The feasibility of computerized cognitive testing as a surrogate measure for assessment center performance

1988

Leilani M. de Saram

University of Central Florida

Find similar works at: https://stars.library.ucf.edu/rtd

University of Central Florida Libraries http://library.ucf.edu

STARS Citation


https://stars.library.ucf.edu/rtd/4272

This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
THE FEASIBILITY OF COMPUTERIZED COGNITIVE TESTING
AS A SURROGATE MEASURE FOR ASSESSMENT CENTER PERFORMANCE

BY

LEILANI M. DE SARAM
B.S., Florida State University, 1985

THESIS

Submitted in partial fulfillment of the requirements
for the Master of Science degree in
Industrial/Organizational Psychology
in the Graduate Studies Program
of the College of Arts and Sciences
University of Central Florida
Orlando, Florida

Summer Term
1988
ABSTRACT

The use of microcomputerized cognitive testing of personnel has thus far been limited to military and environmental stressor applications. The investigation of such testing as a substitute for assessment center performance represents a novel application of this technology. An examination of the relationship between performance on a cognitive test battery and assessment center performance was conducted.

A battery of selected microcomputerized tests measuring cognitive constructs was administered to assessment center candidates in law enforcement occupations. Contrary to expectations, the computerized cognitive measurements did not exhibit significant correlations with assessment center performance measures. It was speculated that the two measures assessed diverse domains of performance, thus limiting the degree of potential overlap. Range restriction of the criterion measure was cited as a methodological concern in this study.
ACKNOWLEDGEMENTS

I wish to extend my deep appreciation to my thesis committee, Dr. Norman Lane, Dr. William Wooten, and especially Dr. Janet Turnage, my chair, for their sound guidance and genuine support of this project. I also wish to thank Tom Rockhill, Maurie Bosse, and Andrea Williams of the Orange County Sheriff's Office for providing the personnel, facilities, and help needed to conduct my research. A special thank you goes to the staff of Essex Corporation, whose resources and chief programmer, Martin G. Smith, made this research possible. I am grateful to my friends, and in particular, my family for their continued support, beyond the scope of this project.
# TABLE OF CONTENTS

**LIST OF TABLES** ................................................................. v

**INTRODUCTION** .............................................................. 1
  Assessment Centers ....................................................... 2
  Computerized Cognitive Assessment ................................. 8
  Surrogate Measures ....................................................... 14
  Problem Statement ........................................................ 17
  Hypotheses ........................................................................ 18

**METHOD** .............................................................. 19
  Subjects ............................................................................ 19
  Materials ........................................................................... 19
  Procedure ........................................................................... 23
  Apparatus .......................................................................... 25

**RESULTS** .......................................................... 26

**DISCUSSION** ............................................................. 30

**APPENDIX** .............................................................. 34

**REFERENCES** ............................................................. 36
LIST OF TABLES

1. Means, Standard Deviations, and Test-Retest Reliabilities of Predictor Variables ............... 28
3. Intercorrelations Between Assessment Center Ratings and Test Average Number Correct Scores ... 29
INTRODUCTION

The use of assessment centers as a tool in selection is widespread and growing (Gaugler, Rosenthal, Thornton, & Bentson, 1987). However, the costs of implementing such programs in terms of time, development, and facilities remain an important issue. This paper proposes the examination of computerized cognitive testing as a surrogate measure for assessment center participation in order to provide a cost-effective, yet valid, alternative to assessment centers for selection purposes.

There is a general lack of research regarding the utility of computerized cognitive testing as a predictor of performance in occupations outside of the military. This type of measure, though non-traditional, could be shown to be a good predictor of assessment center performance. It would be worthwhile to conduct such research under the rationale of a surrogate measures approach. In surrogate measurement, computerized tests, which do not simulate actual job tasks, but tap cognitive constructs essential to job performance effectiveness, would be administered to assessment center candidates. The correlation of test performance scores with assessment center ratings would
confirm the utility of computerized cognitive tests as a potential surrogate for assessment center performance.

Assessment Centers

Assessment centers evaluate employees for the purpose of generating personnel decisions, such as selection, promotion, or development. Multiple assessment techniques (e.g., situational exercises, job samples, etc.) are employed. Trained raters evaluate candidate performance on the exercises according to a number of predetermined dimensions, in order to generate the dimension and overall scores upon which a final recommendation is made (Thornton & Byham, 1982).

The use of assessment centers, particularly the use of situational exercises, appears to be accepted by participants and supervisors alike (Thornton & Byham, 1982). It is presumed that this is due to its high face validity. Further support of the process was garnered by Huck (cited in Joiner, 1984), who in 1974 integrated the results of 50 assessment center studies to find that while random selection of a successful employee occurs at a probability of 15 percent, and managerial nomination of a successful employee occurs 35 percent of the time, assessment center recommendations plus managerial
nomination increase the probability of selecting a successful employee to 76 percent!

The true effectiveness of any evaluation technique, however, depends on its methodological soundness and psychometric properties (Thornton & Byham, 1982). This requires a review of assessment center reliability and validity, as well as an examination of assessment centers which employed a cognitive testing component. Interrater reliability reviewed across 150 studies of assessment centers is high. The majority of studies report coefficients of $r > .80$ for selected dimensions, such as planning and organizing, communications effectiveness, and decision-making skill (Thornton & Byham, 1982).

