# POST ENTRY-LEVEL LAW ENFORCEMENT

# TEST BATTERY

# USER'S MANUAL

## 1996

# POST ENTRY-LEVEL LAW ENFORCEMENT

# TEST BATTERY

## USER'S MANUAL

1996

Richard Honey
Bill Dyer

CALIFORNIA COMMISSION ON PEACE OFFICER STANDARDS AND TRAINING

Copyright California Commission on Peace Officer Standards and Training 1996

Published February, 1996

For information about copies of this publication, contact:

POST Media Distribution Desk
1601 Alhambra Boulevard
Sacramento, CA 95816
(916) 227-4856

## COMMISSIONERS

| | |
|---|---|
| Devallis Rutledge<br>Chairman | Deputy District Attorney<br>Orange County District<br>Attorney's Office |
| Manuel Ortega<br>Vice Chairman | Chief of Police<br>Placentia Police Department |
| Sherman Block | Sheriff<br>Los Angeles County<br>Sheriff's Department |
| Collene Campbell | Public Member |
| Jody Hall-Esser | Chief Administrative Officer<br>City of Culver City |
| George W. Kennedy | District Attorney<br>Santa Clara County |
| William B. Kolender | Sheriff<br>San Diego County<br>Sheriff's Department |
| Marcel L. Leduc | Sergeant<br>San Joaquin County<br>Sheriff's Department |
| Ronald Lowenberg | Chief of Police<br>Huntington Beach<br>Police Department |
| Daniel E. Lungren | Attorney General |
| Raquel Montenegro | Professor of Education (Retired) |
| Lou Silva | Officer<br>Oakland Police Department |
| Dale Stockton | Detective Sergeant<br>Carlsbad Police Department |
| Rick TerBorch | Chief of Police<br>Arroyo Grande Police Department |
| Norman C. Boehm | Executive Director |

# I N D E X

Page

INTRODUCTION.....

**SECTION ONE: DESCRIPTION OF THE READING AND WRITING TEST** .....

Rationale and Intended Use of the Test .....

Overview of the Test.....
Abilities Measured by the Test .....
Writing Ability.....
Reading Ability.....

Test Characteristics.....
Test Reliability.....
Limitations of the Test Battery .....

**SECTION TWO: OVERVIEW OF TEST DEVELOPMENT AND VALIDATION**.....

Test Development Procedures .....
Professional & Legal Standards Adhered to in Developing the Test Battery.....
Writing Ability: Survey of Writing Mechanics .....
Reading Ability: Source of Information Questionnaire .....
Development of Test Items Validity of the Test Battery .....
Relationship of Test with Performance in the Academy and on the Job - POST Validation Studies .....
Differential Prediction .....
Validity Generalization .....

**SECTION THREE: ADMINISTRATION, SCORING AND INTERPRETATION**.....

Test Administration Overview
Administration and Scoring of the Test Battery .....

Scoring and Interpretation.....
Scoring Procedure.....
Test Score Reporting.....
Use of Test Results .....
Adverse Effect.....
Use of Test Battery to Rank Candidates .....

Page

Test Norms .....
Applicant Test Norms .....
Cadet Test Norms .....

Utility of the Test .....
Test Utility .....
Expectancy Tables .....
Utility in Predicting Academy and On-the-Job
Performance .....

## INTRODUCTION

This manual describes the POST Entry-Level Law Enforcement Test Battery and provides information regarding its development, validation, and appropriate use. It is divided into three basic sections: the first describes the test, the second the development and validation of the test, and the third contains information regarding actual test use. If, after reading the manual you have any questions, or if you would like clarification on some aspect of the testing process, please contact the Standards and Evaluations Services Bureau at the California Commission on Peace Officer Standards and Training (POST) in Sacramento.

## SECTION ONE: DESCRIPTION OF THE READING AND WRITING TEST

### Rationale and Intended Use of the Test

For law enforcement officers, the ability to read and the ability to write are essential for both successful completion of academy training and successful performance on the job. On these points there is no disagreement. Every major study of the entry-level law enforcement job in California has reaffirmed this conclusion. Yet, in spite of the undeniable importance of reading and writing ability for success in the law enforcement profession, many academy students continue to read and write poorly as demonstrated by their reading scores on academy entrance exams and by the reports that they write in the academy.

In response to this condition, POST undertook the development of a battery of tests designed specifically to assess the reading and writing skills of entry-level law enforcement applicants (Honey and Kohls, 1981). The POST Entry-Level Law Enforcement Test Battery is the end result of this effort. The research on this test battery has been extensive. Over five years were invested in the original development of the exam, and since then, research and evaluation have been ongoing.

### Overview of the Test

The POST Entry Level Peace Officer Reading and Writing Test consists of two booklets. Book A contains four segments: three 15 item, multiple-choice writing segments, and one 20 to 30 item, multiple-choice reading segment. The three writing segments measure clarity, vocabulary and spelling skills. Book B contains a 350 word cloze reading test with 40 deleted words.

To score the test battery, the three writing tests are combined into one standardized score and the two reading tests are combined into one standardized score. These two combined scores are then restandardized to create one total standardized test score. The exam requires a maximum of two and one half hours test time plus one half hour for instructions.

**Writing Ability**

The clarity sub-test requires the test-taker to identify which of two sentences is correctly written. As the segment name implies, this section is intended to measure the test-taker's ability to identify clear writing. Typical, incorrect sentences manifest problems with unclear reference, misplaced modifiers, fragments, run-ons, etc. No uncommon sentence faults are included.

Example Clarity Item:

1. a. Bill saw him come out of the bank.
   b. Bill saw John come out of the bank.

In this example, "b" is the correct choice. Alternative "a" is incorrect because the pronoun "him" is unreferenced.

The vocabulary test uses a standard four alternative, multiple-choice format. One word in a sentence is underlined and the test-taker is instructed to select the alternative that is the most accurate synonym or definition. Words are included in the spelling and vocabulary tests that are likely to be encountered in law enforcement work, but which are not law enforcement specific or learned in training or on the job.

Example Vocabulary Item:

1. The officers secured the <u>perimeter</u> of the crime scene.

   a. area
   b. contents
   c. boundary
   d. integrity

Choice "c", boundary, is the appropriate synonym.

The spelling test also utilizes a four alternative, multiple-choice format. A sentence is given with one word omitted. A line indicates the location of the omitted word. Four different spellings of the same word are given. The alternative spellings are arrayed under the item.

Example Spelling Item:

1. He could not _____ fact from fiction.

   a. separate
   b. seperate
   c. separat
   d. seperat

Choice "a", separate, is the correct spelling.

**Reading Ability**

Two types of reading comprehension tests are utilized. One is comprised of a 20 item read-a-passage answer-a-question multiple-choice test (two early forms of the test have 30 item multiple-choice reading tests), and the other is a 40 item cloze test. The multiple-choice test utilizes passages taken from a variety of sources and varying in length from a single paragraph to a page. Some come directly from law enforcement materials; others are non-law enforcement narratives. Neither test requires any prior law enforcement knowledge in order to correctly answer the questions or complete the blanks.

The multiple-choice reading test measures the test-taker's ability to either distill factual information from a passage or to draw logical conclusions from the textual material. The multiple-choice format is very effective for measuring specific reading sub-skills (Farr, 1969).

Example Multiple Choice Reading Item:

<u>THE FOLLOWING PASSAGE SHOULD BE USED TO ANSWER QUESTION 1</u>

In all societies there are problems - problems of poverty, security, and literacy to name but a few. In response to these problems, societies have developed numerous means with which to deal with them. One way that societies address certain types of problems is by the division of labor. In this division of labor, certain functions and people are extended a license to perform certain activities that others may not. The license is legislatively mandated, and the work provided by the licensed group may be performed by no others.

1. Which of the following can be concluded from the above passage?

    a. The issuance of licenses is one of societies' ways of solving basic problems.

    b. Every division of labor has a corresponding license.

    c. The problems of poverty, security, and literacy are greater in some societies than others.

    d. Licenses have tended to evolve in response to various intrinsic societal problems.

In this example, choice "a" represents the most appropriate conclusion that can be drawn from the preceding passage.

