abilities, with no mention of a $g$ factor.

However, the pendulum seems to have swung back in more recent literature. Bennett, Seashore and Wesman (1966) did some research using data from the Differential Aptitude Test (DAT) Battery. This set of eight tests, first published in 1947, has been widely used for educational and vocational counseling in high schools for the past 35 years. According to their titles, the eight tests measure: verbal reasoning, numerical ability, abstract reasoning, clerical speed and accuracy, mechanical reasoning, space relations, spelling, and English usage. However, the average intercorrelation of the tests is .40. Using only the common factor shared by the eight tests, Bennett et al. were able to account for 45% of the total variance. This is about 90% of the variance
accounted for by ability tests specially tailored for each high school. They concluded that a general ability test applied equally to each school loses some validity, but only a small amount.

Other intelligence researchers agree with these findings. Schmidt and Hunter (1978) supported the idea of a general factor in their validity generalization research. They stated that the general factor could account for all abilities and the differences found in validity being attributed to specific abilities were actually due to sampling error. More recently, Ree et al. (1991a, 1991b, 1992, 1994) conducted a series of experiments with United States Air Force pilots in which they consistently found a salient g factor. Larson and Saccuzzo (1989) also supported Spearman's general factor with minor clarifications. Their research found that "g appears related to the ability to flexibly and consistently reconfigure the contents of working memory" (p. 5).

Miller and Vernon (1992) investigated the notion of general intelligence across three distinct batteries of ability tests. Their results demonstrated that "significantly correlated general factors can be extracted from distinct batteries of tests" (p. 29). In a large sample meta-analytic study, Hunter (1983b, 1983c, 1985) found that the validity of specific aptitude measures, such as verbal or quantitative tests, stems from their measurement of general mental ability, or g.

While intelligence has a rich history of research, the subject hasn’t been given much attention in the past decade. However, with the controversy
created by Herrnstein and Murray's 1994 book entitled *The Bell Curve,* intelligence and its relevance to society has come back into the spotlight.

Recently, 52 well-known experts in the field of mental abilities came together and unanimously decided upon a mainstream, modern definition of intelligence.

Intelligence is the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or "test-taking smarts". Rather, it reflects a broader and deeper capability for comprehending our surroundings.

(Arvey et al., 1994, p. 67)

**Testing Intelligence**

Taking Arvey's definition, how does one go about testing or measuring the concept of intelligence? As already described, Binet was the first to undertake this challenge with modern techniques. Before developing his "modern" intelligence test for the mentally challenged, he tried many other approaches to measuring intelligence. He examined cranial, facial and hand form as well as handwriting (Anastasi, 1982). After these indirect methods proved unsuccessful, Binet decided to try to measure intelligence more directly.

To assist in assessing the mentally challenged, Binet developed an intelligence test called the 1905 Scale consisting of 30 items that were linked to a child's age and were administered in order of increasing difficulty. The test was designed to measure judgment, comprehension, and reasoning, which
Binet regarded as essential components of intelligence. He used the information gathered to determine a child’s mental age. Binet made several revisions and translated his test into English. This became the basis for the *Stanford-Binet Test of Intelligence* (Gould, 1981).

Soon after the development of Binet’s test, Stern (1912) demonstrated that it was possible to calculate an intelligence quotient by dividing a person’s mental age by his/her chronological age. This ratio, multiplied by 100, was called the “intelligence quotient” or “I.Q.”.

Large scale adult group testing began in earnest in the military during World War I. Robert Yerkes, in conjunction with other psychologists, designed the Army Alpha to help assign soldiers to war-time positions that best suited their intellectual capabilities. Yerkes quickly discovered that approximately 30 percent of the recruits were illiterate, mostly due to recent immigration or lack of available formal schooling. To more accurately test illiterate soldiers, Yerkes and his colleagues developed the Army Beta, a special cognitive ability test for those who couldn’t read English (Muchinsky, 1983).

The testing process was slow and the war ended within six months. Because the Army Alpha and Army Beta were only used for a short time, the intelligence testing program didn’t contribute as much to the war as Yerkes would have liked. Even though 1,726,000 individuals were tested in the program, actual use of the results was minimal (Thorndike & Lohman, 1990).
While psychology's actual impact on the war effort was not substantial, the recognition given to the field of psychology was a great impetus for the profession. Psychologists were regarded as people who could make a valuable contribution to society; consequently, applied psychology emerged from the war as a recognized discipline.

Throughout the 1920s, the testing movement underwent a tremendous growth spurt. Group intelligence tests were being developed for all ages, from pre-school children to graduate students. Teachers began to give intelligence tests to their classes while college students were routinely examined prior to admission. Soon the general public became IQ-conscious (Anastasi, 1982).

World War II began in 1941 and psychologists were again called on to assist with testing. Between the two wars, psychologists had studied the problems of employee selection and placement and had refined their techniques. Thus they were more prepared for World War II than World War I (Meier, 1994). While new selection tools such as stress tests and job knowledge tests were used, intelligence and aptitude tests were again the heart of the selection and placement process. Each branch of the service (the Army, Air Force, and Navy) developed a testing program to identify those who were or were not fit for military duty. Recruits found fit for service were then classified where their talents would be of greatest value to the war effort.
The military testing programs of World War II introduced the widespread application of test batteries designed to assess different functions (Thorndike & Lohman, 1990). For example, the Army used an overall screening device similar to the Army Alpha called the Army General Classification Test (AGCT). During the course of the war, more than 9 million men were given this battery of tests.

Together, the two World Wars were largely responsible for the development and expansion of modern industrial psychology. World War I helped form the profession and give it social acceptance; World War II helped develop and refine it. After 1946, industrial psychology experienced a splintering effect. Sub-specialties formed and intelligence testing branched off with employment testing to form "Personnel Psychology".