In contrast, the internal consistency estimates of assessment centers are low (Thornton & Byham, 1982). This measure refers to the relationship among ratings of the same dimension across different exercises. Reported internal consistency coefficients range from $r = .04$ to $r = .69$ in studies conducted by Hinrichs and Haanpera (1976), Huett (1975), and Neidig et al. (1978) (as cited in Thornton & Byham, 1982). The internal consistency problem should be noted when reviewing validity data. This represents a deficiency in the operationalization of the center that affects its validity (Venkatraman & Grant, 1986). The magnitude of validity is dependent on the component of reliability.
Thornton and Byham (1982) provide extensive reviews of assessment center validity studies. Dimension ratings in several studies have been shown to be valid predictors of overall job performance of managers ($r > .30$). Overall ratings of potential in the landmark Management Progress Study were predictive of the future success of AT&T managers (Bray, Campbell, & Grant, 1974). Reports of overall ratings as valid predictors of performance, however, are inconsistent. Thornton & Byham (1982) report: (a) in validity studies with no feedback to candidates, all results support overall ratings as predictive of performance criteria across different jobs, time periods and subject samples; (b) of studies with control groups, assessed groups were rated higher in job performance criteria than control groups; and (c) in correlational studies with feedback, overall ratings are valid predictors of overall job performance. The validity coefficients corresponding to the studies described range from $r = .13$ to $r = .51$.

Recent research in assessment center validity reports contrasting results. Gaugler, Rosenthal, Thornton and Bentson (1987) identified three moderators in a meta-analysis of assessment center validity: Multiple ratings of candidates, psychologists as assessors, and the methodological soundness of the studies. It is possible that the earlier validity studies did not achieve the
latter criterion. Methodological soundness refers to the quality of the studies included in the meta-analysis. The predictive validity of assessment center evaluations was moderated by the research methodology applied. Studies must possess adequate sample representativeness, and differences in motivation, job experience, and training must be addressed to ensure predictive validity (Gaugler et al., 1987).

In an extensive examination of the predictive validity of assessment center ratings, Turnage and Muchinsky (1984) found that neither ratings nor traditional predictors, such as tenure, education, or tests of general ability, were related to supervisory job performance, although assessment centers were predictive of promotability. Low criterion reliability and low predictor reliability were posited as contributors to lack of predictive validity of the performance criterion.

In addition, the practice of discussion ("team meetings") in preparing final ratings has been questioned. Wingrove, Jones and Herriot (1985) noted no difference in the predictive validity of pre- versus post-discussion ratings of training suitability. The team meeting concept has come under scrutiny.

Finally, the lack of discriminant validity evidenced in current research points to a need for revisions in the multi-trait approach inherent to assessment centers.
Turnage and Muchinsky (1982) found little discriminability among specialized abilities, skills, or traits in assessment. Sackett and Dreher (1982) reported a correlation of zero for the same trait rated across different exercises. Their key finding was that the factor patterns loaded on exercises rather than dimensions. Within-exercise correlations were high, indicating a degree of generalizability among dimensions rated for a particular exercise.

In 1987, Robertson, Gratton, and Sharpley replicated the Sackett and Dreher findings. Due to the high correlations of dimensions within each exercise, it was suggested that exercises be designed to tap into generalized factors. This proposition was further supported by analyses of the General Motors Corporation Supervisory Selection assessment centers data for over 11,000 candidates (Outcalt, 1988). Paralleling earlier findings, assessors were apparently rating a global characteristic associated with each exercise rather than dimensions across exercises. It would seem plausible that microcomputerized tests, designed to tap into these global characteristics, would correlate with acceptable assessment center performance.

Comparisons of cognitive testing and assessment center ratings as valid predictors of performance are mixed. The cognitive tests used thus far are pencil-and-paper tests of
ability in combination with assessment center exercises. Thornton and Byham (1982) cite two studies in which assessment center ratings provided a unique contribution to the predictive validity of the performance criterion over the use of pencil-and-paper tests alone (Bray & Campbell, 1968; Bray & Grant, 1966).

The General Motors Study analysis (Outcalt, 1988), however, challenges these earlier findings. For 1200 candidates, a mechanical and a numerical test correlated with observations of performance with $r = .10$ and $.17$ ($p < .001$), respectively. The overall assessment center rating had a lower correlation with the same performance criterion, $r = .07$ ($p < .01$, Outcalt, 1988). In addition, the General Motors study showed a factor loading of pencil-and-paper tests in all factor analyses, and they appear in all multiple regressions as significant (Smith, 1988). The results of validities corrected for unreliability of predictor and criterion and range restriction, showed the numerical test alone had the highest predictive validity of all 22 assessment center variables for the performance of all units in the study but one. This particular unit, however, had extended assessor training requirements, documentation requirements, etc. It was recommended that the pencil-and-paper tests be used as a screening device for future assessment center candidates.
The preceding findings suggest a need for reassessment of the constructs being measured and the possible utility of cognitive testing. Microcomputerized testing, which accesses specialized domains, can offer a plausible way to assess constructs related to performance.

The proposal to investigate computerized cognitive testing as a surrogate measure is prompted by the above recommendation. More compelling still is the discovery that the use of microcomputerized cognitive testing of information processing domains (e.g., symbolic manipulation, reasoning) has not previously been conducted in an assessment center setting.

Computerized Cognitive Assessment

Meta-analysis has shown cognitive ability testing to be positively related to a number of job performance criteria (Hakel, 1986). Such tests are fair to minority and disadvantaged groups in that they do not underestimate the expected performance of these groups (Schmidt & Hunter, 1981). However, the use of cognitive testing for management selection is rare (Robertson & Makin, 1986).

