The assessment of writing ability: A comparative cost effectiveness study of indirect and direct measures

Kristine Marie Smith

Follow this and additional works at: http://scholarworks.lib.csusb.edu/etd-project
Part of the Rhetoric and Composition Commons

Recommended Citation
http://scholarworks.lib.csusb.edu/etd-project/640

This Project is brought to you for free and open access by the John M. Pfau Library at CSUSB ScholarWorks. It has been accepted for inclusion in Theses Digitization Project by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
THE ASSESSMENT OF WRITING ABILITY: A COMPARATIVE COST EFFECTIVENESS STUDY OF INDIRECT AND DIRECT MEASURES

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

In Partial Fulfillment
of the Requirements for the Degree
Master of Public Administration

by
Kristine Marie Smith
June 1993
THE ASSESSMENT OF WRITING ABILITY: A COMPARATIVE COST EFFECTIVENESS STUDY OF INDIRECT AND DIRECT MEASURES

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

by
Kristine Marie Smith
June 1993

Approved by:

Dr. Naomi Caiden
Chair, Public Administration

Dr. Chuck Christie
Professor, Public Administration

Dr. Clifford Young
Professor, Public Administration

6/7/93
ABSTRACT

The efficient and effective measurement of writing ability is critical to the successful selection of entry level Deputy Sheriff candidates for the County of San Bernardino. There is much theoretical debate on the best method to measure writing ability with the two major methods being indirect and direct assessment. These two methods were compared using a cost effectiveness approach focusing on test reliability, the correlation between performance on the two test instruments and cost. Data was collected on a sample group of 149 entry level Deputy Sheriff applicants.

Results of the study indicated that the test reliability of the direct instrument was unexpectedly higher than the indirect instrument, the correlation between performance on the two instruments was statistically significant at the .01 level, and that there was a clear cost advantage to using the indirect measure. Given that the reliability of the indirect instrument can be improved by selecting more test questions of an appropriate difficulty level, the County of San Bernardino will continue its current test plan which uses the indirect method of writing assessment only. The direct method of writing assessment using a holistic rating approach will not be included in the test plan for entry level Deputy Sheriff. Although no significant changes are being made based on the research, the study proved valuable because it provided
justification for the current test plan, developed expertise in the area of direct writing assessment, and demonstrated that the Employment Division can operate within the greater scope of management concerns by integrating cost considerations with test planning.
# TABLE OF CONTENTS

**ABSTRACT** ........................................................................................................ iii

**LIST OF TABLES** ................................................................................................. vi

**RESEARCH METHOD** .......................................................................................... 1

  - Problem ........................................................................................................... 1
  - Theory ............................................................................................................ 3
  - Hypotheses ....................................................................................................... 5
  - Operational Definitions .................................................................................... 6
  - Research Design ............................................................................................... 8
  - Instruments ...................................................................................................... 10
  - Data Collection Procedures ........................................................................... 10
    - Sample ......................................................................................................... 10
    - Scoring ........................................................................................................ 11
  - Data Analysis .................................................................................................. 15
    - Summary Statistics ....................................................................................... 15
    - Reliability .................................................................................................... 16
    - Correlation ................................................................................................... 17
    - Cost ............................................................................................................... 18
    - Adverse Impact ........................................................................................... 20
  - Discussion ....................................................................................................... 21

**APPENDICES** .................................................................................................... 24

  - Appendix A POST Scoring Guide ................................................................. 24
  - Appendix B Rating Scales ............................................................................... 26
  - Appendix C Scoring Guide ............................................................................. 28

**REFERENCES** .................................................................................................... 30
LIST OF TABLES

Table 1 Summary Test Statistics..........................16
Table 2 Reliability Estimates..............................16
Table 3 Direct Measure Scoring Time.......................19
Table 4 Cost Comparison for Total Test Group.............20
Table 5 Adverse Impact Analysis............................21
RESEARCH METHOD

PROBLEM

The primary function of the San Bernardino County Employment Division is to attract and identify the most qualified applicants for County jobs. To successfully perform this function, the knowledge, skills and abilities required to perform a given job must be determined and assessed using valid, reliable and cost effective selection instruments. Although the use of effective selection procedures has long been viewed by large private sector firms as making good business sense, the primary impetus for selection systems in the public sector has been the concept of merit. However, as governmental agencies continue to experience fiscal stress, the value of selection from an economic perspective has been magnified. As a departmental concern, the human resources manager must assure that resources are maximized by utilizing efficient selection procedures. As an organizational concern, all managers depend upon effective selection procedures to reduce the cost of poor selection decisions.

These concerns are increasingly becoming issues of interest to top level executives. In his book, Staffing the Public Service, Albert Maslow noted this trend emerging in the 1980's and explained that, "Executives will look to the personnel professional for informed advice on many matters... Upper management will seek evaluation of its policies and
practices in terms of cost effectiveness and their impact on human resources" (Maslow, 1989). Proactive personnel agencies will not wait until their organizations are crippled by a lack of funding and thus, unable to support selection research. Instead, they will undertake such projects as soon as possible to demonstrate their ability to contribute to the organization in difficult times.

It was this trend that prompted a study of the assessment of writing ability for the entry level Deputy Sheriff classification. This classification was selected for study because it provided an opportunity to address both efficiency and effectiveness concerns. From the standpoint of efficiency, the entry level Deputy Sheriff represents one of the County's largest job classifications, and similarly one of the largest applicant groups. As such, the selection process utilized is one of the most resource intensive examinations administered by the Employment Division.