The field of mental testing was relatively quiet for the next couple of decades. As Oscar Buros reflected on his 50-year career in testing: "Except for the tremendous advances in electronic scoring, analysis, and reporting of test results, we don't have a great deal to show for fifty years of work" (Thorndike & Lohman, 1990, p. 85). One of the most significant contributions during this time was the development of several multiple-aptitude batteries including: The Differential Aptitude Tests (DAT), the General Aptitude Test Battery (GATB), and the Armed Services Vocational Aptitude Battery (ASVAB) (Muchinsky, 1983). These tests measure general intelligence and are still being used today.
Additionally, the concept of validity generalization was introduced by Schmidt and Hunter (1978). Many researchers believe test validity is situationally specific; however, Schmidt and Hunter's validity generalization theory suggests that criterion-related validity of the test can be transferred to another setting if the criteria in the two settings are very similar. This theory has been well researched and discussed and still generates much debate today (Crocker & Algina, 1986).

**Testing Intelligence to Predict Job Success**

One may wonder why cognitive ability tests would help predict job success. Going back to our modern definition of intelligence given to us by Arvey et al., ("the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience...") it is understandable that employees would need these qualities to be successful in their jobs. Almost every job imaginable requires these skills to some extent.

Research has demonstrated that assessing cognitive abilities can help organizations improve their employee selection. Many businesses have relied on subjective methods in choosing employees, such as unstructured interviews, reference checks and resume assessments. However, this practice has several drawbacks. One disadvantage is the lack of standardization of such methods. That is, the evaluation process is not identical for all applicants who
apply for a given job. For example, it isn’t unusual for interviewers to ask each job applicant a unique set of questions which measure very different constructs. This lack of consistency results in different information being used to evaluate each candidate. Along with being ineffective and unfair, this process could have legal ramifications for the organization. Additionally, extensive research indicates that these subjective methods have very low validity in predicting job success. A meta-analysis on employment interviews conducted by Reilly and Chao (1982) found unstructured interviews demonstrated a validity of about .19. These results have been duplicated by an additional meta-analysis by McDaniel, Whetzel, Schmidt, Hunter, Mauer and Russell (1988) which also examined the validity of interviews. According to the research, subjective methods are not very accurate in predicting how well employees will actually perform on the job (Arvey, 1979; Reilly & Chao, 1982).

Educational attainment may soon prove to be an ineffective predictor of job performance. Results of the 1992 National Norms Study conducted by Wonderlic Personnel Test Inc. reveal that educational attainment as an indicator of workers’ performance is decreasing in validity and reliability along with the decline in the level of graduates’ abilities. The study also found a sharp drop in the ability level of job applicants in relation to entry level job requirements. While these findings do not hold true for all university graduates, these and other results of the study should prompt employers who rely heavily
on educational attainment in making hiring decisions to explore more valid predictors of job performance. The good news is that more predictive alternatives are available. Objective, standardized, paper-and-pencil general mental ability tests are one option. Scores on cognitive ability tests have been shown over and over to successfully predict various measures of job performance such as supervisor ratings, work samples, and production rates (Hunter, 1983a, 1983c; Schmidt, Hunter, McKenzie & Muldrow, 1979; Thorndike, 1985). Schipmann and Prien (1989) reported an uncorrected correlation of .35 between general mental ability and rate of managerial progression (age-corrected managerial rank). Austin and Hanisch (1990) conducted a large-sample longitudinal study that looked at cognitive ability of high school students and their job success in adulthood. Their study found that mental ability scores obtained by high school sophomores were the best predictor of occupational attainment eleven years later. In another study conducted by Schmidt et al. (1988b), people at all levels of cognitive aptitude improved with job experience; however, differences between higher and lower aptitude personnel persisted. These studies support using a cognitive ability test to predict job success. In fact, cognitive ability scores are often described as the “best available predictor” of job performance (Neisser et al., 1976; Phillips & Dipboye, 1989).
Using valid predictors to select employees will improve an organization's bottom line. Utility analyses conducted on valid selection batteries demonstrate the cost benefits of implementing a testing program (Schmidt et al., 1979). Specifically, several utility analyses have been carried out on cognitive ability tests. Schmidt, Hunter, Outerbridge and Trattner (1986) conducted utility analyses for most white-collar jobs in the federal government. Their results indicate that "selection of a one-year cohort based on valid measures of cognitive ability, rather than on non-test procedures (mostly evaluations of education and experience), produces increases in output worth up to $600 million for each year that the new employees remain employed by the government" (p. 1). Dunnette (1989) reported large utility gains from improved selection (which included a test of mental ability) of electrical power plant operators. The National Research Council on the GATB (Hartigan & Wigdor, 1989) devoted considerable attention to the economic value of the testing program and concluded that utility is substantial for employers. Additionally, Johnson, Zeidner and Scholarios (1990) argue that the potential utility gain from cognitive testing, over and above the utility of other valid selection methods, is larger than suggested by typical validates of general ability.

According to employers' perceptions, mental ability tests, used as part of a selection battery, have improved the quality of job candidates hired in their organizations. A survey completed by HR Strategies (1992) in October of 1991
demonstrated that 43% of the 600 companies surveyed were not satisfied with their “current selection procedure’s ability to identify those possessing necessary job skills.” By contrast, a survey conducted by Wonderlic Personnel Test Inc. in January of 1992, found that of 720 companies who have recently implemented pre-employment tests (including a general ability test), 77% were satisfied with the productivity of employees they were selecting. Less than 2% said that the pre-employment tests were “not improving their selection process.”

Ree and Earles (1991) demonstrated that administering a general mental ability test was the best predictor of job training success for Air Force pilots. In fact, they found that specific ability tests provided very little incremental validity. Ree et al. (1994) reproduced his 1991 findings in another study. Again, they found that a general ability test is the best predictor of Air Force pilot and navigator success.

Other case studies have also demonstrated that testing cognitive ability helps predict job success. Franciscan Health Systems of Dayton, Ohio implemented a “Nursing Assistant Test Battery” which included a test of g. They found great improvements in the quality of care and reduction of turnover which they estimated as a $300,000 annual savings (Thomas & Brull, 1993). Robert Solomon (1993) reported that using The Wonderlic Personnel Test (a general intelligence test) to select front-office employees significantly reduced the number of underperforming workers. The Wonderlic Personnel Test was
also used as part of a battery to discriminate between “best” and “least best” correctional officers in a research study. The battery successfully distinguished “best” from “least best” at the .05 level (Super, Blau, Wells & Murdock, 1993).

The research described above clearly supports using cognitive ability tests to predict job success. These tests demonstrate criterion-related validity and substantial utility.