There is evidence of a global factor, or general reasoning ability as identified by cognitive testing that is predictive of job performance across several occupational types (Campbell, Dunnette, Lawler, & Weick, 1970, as cited in Thornton & Byham, 1982; and Thorndike,
Tests of specialized abilities, however, have been more avidly researched in published literature. Traditional pencil-and-paper tests (e.g., numerical and mechanical ability, intelligence, and aptitude tests) exhibit high reliability estimates ranging from $r = .57$ to $.99$ (Department of Defense, 1972; Educational Testing Service, 1975; and The Psychological Corp., 1969). In terms of validity, tests of specialized abilities (verbal, numerical, and clerical) have been shown to correlate with job performance criteria at a range of $r = .26$ to $.66$ (Thorndike, 1985). However, results of a meta-analysis of validity studies published between 1964 and 1982 revealed that pencil-and-paper tests of specialized aptitudes and mental ability exhibit lower predictive validity for the criterion of job performance than work sample, assessment center, or supervisory evaluation predictors of performance (Schmitt et al., 1984).

A validity generalization model of written cognitive tests for law enforcement occupations rejected the situational specificity hypothesis for reasoning and spatial/mechanical tests in performance prediction (Hirsh, Northrop & Schmidt, 1986). The observed validities of such tests are generalizable to the prediction of job performance in law enforcement occupations. An average of 79% of the variance in performance criteria is accounted
for by cognitive testing (the minimum useful validity level was $r = .10$). It was expected that 86% of the cases utilizing reasoning tests and 69% of those using psychomotor tests achieved the minimum validity criterion of $r = .10$. However, predictors of the criterion of job performance for law enforcement occupations exhibit lower validities than do predictors of job performance for other occupational types. This may be due in part to the importance of the non-cognitive, interpersonal abilities required in law enforcement.

Note that the studies cited thus far involve utilization of pencil-and-paper ability testing. The use of microcomputerized cognitive testing of cognitive processing has been limited to military and biomedical applications. The Environmental Protection Agency, and the U.S. Army, Navy, and Air Force are using cognitive tests recently implemented on low-cost, portable microcomputers (see Kennedy, Lane, & Kuntz, 1987). The use of microcomputerized cognitive tests in law enforcement, public sector occupations represents a novel application of this technology.

Computerized cognitive testing differs fundamentally from traditional pencil-and-paper tests in terms of the ability domains tapped by each. Rather than assessing specialized, knowledge-based skills or aptitudes (e.g., verbal ability) as traditional pencil-and-paper tests do,
computerized cognitive tests tap higher-level information processing skills (Lane & Kennedy, 1988). Assessment of information processing skills (e.g., symbolic reasoning) may closer approximate the basic cognitive functions required in job performance than would pencil-and-paper tests of knowledge and aptitude.

In addition, computerized testing affords the advantage of enhanced operational validity (Byham & Temlock, 1972) over pencil-and-paper tests. In terms of administration and implementation, computerized testing is inherently more precise. Conventional pencil-and-paper testing is subject to a number of methodological problems: excessive administration times, scoring inconsistencies, poor identification of people with extreme abilities, high vulnerability to theft, and unintentional sharing of data (Hakel, 1986). Through programming, computerized testing is automated to provide standardized administrations and computer-timed test sessions, which result in higher reliabilities (Kennedy, Lane, & Kuntz, 1987). The reduced variability in test procedures allows researchers to more easily compare related studies. Scoring is accurate and objective, and a number of response measures are available (e.g., number correct, number attempted, response latency, etc.). The computers are capable of storing large amounts of diverse data and providing feedback for rapid data analysis (Kennedy, Lane, & Kuntz, 1987). Finally, data
storage is secure. Complex filing and accessing systems can be programmed as necessary.

The tests currently implemented on microcomputers are available from two batteries: (1) The Essex Automated Performance Test System (APTS); and (2) The Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB). The details of the development of these batteries can be found in Lane and Kennedy (1988). The APTS tests were originally the most appropriate of 114 tests reviewed for specific criteria and recommended for repeated measures research. Tests were evaluated on three major criteria: (a) rapid stabilization (< 10 minutes practice), (b) high reliability ($r > .707$ for three minutes of testing), and (c) lack of ceiling effects (Bittner, Smith, Kennedy, Staley, & Harbeson, 1985).

Psychometric studies of the APTS tests in comparison to pencil-and-paper tests have been highly conclusive. Kennedy, Wilkes, Lane, and Homick (1985) administered four replications of six APTS tests with pencil-and-paper versions of the same tests. All tests stabilized within four sessions and high reliability was obtained for each ($r > .707$ for three minutes of testing). The computerized tests were comparable in content to the pencil-and-paper tests.

The predictive validity of APTS and pencil-and-paper tests were compared using the Wechsler Adult Intelligence
Scale (WAIS) score as the criterion. For 10 replications of 11 tests, nine of the computerized tests stabilized, and their reliability for three minutes of testing was high, \( r > .76 \). Equivalency of computerized versus pencil-and-paper tests was achieved. The correlation between certain computerized tests and the WAIS identified common variance (Kennedy, Wilkes, Dunlap, & Kuntz, 1987). One should note that conventional paper-and-pencil tests do not tap factorially-pure processing skills as cognitive testing may (Fairbank, 1984, cited in Kyllonen, 1986).