The cloze test is comprised of a passage with 40 words deleted. The first and last sentences of the passage are left intact. Between the first and last sentences, every seventh word is systematically deleted from the text. No word is deleted that cannot be deduced from context. Two deletion strategies are utilized in the test battery. In the first two forms of the test, a fixed length, solid line indicates where a word has been deleted. The test-taker can select any word that is consistent with the meaning of the passage. The first letter of the chosen word is then coded onto the answer sheet. In later forms of the test, a dashed line is presented where a word

has been deleted, with the number of dashes indicating the number of letters in the deleted word. In these forms, in addition to making sense in the passage, the word that the test-taker selects must also have the same number of letters as there are dashes. The fixed length line format is being phased out in favor of the dashed line format. The latter format, by more rigorously limiting the number of correct responses, makes scoring key development easier. In both formats, for a word to be scored as correct, it must be the correct part of speech, and it must fit appropriately with the meaning of the overall passage. In some instances, there is more than one correct response, in others only one word can correctly fill the blank.

The cloze test is printed in a separate, disposable test booklet so that test-takers can actually see the words that they are putting in the blanks thus, providing a more complete "picture" of the textual material being interpreted. Test-takers are instructed to actually write the word in the test booklet and then code the first letter of every word they chose onto a scannable answer sheet.

While the multiple-choice format can measure various sub-skills that comprise reading behavior (Farr, 1969), the cloze format provides a measure of the aggregate of all interacting factors which affect the understanding of a passage by a reader (Taylor, 1953).

## Test Characteristics

Ten forms of the test have been developed, all are presently in use. The test statistics for each of the forms vary slightly as can be seen in the Table 1.

Table 1
Descriptive Statistics for Reading & Writing Tests

| Test Form | Total R/W | Read STD | Write STD | Clarity | Spell | Vocab | M/C Read | Cloze |
|---|---|---|---|---|---|---|---|---|
| Form 220 (N=26,633) Mean | 47.14 | 50.47 | 44.34 | 70.08 | 73.09 | 69.52 | 72.17 | 63.33 |
| SD | 12.23 | 12.48 | 11.41 | 13.98 | 15.99 | 15.51 | 18.30 | 14.02 |
| Form 281 (N=7,045) Mean | 48.25 | 50.80 | 45.90 | 73.71 | 74.24 | 69.48 | 76.93 | 59.34 |
| SD | 11.96 | 11.74 | 11.42 | 13.39 | 15.16 | 16.68 | 15.06 | 15.66 |
| Form 291 (N=558) Mean | 50.65 | 55.97 | 44.99 | 69.20 | 79.42 | 66.06 | 74.88 | 73.47 |
| SD | 12.72 | 14.01 | 11.00 | 14.68 | 14.75 | 15.28 | 16.69 | 19.10 |
| | | | | | | | | |
| Form 400 (N=23,761) Mean | 48.50 | 50.14 | 47.08 | 70.81 | 73.75 | 76.43 | 69.39 | 65.11 |
| SD | 12.35 | 12.31 | 11.81 | 15.23 | 17.03 | 15.24 | 17.67 | 14.17 |
| Form 410 (N=32,846) Mean | 45.80 | 48.35 | 44.07 | 69.90 | 73.44 | 68.55 | 67.65 | 62.88 |
| SD | 12.43 | 12.63 | 11.90 | 14.09 | 16.51 | 16.00 | 17.79 | 14.64 |
| Form 431 (N=24,517) Mean | 48.60 | 50.47 | 46.94 | 72.92 | 77.88 | 69.78 | 70.72 | 64.77 |
| SD | 13.22 | 12.90 | 12.51 | 14.91 | 16.64 | 18.26 | 16.24 | 17.09 |
| Form 441 (N=2,700) Mean | 48.64 | 50.61 | 46.87 | 67.08 | 79.41 | 73.86 | 69.99 | 65.83 |
| SD | 12.91 | 11.67 | 13.18 | 15.77 | 17.81 | 19.31 | 14.72 | 15.82 |
| Form 451 (N=11,542) Mean | 45.34 | 48.80 | 42.78 | 69.07 | 72.08 | 66.81 | 68.68 | 63.17 |
| SD | 11.88 | 11.41 | 11.84 | 15.01 | 16.75 | 17.59 | 15.47 | 14.55 |
| Form 461 (N=1,736) Mean | 47.35 | 53.51 | 41.67 | 67.18 | 69.21 | 68.42 | 74.94 | 67.68 |
| SD | 11.33 | 11.32 | 11.18 | 13.80 | 17.72 | 14.27 | 14.94 | 14.38 |
| Form 471 (N=814) Mean | 49.78 | 54.07 | 45.44 | 72.03 | 70.94 | 73.07 | 73.30 | 70.62 |
| SD | 11.98 | 11.97 | 11.31 | 13.09 | 16.90 | 14.93 | 16.63 | 14.79 |

Note: Approximately 94% were job applicants and 6% were academy students or prospective students. Total Read/Write, Reading STD and Writing STD scores are T-scores calibrated to 1983 validation study sample; subtest scores are raw percent correct scores.

## Test Reliability

The average reliability of the test battery was estimated by the test-retest procedure to be .86. Test battery users will find that test reliabilities will vary from the .84 to .87 for the different test forms. The standard error of measurement (SEM) for the current forms of the test vary from 4.39 to 5.02. Table 2 contains the test reliabilities and standard errors of measurements for those forms with sufficient sample size to calculate test-retest correlations.

Table 2: Test-Retest Reliabilities

| Test Form | 220 | 400 | 410 | 431 | 441 | 451 |
|---|---|---|---|---|---|---|
| Reliability | 0.871 | 0.869 | 0.843 | 0.858 | 0.849 | 0.858 |
| SEM | 4.39 | 4.47 | 4.93 | 4.98 | 5.02 | 4.48 |

## Limitations of the Test Battery

Two test formats are utilized in the Test Battery to assess reading ability--multiple-choice and cloze. Scores generated by these different formats tend to be moderately correlated. Entin and Klare (1978) found an average correlation between the two formats to be .68, and Weiner (1994) found the specific correlation between these two tests in the POST Test Battery to be .63. The magnitude of these correlations indicates that there is both common and unique variance between scores from the two formats. When the above correlations are squared to determine the specific amount of common variance between the two scores, values of 46% and 40% are generated. Thus, slightly less than half of the variance between the two scores is common leaving more than half that cannot be explained by the other test. Since each is explaining slightly different aspects of the reading process, the inclusion of both in the test battery tends to make more robust the ability of the Test Battery to accurately assess reading skills. By including both formats in the Test Battery (and due to the achieved reliabilities and validities of the tests) we are confident that the Test Battery offers a good assessment of reading. Achieving such a robust measure of prediction for writing ability, however, is more problematical.

There are two approaches for assessing writing--direct assessment where the test-taker produces an example of his or her own writing, and indirect assessment where generative writing skills are estimated or inferred from the test-taker's ability to answer questions regarding one or more aspects of writing such as grammar, punctuation, spelling, etc.. Direct assessments are made with essay tests and indirect assessments with multiple-choice tests. Each procedure has its own strengths and weaknesses, its own advantages and disadvantages.

Essays are the best predictor of the "higher level" writing skills such as organization, linear development, sense of purpose, clarity etc. (Cooper, 1984; Breland, 1987). On the down side, essay scores are affected by numerous factors that detract from reliability. The topic selected to

write on affects the quality and score of the essay (Braddock et al., 1963). The mode of writing required by the exam (narration, exposition, etc.) also affects the score (White, 1982). The appearance of an essay (neatness and penmanship) can affect how it is evaluated (Markham, 1976). In addition to these writer and assignment variables, the rater variable also produces a degree of error in scores. In spite of these problems and the costs of scoring, support for essays continues and is, in fact, increasing (Cooper, 1984).

Multiple-choice tests are superior to the essay for measuring the word and sentence level characteristics of writing (i.e. spelling, punctuation, grammar, and word choice), but are suboptimal measures of generative writing skills. Multiple choice tests are subject to much less extraneous variability than are essays and multiple choice tests are significantly less expensive to score.

The debate over which procedure to utilize, or whether to use the two procedures together, has been going on since the turn of the century. With the publication of a study by Godshalk in 1966, which demonstrated that short essay tests could be reliably scored and make a unique contribution to the prediction of a stable measure of writing ability, the essay test has increasingly gained adherents, particularly in the area of educational testing. In the educational realm approximately 95% of all writing assessments contain an essay test of some form (Cooper, 1984). In personnel testing, essays are far less common. While POST has no official figures on the percentage of POST Test Battery users who use a formal essay test the number is probably below 5%.

It should be noted that when the Test Battery was initially developed in 1982, POST was aware of the potential limitations of both the multiple-choice test and the essay test. At that time, POST developed a writing sample test that was scored using the then generally accepted analytic scoring process. Analytical scoring focuses on the correctness of the lexical (word level) and syntactic (sentence level) characteristics of the writing. In addition to being both a time consuming and costly process, the scores generated by the analytical scoring system were found to be unreliable. These deficiencies of the analytic scoring procedure led to the abandonment by POST of the writing sample test.