**SRA’s Adaptability Test**

Another measure of cognitive abilities (or g) used for selection in a number of companies is the SRA Adaptability Test. The Adaptability Test was written by Tiffin and Lawshe in 1943 and has gone through several revisions, with the most current revision being in 1985. The test is a speeded paper-and-pencil test which measures cognitive abilities and mental adaptability. The questions include items similar to those used by Thurstone (1938) in his analysis of “primary mental abilities.” The Adaptability Test was designed to assess skills that are important for successful performance in most management and non-management positions. The test is a general ability measure consisting of 35 questions which collectively measure verbal, numerical, and analytical problem solving abilities. Administration time is 15 minutes.
In the early 1980s, the utility company for which alternative forms are being developed implemented SRA's Adaptability Test (Short Form) as part of several selection batteries. The test is still currently being used in these batteries. Criterion-related validity evidence exists for the Adaptability Test both inside and outside The Company. Various criterion-related validation studies conducted by Science Research Associates, Inc. (SRA) demonstrated the relationship between the Adaptability Test and performance in a wide variety of non-management business and industrial occupations. These studies are documented in the SRA Adaptability Test's Examiner's Manual.

Criterion-related validity and job-component validity studies conducted by The Company demonstrate a significant predictive relationship between Adaptability Test scores and performance in non-management positions (see Table 1 for criterion validity and job component validity coefficients).

While The Company was pleased with the predictive ability of the Adaptability Test, there was a concern that the test questions (but not subject matter) were outdated and not face valid for the utility industry. Additionally, the test had been in circulation at The Company for several years and the answers may have become accessible to new job candidates, compromising the validity of the measure. It was believed that using alternate forms of the test would help maintain fairness and security (Holland & Rubin, 1982). For example, if the same questions were used at each administration of the test in one year,
Table 1

Validation Coefficients for the Adaptability Test

<table>
<thead>
<tr>
<th>Job Title</th>
<th>DOT Code</th>
<th>N</th>
<th>JCV Coefficient (From the PAQ)</th>
<th>Criterion Validity Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cust. Billing Analyst</td>
<td>241.267-034</td>
<td>40</td>
<td>.25</td>
<td>.17</td>
</tr>
<tr>
<td>¹Gas Storage Tech.</td>
<td>930.167-010</td>
<td>25</td>
<td>.33</td>
<td>.48</td>
</tr>
<tr>
<td>Instrument Mech.</td>
<td>710.281-026</td>
<td>25</td>
<td>.33</td>
<td>.48</td>
</tr>
<tr>
<td>Ld. Cust. Serv. Rep.</td>
<td>239.137-014</td>
<td>--</td>
<td>.27</td>
<td>--</td>
</tr>
<tr>
<td>²Planning Assistant</td>
<td>820.361-010</td>
<td>66</td>
<td>.32</td>
<td>.53</td>
</tr>
<tr>
<td>²Planning Aide</td>
<td>820.361-010</td>
<td>66</td>
<td>.32</td>
<td>.53</td>
</tr>
<tr>
<td>²Planning Tech</td>
<td>018.261-010</td>
<td>66</td>
<td>.32</td>
<td>.53</td>
</tr>
<tr>
<td>Sys. Gas Dispatch.</td>
<td>953.167-010</td>
<td>--</td>
<td>.3F0</td>
<td>--</td>
</tr>
</tbody>
</table>

Meta-Analysis 87%³

¹ Validity coefficients are uncorrected and statistically significant.
² Validity was transported from Instrument Mechanic.
³ Jobs are part of a progression that promotes from Assistant to Technician within five years.
⁴ Therefore, the same test battery is used for the entire progression and is only administered at the Planning Assistant level.
⁵ 87% of the variance can be accounted for by sampling error.
⁶ All incumbents in this position (33) were invited to participate in the validation study. Twenty-five employees cooperated with the validation effort.

Note: The Technical Reports for each validation study can provide further information on the predictors, criteria and validity coefficients. Contact the author for more information.
the people taking the test later in the year could have an advantage over those who took it earlier. Alternate forms can also discourage cheating (Crocker & Algina, 1986). Candidates often sit next to one another in a testing session. If the candidates have different test questions, they are prevented from looking at one another’s answers. Lastly, The Company wanted to use automated scoring. The Adaptability Test is fill-in-the-blank, which is not compatible with scannable answer sheets. Because SRA has not published an alternate form of the Adaptability Test, it is the purpose of this study to create three valid alternate forms in a format compatible with automated scoring.

Cronbach and Meehl (1955) describe several lines of evidence that help support construct validity. Calculating correlation matrices and a factor analysis can provide supporting evidence if the correlations between the tests are high and the factor structures are similar. Another possible line of evidence that can be offered is a study of the tests’ internal structures. One of several measures of internal-consistency reliability (homogeneity of the items) should be demonstrated before validity is claimed. Along these same lines, test-retest reliability is also desirable (Rosenthal & Rosnow, 1991).

Spiker and McCandless (1954) offer an additional suggestion: “If a new test is demonstrated to predict the scores on an older, well-established test, then an evaluation of the predictive power of the older test may be used for the new one” (p. 266). Cronbach and Meehl (1955) add that in order for this logic
to be accepted, the two tests must correlate so highly that there is “negligible reliable variance in either test” (p. 285). It is generally agreed that a correlation of approximately .80 is considered sufficient to meet this criterion (Brown, Sherbenou & Dollar, 1982; Martin, Blair, & Bledsoe, 1990). Therefore, if an existing test has been validated and a new test correlates with the existing test at about .80, the validity of the existing test may be extended to the new test.

Criterion and job component validation studies were conducted by The Company demonstrating significant predictive validity of the Adaptability Test (see Table 1 for the validity coefficients). Therefore, validity of the three alternate forms may be further supported by correlating the alternate forms with the original Adaptability Test.

**General Problem Solving Test**

The General Problem Solving Test (GPST) is a speeded paper-and-pencil test which measures cognitive abilities and mental adaptability. It was designed to assess skills that are important for the successful performance of work behaviors in field and clerical positions. The GPST was developed by Company staff as an alternate form of SRA’s Adaptability Test designed with questions that are directly relevant to the utility industry.