The PAB battery consists of 25 tests (implemented on a microcomputer) which tap into information processing, cognition, and perception domains (Englund et al., 1987). Comparisons of the APTS and PAB batteries' psychometric properties have been made. Though PAB tests initially exhibited lower average reliability (\( r = .60 \)) than APTS tests (\( r = .80 \)), most PAB tests exhibit acceptable levels of test-retest reliability if given for slightly longer time periods (Turnage, Kennedy, & Osteen, 1987). These results were replicated by Tabler, Turnage, and Kennedy (1987) using additional PAB tests. A third related study found comparable psychometric results between the APTS and PAB batteries (Lane & Kennedy, 1988).

These studies led to a factor analysis of the APTS and PAB batteries across applications. Lane and Kennedy (1988) report that all tests of the two batteries loaded on three
basic tasks consistently: (1) Motor Speed - the speed of response execution, (2) Reasoning/Symbol Manipulation - the general ability to reason abstractly through the application of rules, and (3) Cognitive Processing Speed - the extent to which the rules governing response generation have been learned, and can be applied more or less rapidly. This latter factor also appears to involve elements of spatial manipulation.

**Surrogate Measures**

Given that microcomputerized cognitive testing does access specialized cognitive functioning, the foundation for a surrogate measures approach to proposed research has been established. The problem of poor internal consistency reliability of assessment center measures has been identified in the preceding review of the literature. Lane, Kennedy, and Jones (1986) documented the need for improved assessment of operational measures in the selection process. However, all assessment center research is subject to criterion distortion problems. Inadequate assessor training, procedural inconsistencies, and psychometric deficiencies, anchored by poor criterion reliability, are prevalent (Turnage & Muchinsky, 1984). Low reliability of the criterion measure greatly reduces the magnitude of the observed predictive validity. Lane et
al. (1986) illustrated the relationship using the correction for attenuation formula developed by Guilford (1954):\

\[
R_t = \frac{r_{xy}}{\left[(r_{xx})(r_{yy})\right]^{1/2}}
\]  

(1)

\(R_t\) is the estimated true relationship, \(r_{xy}\) is the predictive validity, \(r_{xx}\) is the reliability of the predictor, and \(r_{yy}\) is the reliability of the criterion.

Even with a high true relationship coefficient, a low criterion reliability measure in the denominator will consistently reduce the value of the predictive validity in the numerator. The criterion problem is inherent to the process of assessment. Therefore, Lane et al. (1986) suggest that rather than attempting to improve the reliability of the criterion measure, the option is to maximize the reliability of the predictor. A predictor which exhibits reliability somewhat higher than that of assessment centers, and which correlates with the constructs underlying actual performance, can potentially substitute for assessment center evaluations. Surrogate measures have been proposed as that alternative.
Surrogate measures are those which are predictive of a construct we wish to emphasize (e.g., "true" job performance), but do not directly measure that construct (Lane et al., 1986). Rather than being synthetic or job sample measures, surrogates tap into the constructs related to effective performance of the job (e.g., spatial processing).

A surrogate measure must, by definition, meet the following criteria: (a) acceptable correlation with the performance construct, (b) ability to detect performance changes (important for environmental applications), (c) achievement of higher reliability than operational measures, and (d) minimal time requirements with regard to learning the tasks and task practice (Lane et al., 1986). The microcomputerized tests reviewed in the preceding section exhibit these characteristics, with the exception of the as yet undetermined correlation with the performance construct. As stated earlier, the purpose of the proposed research is to examine that relationship in the context of assessment center performance. The greater reliability of surrogates over operational measures may tap more of the variance of the job performance criterion. Specifically, it is expected that the variance of highly complex tasks (assessment center performance) can be predicted from relatively simple microcomputerized tests. The requirement
is that the computerized tests must involve factors common to the operational task.

For the purpose of predicting performance of administrative skills, tests should be chosen which relate to those performance constructs. Tests which load on cognitive factors and functions, such as cognitive speed and symbolic reasoning, would be suggested. The test battery must be assembled to access the combination of cognitive components as required by the position. Test batteries can be weighted accordingly as well.

Problem Statement

The use of microcomputerized cognitive testing of personnel has been limited to military and environmental stressor applications (Lane & Kennedy, 1988). The investigation of such testing as a substitute for assessment center performance represents a novel application of the technology. Due to the costs and uncertain validity for performance prediction associated with assessment centers, there is a need to carefully examine alternate assessment methods. These methods could be used as valid substitutes for assessment centers. Potential users of alternate assessment methods include companies that cannot afford the expenses of assessment center implementation, but possess reliable job analysis data.
It is proposed that by administering a selected battery of microcomputerized tests measuring cognitive constructs during an assessment center, adequate information can be gathered to correlate with the respective assessment center performance measures. The following hypotheses are rendered:

H1: A core test battery consisting of APTS Grammatical Reasoning, PAB Pattern Comparison (simultaneous), and PAB Code Substitution will correlate significantly and in a positive direction with assessment center overall score ratings (OAR). PAB Mathematical Processing may replace PAB Code Substitution as the third significant predictor in this sample.

H2: Significant bivariate correlations will exist between computerized test scores and the unique dimension ratings of judgment, decisiveness, perception, and planning and organizing.
METHOD

Subjects

The subject pool consisted of 27 (25 male, 2 female) in-house assessment center candidates vying for the first-level supervisory promotional position of Sergeant within a metropolitan law enforcement agency. At the time of testing, 21 subjects held the position of Deputy and 6 subjects were Corporals.