In 1988, POST again addressed the issue of writing sample testing. In this attempt to improve the prediction of writing skills, however, POST selected a holistic scoring model. Holistic scoring is a procedure for ranking writing samples relative to a set of exemplar, or bench mark, reports. Readers review a number of reports and select those that they feel best represent the various points on the rating scale. A 6-point scale has become the norm with a "1" report being totally unacceptable and a "6" report being excellent. Unlike analytical scoring that focuses on the word and sentence level of the writing, holistic scoring places primary emphasis on the discourse, or passage level of the writing. Word and sentence level considerations are not totally disregarded; they are, however, evaluated in terms of their affect on meaning, not simply on their individual correctness or incorrectness. Thus, holistic scoring is mainly concerned with such things as the organization and development of a passage, the quality of the ideas presented, the effectiveness of transitions, the use of supporting information etc.

The resulting procedure has proven to be a reliable, valid predictor of generative writing ability

(White, 1982, Weiner, 1994). While holistic scoring is much more cost effective and less time consuming than with previous analytical methods, POST lacks the resources to score large numbers of essays, (POST presently provides it's multiple-choice test battery at no cost to user agencies. This service includes the scoring of the tests.) Consequently, POST does not offer the writing sample test as a component of its entry-level battery. It should be pointed out that locally scored writing sample tests may compromise the quality of the essay assessment.

This discussion of the strengths and weaknesses of the various test formats for assessing reading and writing is presented to give the Test Battery user a more clear understanding of the issues that must be addressed in deciding on a testing strategy for the assessment of language skills. What has been presented here, however, is only a brief description of some of the more important considerations. An in-depth discussion of the issues related to the assessment of reading and writing skills is beyond the scope of this manual. For those wishing additional information on the topic, POST staff will gladly provide additional information regarding test formats, scoring procedures, and other related issues.

## SECTION TWO: OVERVIEW OF THE TEST DEVELOPMENT AND VALIDATION

### Test Development Procedures

#### Professional Standards Adhered To in Developing the Test Battery

The test battery was developed in accordance with recognized professional and legal standards for establishing job-relatedness (Honey and Kohls, 1981). While consideration was given to the content of the job when developing the test, the primary model of validation utilized to establish the job relatedness of the test battery was and is empirical. This approach to establishing validity involves a statistical evaluation of the relationship between scores on a test and various criterion measures of performance. Regarding criterion-related validity Section 14 of the Uniform Guidelines on Employee Selection (1978) states:

> "Generally, a selection procedure is considered related to the criterion, for the purposes of these guidelines, when the relationship between performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance, which means that it is sufficiently high as to have a probability of no more than one (1) in twenty (20) to have occurred by chance."

Empirical validation results are reported in Table 4 and Table 5 of this manual. The test battery has been empirically validated against both training (academic performance in the academy) and job performance (performance in field training).

### Job Analysis

In 1979, POST conducted a statewide job analysis of the entry-level law enforcement position (Kohls, Berner and Luke). The study, which remains the most extensive analysis of the entry-level position undertaken in the State, involved the collection of frequency and importance ratings for over 500 job activities. It has also served as the model for conducting similar analyses in numerous other states. Data were collected from officers (N = 1720) and supervisors (N = 675) in each of 217 police and sheriffs departments.

Not surprisingly, tasks which require the entry-level officer to employ language skills were found to be frequently performed and important to successful performance of the entry-level job. Regarding writing, the analysis that revealed that, on average, approximately 25% of the entry-level patrol officer's time is spent writing or dictating reports.

The 1979 POST job analysis served as the foundation for the research that produced the POST Test Battery. The following additional information regarding the language requirements of the job was collected preliminary to developing and validating the actual tests.

#### Writing Ability: Survey of Writing Mechanics

A Survey of Writing Mechanics was developed and administered as a means of identifying the most significant writing requirements of the entry-level job. Respondents to the survey were

presented with examples of different technical writing errors (e.g., misplaced modifiers, run-ons, ambiguous reference, etc.), and asked to indicate whether the errors were: (1) acceptable in a report, (2) not acceptable but persons making the errors could be remediated, or (3) not acceptable and persons making the error should not be hired.

Officers and first-line supervisors from 7 departments completed the survey (Los Angeles PD, Los Angeles SO, Santa Rosa PD, Sonoma SO, Sacramento PD, Sacramento SO, San Francisco PD). Results of the survey showed that three aspects of writing were prerequisite to satisfactory job performance: clarity, vocabulary, and spelling. The departments sampled indicated that all other errors of punctuation and grammar were either acceptable or could be corrected by a reviewing officer.

**Reading Ability: Source of Information Questionnaire**

A Source of Information Questionnaire was developed to further specify the reading requirements of the job. Respondents to the questionnaire were asked to identify the exact form in which they received information on the job: verbally, a written summary, or in original written form. Responses to the questionnaire were analyzed to identify those materials important to, and actually read by, officers. Completed questionnaires were received from 508 officers from 41 departments.

In conjunction with the return of the completed questionnaires, each department was asked to gather and submit samples of written materials from each of 15 reading categories.[1]

The readability levels of these materials were determined to identify the approximate level of reading proficiency required to adequately perform job duties. Since different readability equations generate different difficulty estimates for the same text, three indices, the Flesch, Smog, and Fog, were utilized in this analysis.[2]

Because the readability information was intended only to provide a rough estimate of the reading difficulty encountered by officers, no attempt was made to either combine the various estimates into a single index, or to identify which of the three was the "best" single estimation.

To compute a readability estimate requires narrative passages of at least 100 words that are comprised of complete sentences. Review of the submitted reading materials revealed that only those from 6 categories met this criterion. Thus, although materials from all 15 categories were determined to be important to job success (based on the original job analysis), reading samples

---

[1] Fifteen categories of reading materials were identified as being important for job success: in-depth narrative, short reports, check box reports, maps correspondence, memos, vehicle bulletins, manuals, case law, legal interpretations, teletype messages, training bulletins, warrants, coded material, and legal codes.

[2] All readability indices rely on essentially two variables in predicting difficulty, one semantic and one syntactic. Semantic difficulty is usually estimated by the number of polysyllables in the passage or by the number of words the passage that are on a list of common or uncommon words. Syntactic difficulty is usually estimated by sentence length. How these factors are weighted in the readability equation accounts for the majority of the fluctuations between estimates.

from only 6 categories were subjected to readability analysis. The results of the readability analysis are summarized in Table 3.

Table 3: Reading Difficulty of Law Enforcement Material by Reading Category

| Category | Number of passages analyzed | Flesch index mean | Smog index mean | Fog index mean |
|---|---|---|---|---|
| Codes | 27 | 17.7 | 17.7 | 26.6 |
| Manuals | 46 | 14.6 | 14.7 | 18.7 |
| Training | 47 | 13.1 | 13.8 | 16.0 |
| Legal | 72 | 12.9 | 14.4 | 17.6 |
| Memos | 49 | 10.8 | 12.0 | 13.7 |
| Reports | 56 | 8.2 | 9.6 | 11.0 |
| Overall | 297 | 12.4 | 13.4 | 16.4 |

**Development of Test Items**

All test items were written to conform with the findings of the supplemental job analyses. Separate tests were written for each of the three components of writing found to be most critical to the job (clarity, vocabulary, spelling), and care was taken to ensure that the specific content of each item was appropriate to the entry-level job.

Once the specifications for item content were identified, items were developed from scratch. Clarity items focused on ambiguity (unclear reference, misplaced modifiers), incomplete expressions, and punctuation (incomplete sentences and run-on sentences). The vocabulary and spelling items were gathered from a variety of sources. The approximate difficulty of the words to be used was identified in "The Living Word Vocabulary" by Dale and O'Rourke which attaches grade level estimates to individual words. Reading passages for the multiple choice reading test were constructed to reflect the range of reading difficulties identified in the readability analysis. The cloze passage was written at the mean reading difficulty.

The two reading tests are grounded in different theoretical notions of what constitutes reading behavior. The multiple-choice test is rooted in the behaviorist tradition and is based on analytical, reductionist assumptions (Farr, 1969). From this theoretical perspective the whole is defined as being the sum of the parts. When applied to the assessment of reading, the behaviorist model focuses on the direct measurement of the discrete sub-skills that theorists have identified as the constituent aspects of reading behavior. In the POST Test Battery the multiple-choice items are designed primarily to assess the ability to distill factual information from a passage, to make logical conclusions based on the passage, and to follow instructions.