The GPST has three alternate forms. Each form contains 35 questions which collectively measure verbal, numerical, and analytical problem solving
abilities with an administration time of 15 minutes for each form. Item types were based on research by Spearman (1904, 1927) and Thurstone (1938).

Spearman theorized that general mental ability, or $g$, runs through all abilities; therefore, a single test that is highly saturated in $g$ could be substituted for a heterogeneous collection of tests trying to measure specific abilities ($s$). Spearman also recognized that similar abilities correlated even higher than what can be attributed to the $g$ factor. Based on this, he proposed that there might be another set of factors that are not as universal as $g$ and not as specific as $s$. These factors, which demonstrate a correlation with some but not all activities, were designated “group factors”. Spearman suggested arithmetic, mechanical and linguistic abilities as possible group factors.

Thurstone (1938) expanded Spearman's list of group factors through extensive research by himself and his students. He proposed about a dozen group factors which he called “primary mental abilities.” His findings have been corroborated by several researchers including Thurstone and Thurstone (1941), French (1951) and Harman (1975).

The identified group factors which were used to write the GPST alternate forms were: Verbal Comprehension (verbal analogies, disarranged sentences, verbal reasoning, and proverb matching); Numerical Facility (speed and accuracy of simple arithmetic computations); and Induction/General Reasoning (number series completions). These specific group factors were chosen
because they are most relevant to the job requirements being considered. Prior research has shown that measurement of these group factors provides an accurate assessment of general mental ability (Anastasi, 1982).

**Hypothesis**

It is hypothesized that validity will be demonstrated for each of the three forms of the General Problem Solving Test using several lines of validity evidence. Specifically, the Adaptability Test and the three forms of the GPST will have a correlation of at least .80, the recommended level for parallel forms.

**METHODS**

**Phase I: Item Development**

The GPST items parallel those found in SRA's Adaptability Test. The item types are similar to those identified by Thurstone's (1938) analysis of "primary mental abilities" that could be reasonably included in a test of this length and type. Although the items in the original Adaptability Test were constructed as fill-in-the-blank questions, the GPST items were written in a multiple-choice format to allow the answer sheets to be computer scanned.

One hundred and twenty-two items were constructed following the format of the Adaptability Test. The test questions were written in accordance with standard item-writing principles to insure the items were truly testing
candidates' knowledge (Kline, 1986). The newly-created items were reviewed by the Personnel Research staff at The Company for readability, clarity, bias and accuracy and minor revisions were made prior to piloting the tests.

Initial Pilot

Two forms (1 and 2) containing all 122 items were compiled for the pilot study. The order of the items in Form 1 was reversed in Form 2 to counterbalance for exhaustion. Form 1 was administered to five human resource employees to ensure the questions made sense to the test-taker and that there was only one correct answer for each question.

In November of 1994, the two forms were piloted with college students. Fifty-one students from a large southwestern university volunteered to take the GPST. The students ranged in age from 18-36 with approximately 85% in their freshman year. All students were compensated $20 and also received extra credit in a psychology class for their participation. Additionally, 101 students from a large midwestern university volunteered to participate and were also compensated $20 each and given class credit. These students ranged in age from 18-35 with 81% in their freshman or sophomore years.

All students were administered either Form 1 or Form 2 with standardized instructions (see Appendix A). The testing sessions were untimed.
but had a maximum time limit of 1 hour and 30 minutes. All but two students finished the test in the allotted time with most finishing within an hour.

Item Analysis

Data from the untimed pilot study were used to compute item statistics. Item discrimination, item difficulty and distracter effectiveness were determined. This information was used to select the final items for the GPST.

1. **Item Discrimination.** This was calculated by using the item-total score point-biserial correlation coefficient due to the dichotomous nature of the item (correct or incorrect) and the continuous nature of the possible total score (0-35) (see Appendix B for the resulting item-total correlations). Because the initial pilot contained 122 questions and only 105 were needed, a method was needed to eliminate items. A top-down approach was used and eight items (which all had a correlation of .15 or lower) were discarded.

2. **Item Difficulty.** For SRA's Adaptability Test, the overall difficulty level is .55 (with p-values ranging from .21 to .96). However, this number was derived using a timed test without multiple choice options. For this untimed, multiple choice pilot study, the difficulty level was .71 (with p-values ranging from .18 to .97). It is assumed that the difficulty will
increase once a time constraint is applied to the test because test takers will have less time to spend on each question.

3. **Distracter Effectiveness.** Distracters for each question were reviewed for frequency of selection. Six additional questions with unchosen distracters were eliminated.

Reliability estimates were also calculated on the pilot test data. Typically, reliability estimates would be calculated on the validation study data. However, due to the speeded nature of the tests administered during the validation study, internal consistency calculations (e.g. KR-20) are inaccurate representations of reliability. A KR-20 estimate will be artificially inflated because unfinished questions will correlate perfectly, regardless of whether the items are homogenous in content (Rosenthal & Rosnow, 1991). Therefore, the parallel forms method was used to determine the reliability of the three General Problem Solving Tests. Reliability estimates were able to be calculated from the initial pilot study because each student completed all three forms.

Individuals' total scores on each form were correlated and the average of those correlations was used as the reliability estimate. Although no hard and fast rules exist for what constitutes a minimally acceptable reliability value, most test developers agree that a reliability coefficient should be .80 or higher (Crocker & Algina, 1986). Reliability estimates calculated for the GPST alternate forms were .97 for Form A, .96 for Form B, and .97 for Form C.
SRA also used the parallel form method to determine the reliability for the Adaptability Test. Two forms were correlated and the resulting reliability was .89.

Development of the GPST Alternate Forms

The remaining 108 questions from the 122 item pool were divided into three alternate forms: GPST A, GPST B, and GPST C, each consisting of thirty-five items. Three questions from the item pool were unused and reserved as replacement questions. Test items were matched as closely as possible with respect to difficulty and discrimination and randomly assigned to one of three alternate forms in order of increasing difficulty. The average difficulty levels of the untimed versions of GPST A, GPST B, and GPST C were .72, .71, and .71, respectively (see "Results" for a re-calculation of item difficulty under timed conditions). Items for all tests were arranged in a spiral omnibus format to match SRA's Adaptability Test.