Materials

All assessment center candidates were required to participate in four exercises: in-basket, problem solving, counseling, and computer test battery. Assessors evaluated performance in the first three exercises only, according to the following prespecified dimensions: judgment, decisiveness, leadership, organizing and planning, written communication, oral communication, perception, and interpersonal (see Appendix). The complex cognitive dimensions (i.e., judgment, decisiveness, perception, and planning and organizing) were the basis of selection of particular tests to be administered. Microcomputerized
tests which load on cognitive demands were administered in the order described:

**PAB Mathematical Processing.** This test was chosen as it assesses a cognitive speed factor (Lane & Kennedy, 1988). It requires the subject to perform addition/subtraction operations as presented on a display, and subsequently respond by indicating whether the arithmetic result is greater or less than five. Subjects respond using the UP and DOWN arrow keys respectively (Shingledecker, 1984). The digits 1 through 9 are used in the problems presented, with a maximum response time of 1.5 seconds. Two arithmetic operations were featured in the problems displayed.

**PAB Pattern Comparison (Simultaneous).** This test was chosen as an indicator of cognitive speed and spatial ability (Lane & Kennedy, 1988). The participant is required to view two adjacent patterns of eight dots and decide whether they are the "same" (identical) or "different" (Klein & Armitage, 1979). The subject is required to press the "S" key for same, the "D" key for different.

**APTS Grammatical Reasoning.** This test was selected as an indicator of symbolic reasoning ability (Lane & Kennedy, 1988). It requires the participant to
respond by pressing "T" for true, and "F" for false with regard to the verity of a statement as presented (Baddeley, 1968). Five types of grammatical transformations are employed on statements describing the relationship between the two letters, "A" and "B," as they are presented on the display. The five transformations are: (1) true versus false statements, (2) active versus passive voice, (3) affirmative versus negative phrasing, (4) mention of "A" versus "B" first, and (5) use of the verb "precedes" versus the verbs "trails" and "follows."

**APTS Non-Preferred-Hand Tapping.** This test assesses a purely motor domain, but is recommended as an indicant of motivation (Kennedy, Lane, & Kuntz, 1987). It served as a "mental break" being placed in the fourth position of seven tests. The subject is required to alternately press two keys with the nonpreferred hand as quickly as possible. This tests added only 20 seconds to the total test administration time.

**APTS Manikin.** This test was chosen as an indicator of symbolic manipulation ability (Lane & Kennedy, 1988). It presents a simulated human figure (sailor) facing either toward or away from the participant. The figure holds different patterns of three hearts,
diamonds or clubs, in each hand. The subject must decide which hand, left or right, of the figure holds the pattern upon which the sailor is standing (Benson & Gedye, 1963). The subject responds by pressing the left arrow key for the left hand, or the right arrow key for the right hand.

PAB Linguistic Processing. This test was selected because it requires cognitive speed functioning (Lane & Kennedy, 1988). The category match version of this test requires the participant to respond true or false, using the "T" and "F" keys, respectively, as to whether both letter pairs displayed meet a specified criterion. The criterion used in this experiment was that both letters within a pair must be of the same category (either consonants or vowels) and that both letters of the second pair are within the same category (the second pair does not have to be of the same category as the first pair) (Posner & Mitchell, 1967).

PAB Code Substitution. This test was chosen as an indicator of cognitive speed. Derived from the WAIS (Wechsler, 1958), it requires the subject to enter a corresponding number that is associated with a probe letter. A nine letter string is presented
continuously on the screen with a constant nine-digit string directly beneath. A single letter is flashed at the bottom of the screen. The subject must enter the number associated with the target letter.

**Procedure**

All candidates for the position of Sergeant were required to attend a three day assessment center. Completion of four exercises, including the test battery, was mandatory. The test battery was approved as an adjunct exercise of the assessment center, bearing no influence on the selection process. Assessor rating of the dimensions specified earlier, and subsequent overall ratings, were utilized in preparing final recommendations for the Sergeant position. The order of completion of the four exercises differed according to candidate schedules. Administration of the computer exercise was as follows:

Day 1: Subjects received an orientation followed by a 40 minute practice session of the test battery. During orientation, subjects were introduced to the administrator, who advised them of the purpose of the study (for research application only), provided assurance of confidentiality of results (by subject identification number only), and alerted them to keyboard features and special key locations. Subjects were then instructed to begin the
computer administered battery of tests which administered practice sessions on their first trial. The format was as follows: (a) instructions for a given test were displayed, (b) thirty seconds of subject-paced practice was given, (c) feedback in terms of percent correct was displayed, (d) three minutes of computer-timed testing was given, and (e) instructions for the next test were displayed. APTS Manikin and APTS Grammatical Reasoning tests utilize two minutes of testing time, due to rapid stabilization qualities. The APTS Nonpreferred Tapping test, being purely a motor task, was administered 10 seconds practice on the first trial and 20 seconds testing. Subjects were permitted to ask questions during the initial session for clarification of instructions.

A "smart system" was employed on all test administrations (Day 1 through Day 3). It alerts the subject to contact the experimenter when five consecutive responses are incorrect. A re-start of that particular test is then conducted by the experimenter. Computer orientation guidelines are detailed in Lane and Kennedy (1988).

Day 2: Subjects were administered the test battery without practice and feedback components. The tests were administered in the same order as the previous session for all subjects. Testing durations were identical to those
described in Day 1, with APTS Manikin and APTS Grammatical Reasoning being two minutes in length, and APTS Nonpreferred Tapping being 20 seconds. Subjects took 17 to 20 minutes to complete the battery.