Agencies that participated in the supplemental job analysis were asked to provide POST with

examples of the various types of written materials that officers encounter on the job. This included general orders, memos, manuals, etc. Individuals also identified, when they completed the Source of Information Questionnaire, which code sections (from the Penal, Vehicle, Health and Safety, Welfare and Institutions, and Municipal Codes) they most frequently read on the job. From these materials, and from test items developed in a consortia setting (Krueger, 1978), POST staff generated read-a-passage, answer-a-question format multiple-choice items.

Items were selected that both reflected the different types of reading found on the job and represented the range of reading difficulty identified in the job analysis (8th to 16th grade level using the SMOG index). Items were analyzed using classical item analysis. Items were selected which maximized reliability while maintaining a sample of items that reflected the diversity of reading tasks found on the job. Items in the .5 -.7 difficulty were considered ideal, but easier and more difficult items were retained in order to retain a variety of reading tasks found on the job.

The cloze procedure, named after the gestalt concept of closure, is based in the cognitive tradition and is characterized by holistic assumptions. It is according to Taylor (1953), the developer of the procedure, a holistic measure of reading behavior. Cloze is a measure of the aggregate influences of all interacting factors which affect the understanding of a passage by a reader (Ruddell, 1963, Goodman, 1969). It is a powerful measure of fundamental language skills. The cloze passages from which words are deleted were written or modified by staff. Passages were law enforcement related but required no prior knowledge of law enforcement to complete.

Starting with the first word in the second sentence (the first sentence is always left intact to get the test-taker started) every seventh word was deleted. This created deletion pattern number one. Then, starting with the second word in the second sentence every seventh word was deleted. This created deletion pattern number two. This procedure was followed until all seven deletion patterns were identified. No word or number that could not be deduced from context was deleted. Once all deletion patterns were identified, the part of speech of every word in the passage was identified. With this information, the representation of each part of speech in the passage and in every deletion pattern was determined. The three deletion patterns that most closely approximated the entire passage, in terms of the representation of the various parts of speech, were retained. These three were then trial administered to identify the pattern with the most variance. That pattern was retained in the test battery.

Both test formats contribute significant, unique variance to the prediction of both academy performance and job success, therefore, both are retained in the test battery.

## Validity Of the Test Battery

Empirical evidence of the job-relatedness of the test battery was examined by correlating test scores with various criteria. Since its introduction in 1982, the validity of the test battery has been evaluated on four occasions.[3] In each of these studies the test battery has met or exceeded the requirements for empirical validity as articulated in the Standards for Educational and Psychological Testing (1985) and the Uniform Guidelines on Employee Selection Procedures (1978). These studies are summarized below. The validity data presented in Tables 4,5, and 6 was generated the 1994 study of the test battery.

### Relationship of Test with Performance in the Academy and On the Job - POST Validation Studies

**1981 Study** - In the original Entry-Level Test Battery validation research conducted by POST (Honey & Kohls, 1981), reading and writing tests and an essay test were administered to 320 basic academy students at five academies near the beginning of training, and measures of their performance in the academy were obtained later, near the end of training. Two criterion were used in this study: the POST Basic Academy Proficiency Test which is a state developed, final exam that all academies are required to administer at the end of training, and a standardized linear composite of all locally developed analytical tests administered during basic training. The first criteria measure is common to all cadets in the study, and the second is unique to each academy. It should be noted that the test battery did not include a traditional multiple-choice reading comprehension test at this time. Also, the essay test in this study was scored using an analytical procedure rather than the currently used holistic approach.

**1983 Study** - A second academy study was conducted in 1983 which served as the basis for the development of a test user's manual. By this time, a multiple-choice reading comprehension test had been developed and added to the battery. The study entailed administering the Entry-Level Test Battery to 480 academy students at 10 academies and then gathering measures of cadet performance in the academy at the end of instruction. The same Proficiency Test scores and linear composite of local scores as were used in the 1981 research were used in this research. Test scores and academy performance measures were then correlated.

**1987 Study** - A predictive criterion-related validation study was conducted in 1986-1987 as a follow-up evaluation of the operational Entry-Level Test Battery (Weiner & Berner, 1987). Test Battery scores were retrieved from POST's operational test program files for 1,270 examinees, and measures of their subsequent performance in basic training and on-the-job were collected. No essay test was included in this study. Academy performance was measured by scores on the POST Basic Academy Proficiency Test. It should be noted that the Proficiency Test was found, in a 1983 study by Norborg, to be predictive of the students' later job performance as measured by specially developed and administered job simulations, performance ratings, and indices of completion of field training and probation (Norborg, 1983). Performance on the job was

[3] Validity data were presented in: 1) Entry-Level Reading and Writing Tests for California Law Enforcement: Final Report, 1981, 2) POST Commission Agenda Item, 1984, 3) POST Basic Course Proficiency Test Study, 1987, 4) POST Commission Agenda Item, 1990, 5) Longitudinal Study of the POST Entry-Level Test Battery, 1992.

evaluated using various measures that included: (a) field training officer (FTO) ratings of officer/trainees' performance of job duties that involve writing--Behaviorally Anchored Rating Scales (BARS) were used to make the evaluations; (b) overall success or failure in completing field training--a dichotomous pass/fail variable; (c) patrol supervisor ratings of tenured officers' performance of job duties that involve writing--again using BARS; and (d) overall success/failure in completing probation.

Success criterion data were collected by POST from 27 basic academies between 1986 and 1987. These academies indicated whether each student graduated, withdrew for academic reasons, withdrew for other reasons, failed for other reasons, or recycled to the next academy. Controlling for these outcomes, prior scores were then correlated with the various criterion measures.

**1994 Study** - This study was a re-analysis of all research data collected since 1980. Writing test and essay test scores were examined with respect to their predictions of performance in basic training (academy instructor ratings of writing ability, overall success/failure in completing basic training, and POST Proficiency Test scores) and in field training (FTO ratings of writing ability and overall success/failure in completing field training). Additionally, fairness analyses were computed using all available data, new expectancy and utility tables were computed, and alternative combinations of subtests were analyzed.

Uncorrected correlation coefficients are reported in the table which follows. These correlation coefficients are probably low estimates since they do not take into account restriction in the range of scores on the test and unreliability in criterion measures.

Table 4 summarizes the relationship between scores on the test battery and performance in the academy as represented by the POST Proficiency Test, Academy-Specific Tests, Instructor Ratings, and Academy Success/Failure; as well as the relationship of test scores with performance on the job as represented by Field Training Officer (FTO) Ratings, Patrol Supervisor Ratings, Field Training Success/Failure, and Probation Success/Failure. The significant correlations for the total Read/Write test battery range from r=.10 to r=.60 and most are significant beyond the p<.001 level.

Table 4
Summary of Validity Evidence for
POST Reading & Writing Tests

| Criterion Measure | R/W Total | Read STD | Write STD | Clarity | Spell | Vocab | M/C Read | Cloze | Essay |
|---|---|---|---|---|---|---|---|---|---|
| **Academy Performance** | | | | | | | | | |
| POST Proficiency Test | .52** | N/A | .40** | .28** | .21*** | .41*** | N/A | .50** | .18* |
| 1981 study[4] | * | .66** | * | * | .09* | .46*** | .61** | * | N/A |
| 1983 study[5] | .60** | * | .41** | .38** | .20*** | .39*** | * | .54** | N/A |
| 1987 study[6] | * | .56** | * | * | .14*** | .36*** | .52** | * | .28*** |

[4]Read/Write Test N=218-320; Essay N=147-149. Reading test includes two cloze tests (no multiple-choice test). Essay test was scored using an analytical method.

[5]N=480.