Phase II: Piloting of the GPST A, GPST B and GPST C

Two hundred and three Company non-exempt employees who had taken the SRA Adaptability Test in the past year were invited to participate in the pilot study. This pilot group was selected because it is part of the population that the GPST may assess in future selection.
One hundred and eighty-one employees volunteered to participate and were administered one form of the GPST on Company time. Participants were told the purpose of this pilot study and that their scores would remain confidential and only be used for research. They were then read standardized instructions (see Appendix C) and given 15 minutes to complete the 35 questions. Within two weeks of the testing session, participants were sent feedback sheets informing them of how many questions they answered correctly and the pilot group’s mean score.

Analyses

Several statistical analyses were calculated to support the validity of the GPST. Descriptive statistics, such as means and standard deviations were determined for both the Adaptability Test and the GPST. Ideally, descriptive statistics should be very similar on all four tests.

Difficulty levels, or p-values, were re-calculated using total score data from the timed conditions. A total score p-value of .50 will maximize variance and best distinguish proficient and poor performers (Crocker & Algina, 1986), which is the goal of selection tests. However, when alternate forms are being created, the difficulty levels of the alternate forms should be similar to that of the original test, regardless of its p-value. In this particular case, the item format of the Adaptability Test and GPST differ in that answer options are
provided in the GPST but not in the Adaptability Test. Given this item format difference, the author expects the GPST forms to have a slightly higher p-value due to correct answers achieved by random guessing or elimination strategies.

Total score of the Adaptability Test and GPST were correlated. Because these tests were intended to be parallel forms, correlations should be much higher than one would expect in a criterion-related study. In fact, correlations should actually look more like reliability estimates and approximate .80.

Group norms of each test were identified to determine comparable scores. This allows a test administrator to equate a total score on the Adaptability Test with a total score on any of the three GPST forms. This information is especially helpful if it is easier to obtain a higher total score on one test than on a parallel form (or vice versa). Group norm comparisons “level the playing field” in that different total scores are considered equal if one test is more difficult than the other.

RESULTS

Data Screening/Cleaning

Following standard research principles, data were screened for anomalies to ensure accuracy of test scores. An extreme outlier was found in Form C (4.1 standard deviations from the mean). This participant’s test results were subsequently eliminated from the data set for all statistical calculations. A
scatter plot of each participant’s scores on the Adaptability Test and relevant GPST form were produced and analyzed for linearity and homoscedacity (see Appendices D, E, and F). Forms A and B showed normally distributed data; however, as stated above, an extreme outlier was found in Form C.

Descriptive Statistics

Means and standard deviations were calculated for both the Adaptability Test and the GPST. As shown in Table 2, the average participants’ scores were about one standard deviation higher on all three forms of the GPST when compared to average scores on the Adaptability Test.

Each form’s difficulty level was re-calculated under the timed conditions. As predicted, the difficulty level increased from the untimed administration but was still less challenging than the original Adaptability Test. The mean of each GPST was approximately 5.5 points, or one standard deviation, higher than the Adaptability Test. This was expected, as participants had multiple choice answer options which allowed them to either randomly guess or recognize correct answers on the GPST but had no options to select from on the Adaptability Test, and thus had to recall correct answers. The multiple choice format makes the GPST an easier test. Once the GPST was corrected for guessing, the difficulty levels were almost exactly equal to that of the Adaptability Test. The difficulty levels can be found in Table 3.
<table>
<thead>
<tr>
<th>TEST</th>
<th>N</th>
<th>MEAN</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptability</td>
<td>180</td>
<td>17.6</td>
<td>5.3</td>
</tr>
<tr>
<td>GPST-Overall</td>
<td>180</td>
<td>22.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST</th>
<th>N</th>
<th>MEAN</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptability</td>
<td>62</td>
<td>17.2</td>
<td>5.7</td>
</tr>
<tr>
<td>GPST-Form A</td>
<td>62</td>
<td>21.7</td>
<td>5.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST</th>
<th>N</th>
<th>MEAN</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptability</td>
<td>57</td>
<td>17.3</td>
<td>5.4</td>
</tr>
<tr>
<td>GPST-Form B</td>
<td>57</td>
<td>21.6</td>
<td>5.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST</th>
<th>N</th>
<th>MEAN</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptability</td>
<td>61</td>
<td>18.3</td>
<td>4.7</td>
</tr>
<tr>
<td>GPST-Form C</td>
<td>61</td>
<td>22.6</td>
<td>5.3</td>
</tr>
</tbody>
</table>
Table 3

Difficulty Levels

<table>
<thead>
<tr>
<th>Test</th>
<th>p-Value (Difficulty Level)</th>
<th>p-Value (Corrected for Guessing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPST Form A</td>
<td>.62</td>
<td>.53</td>
</tr>
<tr>
<td>GPST Form B</td>
<td>.61</td>
<td>.52</td>
</tr>
<tr>
<td>GPST Form C</td>
<td>.64</td>
<td>.54</td>
</tr>
<tr>
<td>Adaptability Test</td>
<td>.55</td>
<td>.55</td>
</tr>
</tbody>
</table>
Adaptability Test total scores were correlated with GPST total scores. Additionally, effect sizes were calculated to determine the proportion of variance accounted for by the Adaptability Test. The resulting correlations and effect sizes are listed in Table 4.

Recent research has highlighted the effects that statistical artifacts have on test validites (Nunnally, 1978; Hunter, Schmidt, & Jackson, 1982). Artifacts include factors such as criterion unreliability, sampling error, and range restriction. The Division 14 Principles (1986) suggest that adjustments to validity may provide a clearer picture of the true operational validity of a predictor and endorses this practice. One guideline should be noted, corrections should only be made to validites which are significant. Also, uncorrected validites should always be presented along with corrected validites.