Day 3: Subjects were administered the battery exactly as described in Day 2.

**Apparatus**

Testing was administered on Zenith ZFL-181-93 laptop computers. These low-cost, portable computers utilize liquid crystal displays and feature display controls which allowed subjects to adjust the contrast, brightness, and viewing angle of the display to personal preference. Each computer weighs 14.8 lbs. and has dimensions of 34.29 cm X 29.85 cm X 8.89 cm (Zenith Data Systems, 1987).
Table 1 shows the means, standard deviations, and test-retest reliabilities of each test using the number correct score among three trials. The averaged number correct score from Trials 2 and 3 served as baseline performance estimates for each test of the battery, with a mean reliability of .82 across tests. Table 2 lists the means, standard deviations, and ranges for assessment center variables.

A backwards multiple regression correlation was performed to determine what combination of tests, if any, was predictive of the dependent variable, OAR of assessment center performance. Utilizing the parameters of $F = 2.72$ and $p < .10$, the analysis reported that all test score variables were dropped from the multiple regression equation as non-significant in the prediction of OAR.

Bivariate correlational analyses of APTS/PAB test scores with dimension ratings, and with OAR, are presented in Table 3. There was no apparent pattern of correlations, and none approached significance.

The practice session (Trial 1) number correct scores were also run in a separate bivariate correlational analysis with dimension and OAR performance ratings. A
non-significant pattern of correlations was obtained in that analysis also.
### TABLE 1
MEANS, STANDARD DEVIATIONS (SD), AND TEST-RETEST RELIABILITIES FOR PREDICTOR VARIABLES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>T1:T2</th>
<th>T1:T3</th>
<th>T2:T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAB Math Processing</td>
<td>55.43</td>
<td>9.79</td>
<td>.77**</td>
<td>.52**</td>
<td>.79**</td>
</tr>
<tr>
<td>PAB Pattern Comparison (Simul)</td>
<td>28.52</td>
<td>3.41</td>
<td>.71**</td>
<td>.76**</td>
<td>.79**</td>
</tr>
<tr>
<td>APTS Grammatical Reasoning</td>
<td>28.59</td>
<td>8.93</td>
<td>.81*</td>
<td>.78**</td>
<td>.76**</td>
</tr>
<tr>
<td>APTS Manikin</td>
<td>55.19</td>
<td>8.02</td>
<td>.49**</td>
<td>.39</td>
<td>.88**</td>
</tr>
<tr>
<td>PAB Linguistic Processing</td>
<td>78.24</td>
<td>15.37</td>
<td>.57*</td>
<td>.44</td>
<td>.79**</td>
</tr>
<tr>
<td>PAB Code Substitution</td>
<td>69.07</td>
<td>6.34</td>
<td>.78**</td>
<td>.79**</td>
<td>.86**</td>
</tr>
<tr>
<td>Non-Preferred Tapping</td>
<td>35.37</td>
<td>10.68</td>
<td>.91*</td>
<td>.85**</td>
<td>.90**</td>
</tr>
</tbody>
</table>

*a Denotes averaged Trial 2 and 3 number correct scores.

* P ≤ .01, one-tailed.

** P ≤ .001, one-tailed.

### TABLE 2
MEANS, STANDARD DEVIATIONS (SD), AND RANGES FOR CRITERION VARIABLES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judgment</td>
<td>4.59</td>
<td>.69</td>
<td>3-6</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>4.78</td>
<td>.64</td>
<td>3-6</td>
</tr>
<tr>
<td>Leadership</td>
<td>4.70</td>
<td>.87</td>
<td>3-6</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>4.44</td>
<td>.70</td>
<td>3-6</td>
</tr>
<tr>
<td>Written Communication</td>
<td>4.89</td>
<td>.89</td>
<td>3-6</td>
</tr>
<tr>
<td>Oral Communication</td>
<td>5.11</td>
<td>.80</td>
<td>3-6</td>
</tr>
<tr>
<td>Perception</td>
<td>4.78</td>
<td>.51</td>
<td>4-6</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>5.22</td>
<td>.89</td>
<td>3-6</td>
</tr>
<tr>
<td>OAR</td>
<td>44.64</td>
<td>4.74</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 1
MEANS, STANDARD DEVIATIONS (SD), AND TEST-RETEST RELIABILITIES FOR PREDICTOR VARIABLES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>T1:T2</td>
</tr>
<tr>
<td>PAB Math Processing</td>
<td>55.43</td>
<td>9.79</td>
<td>.77**</td>
</tr>
<tr>
<td>PAB Pattern Comparison (Simul)</td>
<td>28.52</td>
<td>3.41</td>
<td>.71**</td>
</tr>
<tr>
<td>APTS Grammatical Reasoning</td>
<td>28.59</td>
<td>8.93</td>
<td>.81*</td>
</tr>
<tr>
<td>APTS Manikin</td>
<td>55.19</td>
<td>8.02</td>
<td>.49**</td>
</tr>
<tr>
<td>PAB Linguistic Processing</td>
<td>78.24</td>
<td>15.37</td>
<td>.57*</td>
</tr>
<tr>
<td>PAB Code Substitution</td>
<td>69.07</td>
<td>6.34</td>
<td>.78**</td>
</tr>
<tr>
<td>Non-Preferred Tapping</td>
<td>35.37</td>
<td>10.68</td>
<td>.91*</td>
</tr>
</tbody>
</table>

* Denotes averaged Trial 2 and 3 number correct scores.
* P ≤ .01, one-tailed.
** P ≤ .001, one-tailed.