[6]N=1270.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Current study[7] | .56***<br>.47*** | *<br>.49*** | .42***<br>.34*** | .35***<br>.25*** | | | *<br>.45*** | .44***<br>.40*** | |
| Academy-Specific Test<br>1981 study[a]<br>1983 study[b] | .53***<br>.54*** | N/A<br>.55 | .45***<br>.40*** | .33***<br>.33*** | .25***<br>.19*** | .41***<br>.36*** | N/A<br>.51*** | .47***<br>.44*** | .25***<br>N/A |
| Instructor Ratings<br>Current study[8] | .39*** | .36*** | .35*** | .24*** | .28*** | .26*** | .33*** | .31*** | .31*** |
| Academy success/failure<br>1987 study[9]<br>Current study[10] | .21***<br>.14** | .24***<br>.12** | .14***<br>.12** | .09***<br>(.04) | .09***<br>.13** | .13***<br>.10* | .20***<br>.09* | .21***<br>.13** | N/A<br>.21** |
| **Job Performance** | | | | | | | | | |
| FTO ratings<br>1987 study[11]<br>Current study[12] | .38***<br>.26*** | .27***<br>.21*** | .40***<br>.24*** | .24***<br>.13** | .36***<br>.26*** | .24***<br>.14** | .27***<br>.15** | .18*<br>.23*** | N/A<br>(.09) |
| Patrol Supervisor ratings<br>1987 study[13] | .22*** | .19*** | .20*** | .18*** | .15** | .12** | .17*** | .14** | N/A |
| Field training success/failure<br>1987 study[14]<br>Current study[15] | .10***<br>(.05) | .12***<br>(.06) | .07*<br>(.04) | .10***<br>(.08) | (.03)<br>(-.01) | (.02)<br>(.02) | .09**<br>(.07) | .11***<br>(.03) | N/A<br>(.05) |
| Probation success/failure<br>1987 study[16] | .13*** | .13*** | .10** | .09** | .07* | (.05) | .10*** | .13*** | N/A |

***p<.001; **p<.01; *p<.05 (one-tailed).

[7]Read/Write Test N=13,347; includes all available Read/Write scores matched to Proficiency Test scores obtained between Aug83-Feb92. Essay Test N=227; data collected for current study.

[8]Read/Write Test N=504; Essay N=413. Criterion is FTO mean rating on 4 writing abilities. Essay scored using holistic procedure.

[9]Supplemental study; N=1271. Pass/fail index: Graduated=1; failed/withdrew for academic reasons=0. Correlations are point-biserials.

[10]Read/Write Test N=423; Essay N=295. Pass/Fail index: Completed (C1,2)=1; resigned or terminated due to inadequate writing skills or other KSAs (R2,7 & T2)=0. Correlations are point-biserials.

[11]N=103. Criterion is FTO rating of officer performance of job duties that involve writing.

[12]Read/Write Test N=329; Essay N=292. Criterion is FTO mean rating on 4 writing abilities.

[13]N=382. Criterion is patrol supervisor rating of tenured officer performance of job duties that involve writing.

[14]N=1062. Pass/fail index: Completed (C1-3)=1; Failed to complete (R2,3 or T2,3 or F2,3)=0. Correlations are point-biserials.

[15]Read/Write Test N=403; Essay N=336. Pass/fail index: Completed (C1-3)=1; Resigned or terminated due to inadequate report writing skills, analytical skills, or other KSAs (R3,4,5,7 or T3,7)=0. Correlations are point-biserials.

[16]N=895. Pass/fail index: Completed (C1-3)=1; Failed to complete (R2,3 or T2,3 or F2,3)=0. Correlations are point-biserials.

The study of within-group validity for the POST Reading/Writing test reveals that the test is a valid predictor of the criteria in almost all cases for each of the protected groups. In some cases the number of subjects in the protected group was too small for reliable results to be determined. Table 5 displays the results of the validation studies by protected group.

Table 5
Summary of Within-Group Validity for the
POST Reading & Writing Test Battery

| Criterion Measure | Am. Indian | Asian | Black | Filipino | Hispanic | White | Male | Female |
|---|---|---|---|---|---|---|---|---|
| **Academy Performance** | | | | | | | | |
| POST Proficiency Test | | | | | | | | |
| 1981 study[17] | . | .73** | .62** | . | .49** | .46** | .53** | .62*** |
| 1983 study[18] | . | .68** | .55** | . | .66*** | .57** | .63** | .61*** |
| 1987 study[19] | . | . | * | . | .49*** | * | * | .66*** |
| Current study[20] | .57*** | .52*** | .53** | .43*** | .40*** | .54** | .55** | .49*** |
| | | | * | | | * | * | |
| | | | .40** | | | .45** | .48** | |
| | | | * | | | * | * | |
| Academy-Specific Test | | | | | | | | |
| 1981 study[21] | . | .51* | .57** | . | .47** | .51** | .57** | .56*** |
| 1983 study[22] | . | .66** | .35* | . | .56*** | * | * | .63*** |
| | | | | | | .51** | .54** | |
| | | | | | | * | * | |
| Instructor Ratings | | | | | | | | |
| Current study[23] | . | .44* | .30** | . | .31** | .32** | .43** | .36*** |
| | | | | | | * | * | |
| Overall success/failure | | | | | | | | |
| 1987 study[24] | . | .38** | (.13) | (.15) | .22** | .16** | .24** | .16* |
| Current study[25] | . | (.00) | (.07) | . | (.16) | * | * | (.15) |
| | | | | | | .16** | .13** | |
| **Job Performance** | | | | | | | | |
| FTO ratings | | | | | | | | |
| 1987 study[26] | . | . | . | . | . | .30** | .38** | . |
| Current study[27] | . | (.12) | (.07) | . | (.17) | .25** | * | .20* |
| | | | | | | * | .29** | |
| | | | | | | | * | |
| Patrol Supervisor ratings | | | | | | | | |
| 1987 study[28] | . | . | (.31) | . | .38* | .19** | .22** | (.23) |
| | | | | | | | * | |

[17]Asian N=12, Black N=19, Hispanic N=34, White N=154, Male N=191, Female N=28. Reading test includes two cloze tests (no multiple-choice test).

[18]Asian N=17, Black N=41, Hispanic N=60, White N=346, Male N=405, Female N=75.

[19]Black N=111, Hispanic N=137, White N=953, Male N=1103, Female N=167.

[20]American Indian N=137, Asian N=407, Black N=996, Filipino N=154, Hispanic N=1821, White N=9495, Male N=11149, Female N=1882. Includes all available Read/Write scores matched to Proficiency Test scores obtained between Aug83-Feb92.

[21]Asian N=12, Black N=19, Hispanic N=35, White N=152, Male N=190, Female N=28.

[22]Asian N=17, Black N=41, Hispanic N=60, White N=346, Male N=405, Female N=75.

[23]Asian N=28, Black N=68, Hispanic N=86, White N=303, Male N=356, Female N=145. Criterion is FTO mean rating on 4 writing abilities.

[24]Asian N=48, Black N=107, Filipino N=18, Hispanic N=181, White N=886, Male N=1058, Female N=194. Pass/fail index: Graduated=1; failed/withdrew for academic reasons=0. Correlations are point-biserials.

[25]Asian N=27, Black N=58, Hispanic N=64, White N=258, Male N=342, Female N=79. Pass/Fail index: Completed (C1,2)=1; resigned or terminated due to inadequate writing skills or other KSAs (R2,7 & T2)=0. Correlations are point-biserials.

[26]White N=81, Male N=86. Criterion is FTO rating of officer performance of job duties that involve writing.

[27]Asian N=21, Black N=47, Hispanic N=49, White N=205, Male N=236, Female N=93. Criterion is FTO mean rating on 4 writing abilities.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Field training success/failure<br>1987 study[29]<br>Current study[30] | .<br>. | .<br>(.05) | .20*<br>(.15) | .<br>. | (.02)<br>(.04) | .07*<br>(.05) | .09**<br>(.06) | .17*<br>(.05) |
| Probation success/failure<br>1987 study[31] | . | . | (.15) | . | (.04) | .09** | .12***<br>* | (.15) |

***p<.001; **p<.01; *p<.05 (one-tailed); Correlations are not reported in instances where N was very small (<10).

## Differential Prediction

A differential prediction analysis of error variance, slopes, and intercepts, together with an analysis of residuals derived from the actual versus the predicted criterion performance, was computed to determine whether the Test Battery and the Essay Test are unfair to members of any protected group. Unfairness, in this sense, means the systematic under prediction by the test battery of achievement in the academy or field performance. The analysis was computed for each of the following performance criteria: academy instructor ratings, academy success/failure, academy proficiency test scores, and Field Training Officer (FTO) ratings of officers' writing abilities demonstrated throughout field training.

The results of all analyses indicate that the test battery is not unfair to racial/ethnic minorities in predicting measures of their performance in basic training and field training. That is, their performance was not significantly under predicted by test scores; in fact, their performance was often over predicted. The results for females, however, were mixed. Differential prediction was detected for some performance measures; proficiency scores tended to be over predicted while academy and field training performance ratings were under predicted. These results are displayed in Table 6. In this table only those residual analysis outcomes which produced a significant effect, whether over prediction (+) or under prediction (-) are noted.

[28]Black N=15, Hispanic N=31, White N=311, Male N=337, Female N=37. Criterion is patrol supervisor rating of tenured officer performance of job duties that involve writing.