The validity study of the GPST utilized current employees as participants. These employees have already passed a selection battery and therefore will produce a smaller range of scores than what would be expected in the general population. The GPST may be used to select qualified applicants. For this reason, range restriction corrections were applied to the validity coefficients (see Appendix G for the formula and estimate of true variance). With the correction applied, correlations between the Adaptability Test and GPST Forms A, B, and C increased to .88, .86, and .90, respectively.
Table 4
Correlations and Effect Sizes of General Problem Solving Test Total Scores and Adaptability Test Total Scores

<table>
<thead>
<tr>
<th></th>
<th>Adaptability Test and GPST-Form A</th>
<th>Adaptability Test and GPST-Form B</th>
<th>Adaptability Test and GPST-Form C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrected</td>
<td>.76</td>
<td>.73</td>
<td>.76</td>
</tr>
<tr>
<td>Corrected for Range Restriction</td>
<td>.88</td>
<td>.86</td>
<td>.90</td>
</tr>
<tr>
<td>Effect Size</td>
<td>77%</td>
<td>74%</td>
<td>81%</td>
</tr>
<tr>
<td>N</td>
<td>62</td>
<td>57</td>
<td>61</td>
</tr>
</tbody>
</table>
These corrected validites represent the upper bound of validity for the GPST forms. If these alternate forms are used as a promotional selection tool, range restriction corrections can not be justified and therefore the uncorrected validity coefficients would represent the lower bound of the GPST's validity.

A norm table comparing raw test scores of the Adaptability Test and the GPST can be found in Table 5. As previously explained, group norms identified for each test help determine comparable scores. For example, if the cutoff score is 19 on the Adaptability Test, the cutoff score would be 24 on all GPST forms.

**DISCUSSION**

**Hypothesis**

A construct validation approach may be used to demonstrate the validity of alternate forms. Cronbach and Meehl (1955) discussed how to validate a test by creating a "nomological network" of indirect validity evidence. That is, the validity of a test can be supported if there are several lines of converging evidence all pointing to the same conclusion. Campbell and Fiske (1959) elaborated on this idea by suggesting some ways to obtain independent converging lines of evidence.

In this study, means, standard deviations, reliabilities, correlations and norm tables were used to demonstrate the comparability of the GPST alternate
<table>
<thead>
<tr>
<th>Percentile</th>
<th>Form A</th>
<th>Form B</th>
<th>Form C</th>
<th>Average</th>
<th>General Problem Solving Test</th>
<th>Adaptable Test Form A</th>
<th>Adaptable Test Form B</th>
<th>Adaptable Test Form C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1-1.7</td>
<td>1-1.7</td>
<td>1-1.4</td>
<td>1.0</td>
<td>6.7</td>
<td>9.0</td>
<td>11.3</td>
<td>12.3</td>
</tr>
<tr>
<td>1.5</td>
<td>2.2</td>
<td>2.2</td>
<td>2.1</td>
<td>1.9</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>2.0</td>
<td>2.1</td>
<td>2.1</td>
<td>2.0</td>
<td>2.0</td>
<td>19.2</td>
<td>20.0</td>
<td>18.7</td>
<td>18.7</td>
</tr>
<tr>
<td>2.5</td>
<td>2.1</td>
<td>2.1</td>
<td>2.0</td>
<td>2.0</td>
<td>19.2</td>
<td>20.0</td>
<td>18.7</td>
<td>18.7</td>
</tr>
<tr>
<td>3.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.9</td>
<td>1.9</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>3.5</td>
<td>1.8</td>
<td>1.8</td>
<td>1.7</td>
<td>1.7</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>4.0</td>
<td>1.7</td>
<td>1.7</td>
<td>1.6</td>
<td>1.6</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>4.5</td>
<td>1.6</td>
<td>1.6</td>
<td>1.5</td>
<td>1.5</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>5.0</td>
<td>1.5</td>
<td>1.5</td>
<td>1.4</td>
<td>1.4</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>5.5</td>
<td>1.4</td>
<td>1.4</td>
<td>1.3</td>
<td>1.3</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>6.0</td>
<td>1.3</td>
<td>1.3</td>
<td>1.2</td>
<td>1.2</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>6.5</td>
<td>1.2</td>
<td>1.2</td>
<td>1.1</td>
<td>1.1</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>7.0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>1.0</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>7.5</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>8.0</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>8.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td>9.0</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.6</td>
<td>12.6</td>
<td>13.1</td>
<td>12.7</td>
<td>12.7</td>
</tr>
</tbody>
</table>

*denotes raw scores
forms to the Adaptability Test. The standard deviations for the Adaptability Test and all three alternate forms were similar as expected (approximately 5.5). The mean of each GPST was approximately 5.5 points, or one standard deviation, higher than the Adaptability Test. This result does not impact the ability to use the GPST as alternate forms of the Adaptability Test because the alternate forms all differ from the original form in a standard manner. Therefore, The Company can simply set the GPST cut score one standard deviation higher than the cut score for the Adaptability Test. This allows the passing scores for the Adaptability Test and GPST to still be equivalent. The norm table can also be used to compare mean total scores and confirm the equated passing point.

While there is no minimally acceptable value for alternate form reliability estimates, many test publishers report coefficients ranging in the .80s and .90s for this type of reliability (Crocker & Algina, 1986). The reliabilities for the GPST A, B, and C were determined to be .97, .96, and .97, respectively. These values more than meet the standard cited by Crocker and Algina; in fact, the reliability of the original Adaptability Test was only .89.

Correlations between each GPST total score and the Adaptability Test total score were calculated to help support construct validity. These correlations are sometimes called “coefficients of equivalence” and were expected to resemble reliability coefficients in size; that is, they should be around the .80s (Crocker & Algina, 1986, p. 132). The total score uncorrected
correlations in this study are .76, .73, and .76 for GPST A, B, and C, respectively.

As was discussed in the Results section, the GPST was validated with current employees. If the GPST is used to test applicants, range restriction will exist and test validites should be corrected for statistical artifacts. The corrected correlation coefficients for the three alternate forms increased to .88 (Form A), .86 (Form B), and .90 (Form C). These equivalency coefficients showed a high positive correlation between the Adaptability Test and any form of the GPST. Additionally, calculations of effect size demonstrated that 77%, 74%, and 81% of the variance was accounted for by the Adaptability Test in GPST Forms A, B, and C, respectively. These results support the validity of the alternate forms. Table 6 provides a summary of the analyses comparing the Adaptability Test and the General Problem Solving Test.