### TABLE 2
MEANS, STANDARD DEVIATIONS (SD), AND RANGES FOR CRITERION VARIABLES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judgment</td>
<td>4.59</td>
<td>.69</td>
<td>3-6</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>4.78</td>
<td>.64</td>
<td>3-6</td>
</tr>
<tr>
<td>Leadership</td>
<td>4.70</td>
<td>.87</td>
<td>3-6</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>4.44</td>
<td>.70</td>
<td>3-6</td>
</tr>
<tr>
<td>Written Communication</td>
<td>4.89</td>
<td>.89</td>
<td>3-6</td>
</tr>
<tr>
<td>Oral Communication</td>
<td>5.11</td>
<td>.80</td>
<td>3-6</td>
</tr>
<tr>
<td>Perception</td>
<td>4.78</td>
<td>.51</td>
<td>4-6</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>5.22</td>
<td>.89</td>
<td>3-6</td>
</tr>
<tr>
<td>OAR</td>
<td>44.64</td>
<td>4.74</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 3
INTERCORRELATIONS BETWEEN ASSESSMENT CENTER RATINGS AND TEST AVERAGE NUMBER CORRECT SCORES

<table>
<thead>
<tr>
<th>Rating</th>
<th>Test</th>
<th>PMPNC</th>
<th>PPCNC</th>
<th>AGRNC</th>
<th>AMNC</th>
<th>PLPMC</th>
<th>PCSNC</th>
<th>NPTAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judgment</td>
<td>.13</td>
<td>-.04</td>
<td>-.02</td>
<td>-.14</td>
<td>-.23</td>
<td>.00</td>
<td>-.05</td>
<td></td>
</tr>
<tr>
<td>Decisiveness</td>
<td>.06</td>
<td>-.09</td>
<td>.04</td>
<td>.14</td>
<td>.08</td>
<td>.16</td>
<td>-.24</td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>.15</td>
<td>.07</td>
<td>-.06</td>
<td>-.02</td>
<td>-.28</td>
<td>-.03</td>
<td>-.20</td>
<td></td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>.26</td>
<td>.07</td>
<td>-.12</td>
<td>-.17</td>
<td>-.03</td>
<td>.16</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>Written Communication</td>
<td>.23</td>
<td>-.20</td>
<td>.02</td>
<td>-.18</td>
<td>.17</td>
<td>.03</td>
<td>-.18</td>
<td></td>
</tr>
<tr>
<td>Oral Communication</td>
<td>.07</td>
<td>.01</td>
<td>-.26</td>
<td>-.34</td>
<td>-.18</td>
<td>-.12</td>
<td>-.09</td>
<td></td>
</tr>
<tr>
<td>Perception</td>
<td>.15</td>
<td>-.24</td>
<td>.22</td>
<td>-.27</td>
<td>.09</td>
<td>-.09</td>
<td>-.18</td>
<td></td>
</tr>
<tr>
<td>Interpersonal</td>
<td>.11</td>
<td>-.22</td>
<td>.15</td>
<td>-.28</td>
<td>-.28</td>
<td>-.37</td>
<td>-.20</td>
<td></td>
</tr>
<tr>
<td>GAR</td>
<td>.16</td>
<td>-.10</td>
<td>-.02</td>
<td>-.22</td>
<td>-.14</td>
<td>-.05</td>
<td>-.19</td>
<td></td>
</tr>
</tbody>
</table>

*a* $P \leq .01$, one-tailed.

** $P \leq .001$, one-tailed.

AGRNC - APTS Grammatical Reasoning Number Correct

AMNC - APTS Manikin Number Correct

NPTAP - Non-Preferred Tapping

PCSNC - PAB Code Substitution Number Correct

PLPMC - PAB Linguistic Processing Number Correct

PMPNC - PAB Mathematical Processing Number Correct

PPCNC - PAB Pattern Comparison Simultaneous Number Correct
DISCUSSION

The absence of any identifiable relationships between assessment center performance and the tests used in this study both confirms aspects of earlier findings related to assessment center evaluations and fails to support the stated hypotheses.

In terms of the latter component, assessments of cognitive processing abilities are not related to assessment center performance ratings of dimensions or overall ratings. In no case was an individual test, nor a combination of tests' scores predictive of performance in an assessment center setting. Apparently the elements determining law enforcement assessment center evaluations are not of the cognitive processing domain as assessed by the computerized battery.

This finding is noteworthy in that it supports earlier research which reported evidence of common global factors which ultimately determine assessment center performance (Outcalt, 1988; Robertson et al., 1987; Sackett & Dreher, 1982; and Turnage & Muchinsky, 1982). Unequivocally, cognitive abilities are essential to successful job performance. However, successful assessment center
performance appears to be most heavily dependent on ratings of non-cognitive aspects of performance for this sample.

This speculation is corroborated by the Hirsh et al. (1986) conclusion regarding performance prediction for law enforcement occupations. Lower validity coefficients for predictors of law enforcement job performance, in comparison with other occupational types, may be due to the existence of a non-cognitive, interpersonal performance component inherent to police work alone. The assessment center may be evaluating this non-cognitive element associated with successful law enforcement job performance.