[29]Black N=95, Hispanic N=118, White N=795, Male N=929, Female N=133. Pass/fail index: Completed (C1-3)=1; Failed to complete (R2,3 or T2,3 or F2,3)=0. Correlations are point-biserials.

[30]Asian N=24, Black N=64, Hispanic N=64, White N=236, Male N=275, Female N=126. Pass/fail index: Completed (C1-3)=1; Resigned or terminated due to inadequate writing, analytical, or other KSAs, or performance level unknown (R3,4,5,7 or T3,7)=0. Correlations are point-biserials.

[31]Black N=89, Hispanic N=108, White N=651, Male N=779, Female N=116. Pass/fail index: Completed (C1-3)=1; Failed to complete (R2,3 or T2,3 or F2,3)=0. Correlations are point-biserials.

TABLE 6: Summary of Differential Prediction Analyses

| Criterion Measure | Am. Indian | Asian | Black | Filipino | Hispanic | Female |
|---|---|---|---|---|---|---|
| **Academy Performance** | | | | | | |
| POST Proficiency Test | | | | | | |
| 1981 Study | * | * | + | * | + | + |
| 1983 Study | * | + | + | * | + | + |
| 1987 Study | * | * | + | * | + | + |
| 1994 Study | * | + | + | + | + | + |
| Academy-Specific Test | | | | | | |
| 1981 Study | * | * | + | * | + | + |
| 1983 Study | * | * | + | * | * | + |
| Instructor Ratings | | | | | | |
| 1994 Study | * | * | + | * | * | - |
| Overall Success/Failure | | | | | | |
| 1987 Study | * | * | * | * | * | + |
| 1994 Study | * | * | * | * | * | * |
| **Job Performance** | | | | | | |
| FTO Ratings | | | | | | |
| 1987 Study | * | * | * | * | * | * |
| 1994 Study | * | * | * | * | * | - |
| Patrol Supervisor Ratings | | | | | | |
| 1987 Study | * | * | * | * | + | * |
| Field Training P/F | | | | | | |
| 1987 Study | * | * | * | * | * | * |
| 1994 Study | | | | | | |
| Probation P/F | | | | | | |
| 1987 Study | * | * | * | * | * | * |

+ The positive sign indicates an over prediction of the criterion scores.

- The negative sign indicates an under prediction of the criterion scores.

* All others indicate either differential prediction but no residual difference, or too small a sample for analysis, or no differential prediction.

## Validity Generalization

The results of the criterion-related validation studies presented in this manual are consistent with the cumulative research concerning the prediction of peace officer performance in the academy and on the job. In a study of the validity generalization results for law enforcement occupations encompassing 40 validation studies and 381 validity coefficients (Hirsh, Northrop, Schmidt,

1986) it was demonstrated "that cognitive ability tests are excellent predictors of performance in job training programs for police and detectives..." Schmidt and Hunter found evidence that job knowledge of the type acquired in training programs plays a strong causal role in determining job performance capabilities (Schmidt, Hunter & Outerbridge, 1986). These authors also concluded "that several types of cognitive tests have at least a minimally useful level of validity for predicting job performance of law enforcement personnel." They found similar outcomes for the pooled results across all law enforcement occupations. Further, comparable results were found in an earlier study by Ghiselli (1973). These findings support the conclusion that the validity of the POST Entry-Level Test Battery can be generalized to all of the peace officer jurisdictions using the test.

In another study, Hunter (1983) has found that there is a high correlation between cognitive ability and job performance. This high correlation is in part the result of the direct impact of ability differences on performance. However, it is even more the result of the strong relationship between cognitive ability and job knowledge and the high relevance that job knowledge has to job performance.

Concerning job performance, findings by Schmidt, Hunter, Outerbridge, and Goff (1988) indicate that the validity of tests such as the Post Reading and Writing Test remains relatively constant over time. That is, as the experience of incumbents increases, those who score higher on the test will continue to perform better on the job.

Taken together, these studies demonstrate that cognitive tests such as the POST Entry-Level Test Battery correlate well with peace officer training and job performance criteria, that this correlation is the result of the peace officer's ability to acquire job knowledge, and that those who more effectively gain job knowledge (those who score higher an the entry level cognitive test) will continue to perform better on the job over time. These studies offer considerable support for the use of the POST Test Battery in the selection of peace officer candidates.

## SECTION THREE: ADMINISTRATION, SCORING and INTERPRETATION

### Test Administration Overview

#### Administration of the Test Battery

The test battery takes three hours to administer. Thirty minutes are allowed for the proctor to read the test instructions and for the applicants to code various information on the machine readable forms. The scoring procedure is automated, thus requiring the applicants to code their name, address, social security number, etc., on the forms. Two and one-half hours are allowed for actual completion of the test battery. This is a generous time allowance intended to ensure that almost everyone has time to complete the exam.

### Scoring and Interpretation

#### Scoring Procedure

Presently, there are 10 forms of the test available for use. These tests are designed to be similar so that scores on all tests are comparable.

Each multiple-choice item has one correct alternative. The cloze items have from one to five correct responses.

The percentage correct for each subtest is computed. Then the three writing percent scores are added together to form a total writing score, and the two reading percent scores are added to form a total reading score. These two sets of scores are then standardized to T scores.[32] The two T scores are then combined and restandardized to arrive at a single T score for the total test battery.

To ensure stability in converted scores, the means and standard deviations obtained in the original norming of the test are used in calculating all subsequent T's. This makes all scores in all test administrations directly comparable.

#### Test Score Reporting

Users of the test battery receive a computer printout which lists, for each applicant, mean percent correct and T score for the writing tests, mean percent correct and T score for the reading tests, and total standardized T score for the entire test battery. On the last page of the printout summary information regarding total sample, sample by gender, and sample by ethnicity is provided.

[32] Conversion of T Scores results in a distribution of scores that have a mean (average) of 50 and a standard deviation of 10.

Three versions of this exam list are generated. On one, all information is presented in score order based on the total T score. In another, information is presented in alphabetical order based on the candidate's last name. And, on the third, information is ordered according to candidate social security number. Typically, the user agency receives the score and alpha listings and POST retains the ssn listing. In addition, POST provides a printout of summary arrays displaying distributions from highest to lowest score by race/ethnicity and gender together with score frequencies and cumulative frequency percents. When specifically requested, POST will provide the results on a computer diskette with the data in ASCII format.

**Use of Test Results**

POST research has repeatedly demonstrated that there is a linear relationship, across the entire range of scores, between the scores on the POST Test Battery and performance in the academy and on the job. Therefore, the higher a department sets the pass/fail cut score the higher the expected proficiency of those passing, and the lower the cut score the lower the expected proficiency of those passing. While it is the responsibility of the user agency to set the cut score, POST recommends that agencies set cut scores no lower than 40-45, and that cut scores in the high 40's and 50's be set with caution.

Experience with the test has demonstrated that individuals with scores below 40 will either fail the academy, or perform marginally. As is always the case with test predictions, there will be false negatives and individuals with low test scores will occasionally demonstrate the ability to perform adequately in the academy. In the long run, however, low scoring individuals will perform less well than better scoring individuals, and individuals with scores under 40 will perform unsatisfactorily or marginally.

Care should be exercised, however, when deciding how high to set a pass/fail cut score. For while it is true that higher scores predict higher abilities, higher cut scores often increase adverse impact against some protected groups. Thus, considerations other than just predicted performance should play a role in deciding on a specific cut score.

One important factor to consider when setting cut scores is the standard error of measurement for the test. No test is perfectly reliable, therefore, there is always some error in test scores. The standard error of measurement is an estimate of this error. For the POST test battery forms, the average standard error is 4.7 (see Table 2). This means, for example, that for a score of 50 there is a 68 percent chance that the true score is somewhere between 45.3 (50-4.7=45.3) and 54.7 (50+4.7=54.7). Realizing that scores are never error free, one should be cautious when making distinctions between candidates with extremely close scores.

If the test is to be used as a ranking device, the jurisdiction should consider the use of test score bands. A score band is a set of individual test scores grouped in such a way that all scores within the band are considered equal (Cascio, Outtz, Zedeck, and Goldstein, 1991).

In fiscal year 1991-92 the average cut score set on the POST Test Battery by user agencies was 44.8. This was a slight increase over the preceding fiscal year when the average was 43.4. The average cut score has gone up, however, every year since the test was made available.

**Adverse Effect**

The Uniform Guidelines defines adverse impact as:

> "A selection rate for any race, sex, or ethnic group which is less than for-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be evidence of adverse impact."