Given these various lines of evidence, support has been demonstrated for the hypothesis, namely that the three General Problem Solving Tests are valid alternate forms of the Adaptability Test. However, as Cronbach (1955) points out, you can not say a "test has construct validity, because validation is a lengthy, even endless process" (p. 281). Therefore, the author concludes that the evidence to date is consistent with validity being demonstrated.
Table 6

Summary of Analyses Comparing the Adaptable Test and the General Problem Solving Test

<table>
<thead>
<tr>
<th></th>
<th>ADAPTABILITY</th>
<th>GPST A</th>
<th>GPST B</th>
<th>GPST C</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN (avg. mean)</td>
<td>17.6</td>
<td>21.7</td>
<td>21.6</td>
<td>22.6</td>
</tr>
<tr>
<td>STANDARD DEVIATION (avg. SD)</td>
<td>5.3</td>
<td>5.7</td>
<td>5.6</td>
<td>5.3</td>
</tr>
<tr>
<td>RELIABILITY ESTIMATE</td>
<td>.89</td>
<td>.97</td>
<td>.96</td>
<td>.97</td>
</tr>
<tr>
<td>DIFFICULTY LEVEL</td>
<td>.55</td>
<td>.62</td>
<td>.61</td>
<td>.64</td>
</tr>
<tr>
<td>CORRELATION (CORRECTED)</td>
<td>.88</td>
<td>.86</td>
<td>.90</td>
<td></td>
</tr>
<tr>
<td>EFFECT SIZE</td>
<td>77%</td>
<td>74%</td>
<td>81%</td>
<td></td>
</tr>
</tbody>
</table>
Future Research

It is recommended that future research be conducted on these alternate forms. Specifically, more research demonstrating additional lines of evidence (e.g., comparing supervisor or performance ratings with GPST scores) may further support the claim of validity. Campbell and Fiske (1959) also suggest that one should seek out and investigate situations where the interpretation is not supported. That is to look for disconfirming evidence and “plausible rival hypotheses” (p. 81) instead of only seeking out confirming evidence.

The General Problem Solving Tests are cognitive ability tests. Tests of this nature have consistently been shown to predict job performance (Hunter, 1983a, 1983c; Schmidt, Hunter, McKenzie & Muldrow, 1979; Thorndike, 1985). It is worth noting, however, that such tests predict considerably less than half the variance of job-related measures. Additionally, research has shown cognitive ability tests consistently demonstrate adverse impact on some protected groups (Hunter & Schmidt, 1978; Halpern, 1992; Jensen, 1980). Other individual characteristics such as interpersonal skills and aspects of personality may add incremental validity and exhibit less adverse impact, but at this point we do not have equally reliable instruments to measure them (Neisser, 1976). These other factors, however, should also be assessed when considering candidates for employment. As valuable as tests are, they should only be one component of the hiring process (Soloman, 1993).
Appendix A

Standardized Pilot Study Test Instructions

INSTRUCTIONS

The purpose of this test is to look at a job candidate’s strengths in skills such as math, logical reasoning, and vocabulary. You are taking part in a study to validate this test. That is, test developers want to ensure that the test is measuring the skills it is intended to measure. It is critical that you do your best in order to provide accurate data.

There is no time limit for this test. However, you should not spend more than 1 hour and 30 minutes on this test. It is important that you work as quickly and accurately as possible.

There is no penalty for guessing. Your final score is determined by adding together all correct answers. Therefore, it is to your advantage to answer each question.

Write your answer on the blank line that precedes each question.

No calculators will be allowed, but you may write on the test.

SAMPLE QUESTIONS

_____ 1. What is the last letter of a 4-letter word meaning walking stick?
   a) h      b) t      c) f      d) e      e) d

_____ 2. If Pete walks 3 blocks to work and Lori walks 5 blocks to work, how many more blocks does Lori walk to work than Pete?
   a) 1      b) 2      c) 3      d) 8      e) 15

_____ 3. What number is missing in this series? 1 - 3 - 5 - 7 - (?)
   a) 5      b) 6      c) 7      d) 8      e) 9
### Appendix B

**Item-Total Correlations Calculated on the Pilot Study Data**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>CORR</th>
<th>ITEM</th>
<th>CORR</th>
<th>ITEM</th>
<th>CORR</th>
<th>ITEM</th>
<th>CORR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.18</td>
<td>18</td>
<td>.31</td>
<td>35</td>
<td>.39</td>
<td>52</td>
<td>.37</td>
</tr>
<tr>
<td>2</td>
<td>.38</td>
<td>19</td>
<td>.32</td>
<td>36</td>
<td>.47</td>
<td>53</td>
<td>.34</td>
</tr>
<tr>
<td>3</td>
<td>.10</td>
<td>20</td>
<td>.38</td>
<td>37</td>
<td>.39</td>
<td>54</td>
<td>.38</td>
</tr>
<tr>
<td>4</td>
<td>.42</td>
<td>21</td>
<td>.32</td>
<td>38</td>
<td>.29</td>
<td>55</td>
<td>.42</td>
</tr>
<tr>
<td>5</td>
<td>.11</td>
<td>22</td>
<td>.22</td>
<td>39</td>
<td>.10</td>
<td>56</td>
<td>.04</td>
</tr>
<tr>
<td>6</td>
<td>.42</td>
<td>23</td>
<td>.37</td>
<td>40</td>
<td>.13</td>
<td>57</td>
<td>.19</td>
</tr>
<tr>
<td>7</td>
<td>.28</td>
<td>24</td>
<td>.50</td>
<td>41</td>
<td>.27</td>
<td>58</td>
<td>.24</td>
</tr>
<tr>
<td>8</td>
<td>.28</td>
<td>25</td>
<td>.42</td>
<td>42</td>
<td>.35</td>
<td>59</td>
<td>.31</td>
</tr>
<tr>
<td>9</td>
<td>.37</td>
<td>26</td>
<td>.37</td>
<td>43</td>
<td>.44</td>
<td>60</td>
<td>.31</td>
</tr>
<tr>
<td>10</td>
<td>.17</td>
<td>27</td>
<td>.35</td>
<td>44</td>
<td>.56</td>
<td>61</td>
<td>.28</td>
</tr>
<tr>
<td>11</td>
<td>.37</td>
<td>28</td>
<td>.07</td>
<td>45</td>
<td>.40</td>
<td>62</td>
<td>.29</td>
</tr>
<tr>
<td>12</td>
<td>.18</td>
<td>29</td>
<td>.24</td>
<td>46</td>
<td>.34</td>
<td>63</td>
<td>.58</td>
</tr>
<tr>
<td>13</td>
<td>.26</td>
<td>30</td>
<td>.33</td>
<td>47</td>
<td>.26</td>
<td>64</td>
<td>.46</td>
</tr>
<tr>
<td>14</td>
<td>.42</td>
<td>31</td>
<td>.29</td>
<td>48</td>
<td>.35</td>
<td>65</td>
<td>.31</td>
</tr>
<tr>
<td>15</td>
<td>.39</td>
<td>32</td>
<td>.29</td>
<td>49</td>
<td>.22</td>
<td>66</td>
<td>.42</td>
</tr>
<tr>
<td>16</td>
<td>.12</td>
<td>33</td>
<td>.38</td>
<td>50</td>
<td>.40</td>
<td>67</td>
<td>.23</td>
</tr>
<tr>
<td>17</td>
<td>.31</td>
<td>34</td>
<td>.20</td>
<td>51</td>
<td>.33</td>
<td>68</td>
<td>.27</td>
</tr>
</tbody>
</table>
Appendix B (Continued)