The lack of statistically significant overlap between the computerized test battery and assessment center evaluations should also be examined in a methodological context. The reliability of the test battery was confirmed, though the reliability of assessment center ratings, in terms of internal consistency, has been consistently low across several studies previously cited (see Thornton & Byham, 1982). Assessor evaluations, in turn, are subject to rating errors (e.g., halo, central tendency, etc.). All contribute to criterion unreliability.

Range restriction compounds the problem of criterion unreliability. Due to the "hurdle" system utilized to select assessment center candidates (technical knowledge test, tenure requirements, and a multiple choice in-basket
test), the final group of subjects were severely limited in terms of variability. Range restriction of sample subjects results in a demand for a greatly increased sample size in order to detect a significant relationship among variables (Schmidt, Hunter, & Urry, 1976). An additional source of range restriction is evident in terms of the scaling of assessment center ratings. Use of the full range in the 1 to 7 point scale in determining dimension and subsequent OAR scores was not evident. Ratings appear to exhibit the error of central tendency. Mid-range ratings abound, thus excluding the high and low extremes of the ratings parameters. Only the OAR rating exhibited a standard deviation greater than .89.

It is suspected that the unreliability of the criterion sharply reduced the power of the study to detect relationships of statistical significance, given limited sample size (N = 27). This speculation, combined with the likelihood that the content of the two respective measures is fundamentally different (cognitive versus non-cognitive), comprises a factor for consideration. This factor may be the chief explanation for the absence of an identifiable relationship between predictor and criterion.

The findings of this research indicate that computerized cognitive testing does not exhibit acceptable correlations with assessment center performance in the law
enforcement domain, which precludes its use as a surrogate measure in that context. The extent to which these results would generalize to assessment centers for other occupational types is unknown. It is recommended that the relationship between computerized cognitive testing and assessment center performance criteria be investigated within occupational types demanding high level information processing abilities.

The findings also suggest the need for an examination of the relationship between computerized testing and actual job performance estimates. Traditional measures, such as education or general ability tests, do not appear to be predictive of supervisory job performance (Turnage & Muchinsky, 1984). Computerized cognitive testing could be examined in various occupational settings as a non-traditional predictor of job performance. Areas considered for further research with this sample include: diagnostic applications (e.g., correlation with psychological or personality assessments); environmental applications (e.g., performance decrements under varying stressors encountered in law enforcement); and relatedness to other law enforcement tests (e.g., correlation with the screening test utilized in assessment center candidate selection).
ORANGE COUNTY SHERIFF'S OFFICE

DEFINITION OF SKILLS TO BE MEASURED

LEADERSHIP: THE ABILITY TO TAKE CHARGE; TO DIRECT COURSES OF ACTION; TO PROVIDE GUIDANCE TO SUBORDINATES IN MEETING GOALS AND OBJECTIVES; TO INITIATE ACTION; TO ENSURE COMPLIANCE WITH STANDARDS AND TO ENCOURAGE CONFIDENCE AND PRIDE IN WORK.

INTERPERSONAL: THE ABILITY TO ACT IN A SENSITIVE MANNER REGARDING THE NEEDS, FEELINGS AND CAPABILITIES OF OTHERS; TO ADVISE SUBORDINATES OF CHANGES; TO TACTFULLY DEAL WITH SENSITIVE ISSUES; TO CONSTRUCTIVELY CRITICIZE; TO ESTABLISH RAFFORT WITH OTHERS; TO LISTEN PRODUCTIVELY TO OTHERS.

ORGANIZING AND PLANNING: THE ABILITY TO ESTABLISH AND FOLLOW ORDERLY COURSES OF ACTION FOR SELF AND OTHERS; TO KEEP ORDERLY RECORDS; TO EFFECTIVELY PLAN WORK SCHEDULES; TO ESTABLISH OBJECTIVES AND PRIORITIES.

PERCEPTION: THE ABILITY TO IDENTIFY, UNDERSTAND AND INTEGRATE INFORMATION RELATED TO A SITUATION OR PROBLEMS; TO OBSERVE AND RECORD FACTS; TO EVALUATE INFORMATION OBJECTIVELY AND COMPLETELY; TO IDENTIFY PROBLEMS AND NEEDS.

JUDGMENT: THE ABILITY TO MAKE SOUND AND LOGICAL DECISIONS; TO APPLY PRINCIPLES TO SOLVE PRACTICAL PROBLEMS; TO DETERMINE WHEN TO CONTACT A SUPERIOR AND WHAT TO TELL HIM/HER; TO DRAW VALID CONCLUSIONS FROM AVAILABLE INFORMATION.

DECISIVENESS: THE ABILITY TO MAKE DECISIONS AND TAKE ACTION IN A TIMELY MANNER AND TO DEFEND DECISIONS WHEN CHALLENGED.

ORAL COMMUNICATION: THE ABILITY TO CLEARLY PRESENT AND EXPRESS INFORMATION ORALLY; TO UTILIZE EFFECTIVE ORAL SKILLS SUCH AS EYE CONTACT, GESTURES, VOICE INFLECTION, AND APPROPRIATE VOCABULARY IN COMMUNICATING WITH OTHERS.

WRITTEN COMMUNICATION: THE ABILITY TO CLEARLY PRESENT AND EXPRESS INFORMATION IN WRITING; TO UTILIZE EFFECTIVE WRITING SKILLS SUCH AS CORRECT GRAMMAR, PUNCTUATION, SPELLING, TRANSITION, SENTENCE AND PARAGRAPH STRUCTURE IN ORDER TO CLEARLY AND CONCISELY PRESENT WRITTEN INFORMATION.
REFERENCES