From the preceding definition it is clear that adverse impact is a function of the relationship between the passing rates of one group relative to another. The adverse impact of a given administration of the Read/Write test is dependent upon the pass/fail score which the using jurisdiction sets. Thus, adverse effect is test administration specific and not something that can be addressed with any accuracy in general terms. This can explain how in one test administration adverse impact may be present and in another it
may not be, even when the same test is used in both instances.

POST cannot predetermine this effect. However, POST does provide the jurisdiction with a set of score distributions according to gender and ethnicity to aid the user in the calculation any such impact.

Regardless of any adverse effect, however, it is important to note that the previously reported differential prediction analysis demonstrates that the test battery is not "unfair" to racial/ethnic protected group individuals. Differences in test score indicate differences on job related skills.

**Use of the Test Battery to Rank Candidates**

It is recommended that users of the test battery give serious consideration to using the scores to rank order candidates. The results reported in Tables 4, and 5 show that scores on the test battery are highly predictive of success in the academy and on the job. As test scores increase, so does performance on the criterion measure. This relationship is shown graphically in Figure 1, and indicates that optimal use of the test battery is achieved by rank ordering candidates.

Figure 1
The Relationship of Scores on the POST Test
Battery with Academy Performance and Job Performance

In scoring a test pass/fail, one assumes that beyond a particular score on the test (the cut score), increasing test scores no longer correspond to increasing performance on the criterion. Graphically this assumption is represented in Figure 2.

Figure 2
The Assumed Relationship of Test Scores and
Criterion Performance when Pass/Fail Scoring is Utilized

The assumption depicted in Figure 2 is incorrect for the POST test battery, and to score the test battery pass/fail is to ignore valuable information regarding individuals who score above the minimum cut score. For example, if the cut score were set at 40 and the test battery was used pass/fail, then individuals with scores of 40 through 75 would functionally be considered to be of equal ability. This is not the case.

To set a cut score requires knowledge of how the test actually performs on the population for which it was designed. Table 7, presents information regarding test norms for both the applicant population and the cadet population. Information regarding the former group is necessary because it provides the test user with estimates of passing rates, relative adverse affect, etc. Information regarding the latter group is important as it provides the test user with the information necessary to predict the academic achievement of those who actually attend an academy. The percentiles for each test form are listed in Appendix 1.

## Test Norms

### Applicant and cadet test norms

Table 7 summarizes applicant test performance on the test battery for the period extending from 1983 to 1994. Information on both the cadet population and on applicant population is presented.

Table 7: Norms for Total Applicant and Cadet Groups

## PERCENTILES FOR ALL FORMS

| %tile | SCORES | | %tile | SCORES | |
|---|---|---|---|---|---|
| | APPLI-CANTS | CADETS | | APPLI-CANTS | CADETS |
| 100 | 78 | 76 | 53 | 49 | 49 |
| 99 | 70 | 71 | 52 | 49 | 48 |
| 98 | 69 | 69 | 51 | 48 | 48 |
| 97 | 68 | 67 | | | |
| 96 | 67 | 67 | | | |
| 95 | 66 | 66 | | | |
| 94 | 65 | 65 | | | |
| 93 | 64 | 64 | | | |
| 92 | 64 | 64 | | | |
| 91 | 63 | 63 | | | |
| 90 | 63 | 62 | | | |
| 89 | 62 | 62 | | | |
| 88 | 62 | 62 | | | |
| 87 | 61 | 61 | | | |
| 86 | 61 | 61 | | | |
| 85 | 60 | 60 | | | |
| 84 | 60 | 60 | | | |
| 83 | 59 | 59 | | | |
| 82 | 59 | 59 | | | |
| 81 | 59 | 58 | | | |
| 80 | 58 | 58 | | | |
| 79 | 58 | 58 | | | |
| 78 | 57 | 57 | | | |
| 77 | 57 | 57 | | | |
| 76 | 57 | 56 | | | |
| 75 | 56 | 56 | | | |
| 74 | 56 | 56 | | | |
| 73 | 56 | 56 | | | |
| 72 | 55 | 55 | | | |
| 71 | 55 | 55 | | | |
| 70 | 55 | 54 | | | |
| 69 | 54 | 54 | | | |
| 68 | 54 | 54 | | | |
| 67 | 54 | 53 | | | |
| 66 | 53 | 53 | | | |
| 65 | 53 | 53 | | | |
| 64 | 53 | 52 | | | |
| 63 | 52 | 52 | | | |
| 62 | 52 | 51 | | | |
| 61 | 52 | 51 | | | |
| 60 | 51 | 51 | | | |
| 59 | 51 | 51 | | | |
| 58 | 51 | 50 | | | |
| 57 | 50 | 50 | | | |
| 56 | 50 | 50 | | | |
| 55 | 50 | 49 | | | |
| 54 | 49 | 49 | | | |

| | | | | | |
|---|---|---|---|---|---|
| 50 | 48 | 48 | 25 | 38 | 37 |
| 49 | 48 | 47 | 24 | 38 | 37 |
| 48 | 47 | 47 | 23 | 37 | 36 |
| 47 | 47 | 47 | 22 | 37 | 36 |
| 46 | 47 | 47 | 21 | 36 | 35 |
| 45 | 46 | 46 | 20 | 36 | 35 |
| 44 | 46 | 46 | 19 | 35 | 34 |
| 43 | 45 | 45 | 18 | 34 | 34 |
| 42 | 45 | 45 | 17 | 34 | 33 |
| 41 | 45 | 45 | 16 | 33 | 32 |
| 40 | 44 | 44 | 15 | 33 | 31 |
| 39 | 44 | 44 | 14 | 32 | 31 |
| 38 | 44 | 43 | 13 | 31 | 30 |
| 37 | 43 | 43 | 12 | 30 | 29 |
| 36 | 43 | 43 | 11 | 30 | 28 |
| 35 | 42 | 42 | 10 | 29 | 27 |
| 34 | 42 | 42 | 9 | 28 | 26 |
| 33 | 42 | 41 | 8 | 27 | 25 |
| 32 | 41 | 41 | 7 | 26 | 24 |
| 31 | 41 | 40 | 6 | 24 | 23 |
| 30 | 40 | 40 | 5 | 23 | 21 |
| 29 | 40 | 40 | 4 | 21 | 19 |
| 28 | 40 | 39 | 3 | 19 | 17 |
| 27 | 39 | 39 | 2 | 16 | 14 |
| 26 | 39 | 38 | 1 | 12 | 11 |

**Cadet Test Norms**

During the period from April 1982 to September 1982, the test battery was administered to a total of 480 cadets from 10 academies (Santa Clara Valley Regional Training Center, Golden West College, Los Angeles PD, Los Angeles SO, San Francisco PD, Alameda SO, Riverside SO, Santa Rosa Junior College, Gavilan College, and California Highway Patrol). The resultant data were analyzed both to obtain normative information regarding the test and to generate validation information.

Table 8 compares the original 1982 norms with the scores achieved by academy cadets ten years later.

Table 8: Cadet Means for the Original Norm Group and for Current Academy Cadets

| | Total | Black | White | Asian | Hispanic | Filipino | Other |
|---|---|---|---|---|---|---|---|
| Original Sample (1982) | | | | | | | |
| Mean | 49.8 | 46.2 | 51.1 | 47.8 | 45.5 | na | na |
| Standard Deviation | 10.0 | 10.0 | 9.6 | 10.3 | 10.7 | -- | -- |
| Sample Size | 480 | 41 | 346 | 17 | 60 | -- | -- |
| Current Sample (1992) | | | | | | | |
| Mean | 52.8 | 46.5 | 55.6 | 51.8 | 47.8 | 47.5 | 46.9 |
| Standard Deviation | 10.1 | 9.3 | 9.0 | 11.7 | 10.3 | 13.6 | 2.7 |
| Sample Size | 560 | 77 | 352 | 32 | 85 | 6 | 5 |

As indicated in the table, test performance overall, and for all ethnic groups for which we have data, improved between 1982 and 1992. The most probable explanation for this improvement was the enactment by POST, in 1982, of a requirement that all law enforcement departments utilize a reading and writing test as a component of the entry-level selection process.

## UTILITY OF THE TEST

### Test Utility

A statistically significant linear relationship exists between Test Battery scores and measures of performance, both in the basic academy and in field training. That is, a significant relationship was found between Test Battery scores and knowledge of the basic training curriculum as measured by Proficiency Test scores. Further, examinees who score lower on the Test Battery tend to be rated lower in their demonstrated writing abilities, while higher scoring examinees tend to be rated higher on these abilities. Test Battery scores are similarly related to overall success or failure in completing basic on-the-job training.