Item-Total Correlations Calculated on the Pilot Study Data

<table>
<thead>
<tr>
<th>ITEM</th>
<th>CORR</th>
<th>ITEM</th>
<th>CORR</th>
<th>ITEM</th>
<th>CORR</th>
<th>ITEM</th>
<th>CORR</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>.35</td>
<td>70</td>
<td>.53</td>
<td>71</td>
<td>.36</td>
<td>72</td>
<td>.34</td>
</tr>
<tr>
<td>73</td>
<td>.18</td>
<td>74</td>
<td>.30</td>
<td>75</td>
<td>.32</td>
<td>76</td>
<td>.21</td>
</tr>
<tr>
<td>77</td>
<td>.22</td>
<td>78</td>
<td>.44</td>
<td>79</td>
<td>.34</td>
<td>80</td>
<td>.46</td>
</tr>
<tr>
<td>81</td>
<td>.29</td>
<td>82</td>
<td>.10</td>
<td>83</td>
<td>.34</td>
<td>84</td>
<td>.43</td>
</tr>
<tr>
<td>85</td>
<td>.45</td>
<td>86</td>
<td>.33</td>
<td>87</td>
<td>.26</td>
<td>88</td>
<td>.13</td>
</tr>
<tr>
<td>89</td>
<td>.27</td>
<td>90</td>
<td>.38</td>
<td>91</td>
<td>.26</td>
<td>92</td>
<td>.22</td>
</tr>
<tr>
<td>93</td>
<td>.39</td>
<td>94</td>
<td>.40</td>
<td>95</td>
<td>.42</td>
<td>96</td>
<td>.23</td>
</tr>
<tr>
<td>97</td>
<td>.03</td>
<td>98</td>
<td>.41</td>
<td>99</td>
<td>.22</td>
<td>100</td>
<td>.34</td>
</tr>
<tr>
<td>101</td>
<td>.25</td>
<td>102</td>
<td>.21</td>
<td>103</td>
<td>.12</td>
<td>104</td>
<td>.16</td>
</tr>
<tr>
<td>105</td>
<td>.37</td>
<td>106</td>
<td>.34</td>
<td>107</td>
<td>.37</td>
<td>108</td>
<td>.36</td>
</tr>
<tr>
<td>109</td>
<td>.42</td>
<td>110</td>
<td>.43</td>
<td>111</td>
<td>.42</td>
<td>112</td>
<td>.19</td>
</tr>
<tr>
<td>113</td>
<td>.38</td>
<td>114</td>
<td>.35</td>
<td>115</td>
<td>.45</td>
<td>116</td>
<td>.44</td>
</tr>
<tr>
<td>117</td>
<td>.29</td>
<td>118</td>
<td>.27</td>
<td>119</td>
<td>.29</td>
<td>120</td>
<td>.33</td>
</tr>
<tr>
<td>121</td>
<td>.20</td>
<td>122</td>
<td>.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C

Standardized Validation Study Test Instructions

INSTRUCTIONS

The purpose of this test is to look at a job candidate’s strengths in skills such as math, logical reasoning, and vocabulary.

The time limit for this test is 15 minutes. This is a highly speeded test so it is important that you work as quickly and accurately as possible.

There is no penalty for guessing. Your final score is determined by adding together all correct answers. Therefore, it is to your advantage to answer each question.

Write your answer on the blank line that precedes each question.

No calculators will be allowed, but you may write on the test.

SAMPLE QUESTIONS

1. What is the last letter of a 4-letter word meaning walking stick?  
   a) h  b) t  c) f  d) e  e) d

2. If Pete walks 3 blocks to work and Lori walks 5 blocks to work, how many more blocks does Lori walk to work than Pete?  
   a) 1  b) 2  c) 3  d) 8  e) 15

3. What number is missing in this series?  1 - 3 - 5 - 7 - (?)  
   a) 5  b) 6  c) 7  d) 8  e) 9
Appendix D

Scatterplot of GPST and Adaptability Test Total Scores - Form A

Legend:  
A = 1 Observation  
B = 2 Observations  
C = 3 Observations
Appendix E

Scatterplot of GPST and Adaptability Test Total Scores - Form B

Legend: A = 1 Observation
B = 2 Observations
C = 3 Observations
Appendix F

Scatterplot of GPST and Adaptability Test Total Scores - Form C

Legend: A = 1 Observation
       B = 2 Observations
       C = 3 Observations

© denotes outlying test score deleted from data set
Appendix G

Range Restriction Formula

\[ r_c = \frac{r_{xy} (obs)}{\sqrt{r_{xx}} \sqrt{r_{yy}}} \]

Where:

- \( r_c \) = corrected correlation
- \( r_{xx} \) = restricted correlation (observed)
- \( r_{xy} \) = unrestricted correlation of the predictor
- \( r_{yy} \) = unrestricted correlation of the criterion (estimated at .89)
BIBLIOGRAPHY