While the empirical validity results are important in that they document the job-relatedness of the Test Battery, there are additional factors which affect the practical utility of the battery, namely, the base rate for successful job performance (the percentage of employees who would be successful without using the test as a screening device) and the passing rate resulting from the cut score used with the test (selection ratio). This is the classic Taylor-Russell model for assessing the utility of a test (Cascio, 1982). Under this model, when validity is held constant, a test will have maximum utility when the base rate is near 50% and when the selection ratio is low (few are selected). As the selection ratio approaches 100% and/or the base rate departs from 50%, the usefulness of the test decreases until there is little or no gain realized from administering the test.

Thus, it is possible for a highly valid test to have low utility when either the selection ratio or the base rate for successful job performance is high. Likewise, it is possible for a marginally valid test to offer substantial utility when the selection ratio is low and the base rate is near 50%.

The following information is helpful in anticipating the effect which use of the test will have on performance in the academy and on the job.

### Expectancy Tables

Estimates of the utility of Test Battery scores in predicting academy and field training performance were computed for the total sample in those instances where significant validities were obtained in the 1994 study. Five levels of Test Battery performance were selected for the analyses representing the lower to upper middle range of test performance in 5-point increments (35-55). In the following tables, two of the five participating agencies used the POST Read/Write Battery and the remaining agencies used alternative measures of reading and writing abilities in their operational hiring practices, the true utility may be under represented as lower scoring examinees at these agencies were less likely to be hired. Thus, the results of these analyses should be viewed as gains relative to existing selection practices.

**Utility in Predicting Academy and On-the-Job Performance**

The utility of Test Battery scores in predicting Academy Proficiency Test scores, academy instructor ratings, and FTO ratings of writing ability for the total 1994 validation sample are summarized in Table 9. The table presents the percentages of academy students performing adequately and above average on the Proficiency Test, and the percentages of academy students and officers in field training rated as satisfactory or above average, who scored above and below each of the five Test Battery score levels as a pass/fail score.

The difference between the two columns for each of the criterion presented in Table 9 is a function of the manner in which satisfactory performance was defined. In predicting academy Proficiency Test scores, adequate performance was defined as a Proficiency Test score greater than, or equal to the 25th percentile (44.1), or higher and above average performance was defined as a score greater than, or equal to the median of the test (52.4). When considering academy instructor ratings and FTO ratings, a mean rating corresponding to adequate (3) on a 5-point rating scale was used to classify academy students and officers in field training as either minimally acceptable or less than acceptable with regard to writing ability. The median composite rating for the total sample (3.7) was used to classify students and officers as above average or below average in writing ability. Thus, the left column of Table 8 focuses on the utility of Test Battery scores in identifying academy students who demonstrate minimally acceptable academic knowledge and writing ability, while the right column focuses upon above average performance. Results pertaining to the former are offered to reflect the minimum standards concerns of POST (i.e., screening out candidates who do not possess minimum requisite abilities). Results pertaining to the latter are intended to reflect the concerns of local hiring authorities (i.e., to select the best qualified candidates).

TABLE 9: Empirical Expectancy Tables

| --A-- | | --B-- |
|---|---|---|
| **MIN ACCEPTABLE PERFORMANCE % GAIN OVER BASE RATE** | | **ABOVE AVERAGE PERFORMANCE % GAIN OVER BASE RATE** |
| **Predicting Academy Proficiency Test Scores (N=13,347)** | | |
| Pass/Fail Score | Base Rate=76.9% | Base rate=50.8% |
| 55 | 16.8% | 37.1% |
| 50 | 11.6% | 23.9% |
| 45 | 7.0% | 13.3% |
| 40 | 3.8% | 6.4% |
| 35 | 1.8% | 2.8% |
| **Predicting Academy Instructor Ratings (N=504)** | | |
| Pass/Fail Score | Base Rate=92.1% | Base=53.8% |
| 55 | 4.8% | 29.4% |
| 50 | 3.7% | 20.2% |
| 45 | 3.1% | 6.8% |
| 40 | 2.1% | 9.6% |
| 35 | 0.9% | 3.3% |
| **Predicting FTO Ratings (N=329)** | | |
| Pass/Fail Score | Base Rate=93.9% | Base Rate=50.2% |
| 55 | 2.9% | 24.8% |
| 50 | 2.3% | 11.3% |
| 45 | 2.2% | 5.3% |
| 40 | 0.6% | 2.5% |
| 35 | 0.1% | 1.0% |

The results (9A - top section) are interesting from several standpoints. With regard to the prediction of adequate performance, Test Battery total scores were found to produce monotonic gains in Proficiency Test performance ranging from 1.8% to 16.8%. Gains in above average performance on the Proficiency Test ranged from 2.8% to 37.1% (9B - top section).

The Test Battery scores in the range studied were found to distinguish between students performing above and below a minimum acceptable level. However, because of the high base rate for adequate writing ability as measured by academy instructor ratings (92.1%), there was little room for improvement in predicting the criterion defined in this way (9A - middle section). Accordingly, the relative percentage gains in academy students identified as adequate performers were small, ranging form less than 1% for a score of 35, to a gain of 4.8% for a score of 55. The relative gains increased monotonically with test score level consistent with the observed significant correlation between test scores and academy ratings.

When the focus is shifted from adequate performance to above average performance (9B - middle section), substantially larger gains in academy performance are realized at each Test Battery score level. For the total sample, percentage gains in academy students rated as above average in writing ability were found to increase monotonically with test score, ranging from 3.3% for a score of 35, to 29.4% for a score of 55.

Again, when the focus is shifted from adequate performance to above average performance (9A and 9B - bottom section), the gains in field training performance associated with Test Battery scores are increased. The results in indicate that for the total sample, such gains increase monotonically and range from 1% for a score of 35, to 24.8% for a score of 55. Also as above, gains in performance were statistically significant only in the 45-55 score range.

## References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). Standards for Educational and Psychological Testing. Washington, D.C.: American Psychological Association.

Braddock, R., Lloyde-Jones, R., & Schoer, L. (1963). Research in Written Composition. Urbana, IL: National Council of Teachers of English.

Breland, Hunter M. (1987). Assessing Writing Skills. New York: College Entrance Examination Board.

Cascio, W., Outtz, J., Zedeck, S. & Goldstein, I. (1991). Statistical implications of six methods of test score use in personnel selection. Human Performance. 4:233-264.

Cooper, P. (1984). The Assessment of Writing Ability: A Review of Research. Princeton, N.J.: Educational Testing Service.

Entin, Eileen B. and Klare, George R. (1978). Some inter-relationships of readability, cloze and multiple choice scores on a reading comprehension test. J of Reading Beh. 4:417-437.

Farr, R. (1969). Reading: What Can Be Measured? Newark: International Reading Association Research Fund.

Ghiselli E. (1973). The validity of occupational aptitude tests in personnel selection. Personnel Psy. 26:461-477.

Goodman, K.S. (1967). Reading: A psycholinguistic guessing game. J of the Reading Specialist. 4:126-135.

Honey, R.A. & Kohls, J.W. (1981). Entry-level Reading and Writing Tests for California Law Enforcement: Final Report (Technical Report No. 4). Sacramento: Commission on POST.

Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job performance. J of Voc Beh. Academic Press Inc. 29:340-362.

Kohls, J.W., Berner, J. & Luke, L. (1979). California Entry-Level Law Enforcement Officer Job Analysis (Technical Report No. 1). Sacramento: Commission on POST.

Krueger, K. (1978). Continuous Updating System Project: Development of a Reading Skills Examination. Phase I. Sacramento: The Selection Consulting Center.

Longo, J.A. (1982). The fry graph: Validation of the college levels. J of Reading. 26:229-34.

Ruddell, R.B. (1963). The Effect of Oral and Written Patterns of Language Structure on Reading Comprehension. Unpublished doctoral dissertation, University of Indiana.

Schmidt, F., Hunter, J. & Outerbridge, A. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. J of Applied Psy. 71(3):432-439.

Schmidt, F. (1986). Validity Generalization Results for Law Enforcement Occupations. Personnel Psychology, Inc. 39:399-420.

Taylor, W.L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly.

Uniform Guidelines on Employee Selection Procedures. (1978). Washington, DC: Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice and Department of Labor.

Weiner, J. (1994). POST Reading and Writing Test Validation Research. Sacramento: Commission on POST.

White, E.M. (1982). Some issues in the testing of writing. Notes from the National Testing Network in Writing. (October) p.17. New York: Instructional Resource Center of CUNY.

# APPENDIX 1