Therefore, the effect of inefficiencies in test administration and scoring may significantly impact the work of staff. Given this, it is essential for the smooth operation of the Employment Division that the Deputy Sheriff selection process be conducted with maximum efficiency. The need for efficiency, however, cannot be met through sacrifices in testing effectiveness.
THEORY

Writing ability is one of the dimensions identified as critical to job success by the California Commission on Police Officer Standards and Training (Honey & Kohls, 1982). As such, the County is mandated to assess this area in its selection process. The nature of the training academy and the day-to-day work of a Deputy Sheriff also dictate that candidates must possess basic writing skills to be successful. Therefore, even without the POST mandate, the County needed to effectively assess writing ability. By identifying those candidates who are unlikely to complete the academy or perform adequately on the job, the high costs of turnover are reduced. The cost of turnover is a significant issue for the entry level Deputy Sheriff due to the large investment in screening and training employees.

Writing ability may be assessed by using either indirect or direct methods. Indirect methods are typically multiple choice format examinations that include questions in the areas of spelling, punctuation, grammar and editing (Hoffman & Holden, 1990). In contrast, direct methods of writing assessment require the examinee to produce original written text. There has been a great deal of debate about which method is preferable focusing on test validity, reliability and cost effectiveness.
Indirect measures have been favored in the past by employers because they have demonstrated predictive and concurrent validity, tend to have high reliability and are extremely cost effective due to the use of machine scorable tests (Honey, 1990; White, 1985; Quellmalz, Capell & Chou, 1982). Critics, on the other hand, argue that indirect measures are weak in construct and content validity and that such measures fail to evaluate unity, content or organization (Ackerman & Smith, 1989). Conversely, the strengths of the direct assessment of writing ability are these same areas: content validity and the ability to evaluate unity, content and organization (Friend & Denning, 1990; Honey, 1990; Wansor, 1986). Direct measures of writing ability are typically scored using rating scales or point systems based on identifying attributes in the writing sample (Mullis, 1984). All of these techniques involve the evaluation of writing by other individuals. This introduction of the human factor increases both the time required to score the samples and the subjectivity of the scoring process (McKinney, 1990). These disadvantages of direct assessment including the high costs associated with scoring and a tendency towards low inter-rater reliability have made many employers hesitant to utilize them (McKinney, 1990; Friend & Denning, 1990).

As a result, there has been a movement towards a holistic model of writing assessment (Friend & Denning, 1990). This
model of writing assessment, defines writing as an overall process and is purported to increase the reliability and cost effectiveness of direct writing assessment, therefore, possibly making this a more viable alternative (Honey, 1990; White, 1985). Essentially, holistic rating of direct writing involves providing a single score based on the total impression provided by the writing sample (Mullis, 1984). This method greatly reduces the time required for scoring of essay tests and, thus, the costs associated with direct writing assessment (Huot, 1990; White, 1985). The issue of reliability is addressed by extensive training and monitoring of graders (White, 1985).

The current study is a comparison of indirect and direct, holistically scored, measures of writing ability relative to reliability and costs. Edward White, author of Teaching and Assessing Writing, indicates that studies have been conducted to analyze the relationship between scores on indirect and direct writing tests (White, 1985). However, this study was slightly different in that the research was done in an employment setting with consideration given to organizational issues.

HYPOTHESES

The following three hypotheses were tested using a descriptive and correlational research design:
1. The test reliability estimate obtained from an indirect assessment of writing will be higher than the reliability estimates obtained for a direct assessment utilizing a holistic scoring system.

2. Scores obtained from indirect and direct measures of writing ability for Deputy Sheriff positions will have a statistically significant, positive correlation.

3. The cost per examinee of administering a direct writing assessment test using holistic scoring will be greater than the cost of administering a machine scorable, indirect assessment instrument.

OPERATIONAL DEFINITIONS

The following definitions are provided to clarify the hypothesis and the context of the research:

Indirect measure of writing assessment: A paper and pencil test in multiple-choice format, which includes questions on spelling, punctuation, grammar, and editing.

Direct measure of writing assessment: A test which requires the test taker to produce original written text in response to a specific assignment.
**Concurrent Validity:** An approach to validation in which predictor and criterion information are obtained for present employees at approximately the same time (Maslow, 1989).

**Construct Validity:** A demonstrated relationship between underlying traits inferred from behavior and a set of test measure related to those constructs (Maslow, 1989).

**Content Validity:** A relationship between job performance and a test that is self-evident because the test includes a representative sample of job tasks (Maslow, 1989).

**Predictive Validity:** A demonstrated relationship between test scores of applicants and some future behavior on the job (Maslow, 1989).

**Reliability:** The consistency of measurement.

**Correlation:** The measure of the relationship between the mean scores obtained by a test group on indirect and direct measures on writing ability. For the purposes of this study, the correlation between performance on the two instruments was computed as part of the automated test scoring process.

**Criterion:** A measure of job performance.
Mean Score: The average test score obtained by the sample group.

Efficiency: The extent to which the test may be administered with a minimum input of resources such as staff time.

Effectiveness: The performance of a test in terms of validity and reliability. In this case, the tests were considered equal in validity given the use of the content validation model. Therefore, the primary effectiveness indicator in this study is test reliability.

Cost: The amount of resources dedicated to administer and score a test. Resources are essentially the staff time dedicated to complete a test process given that there were no large equipment or supply expenditures related to testing. Costs were determined by assigning a value to clerical and professional staff time and computing the amount of time required to score the test for the sample group.

RESEARCH DESIGN

To test the hypothesis, a cross-sectional research design was employed using a descriptive and correlational methodology. The cross-sectional design was used because, unlike a longitudinal study where a variable is measured to determine changes over time, the variables being measured were
both independent and needed only single event descriptive data to test the hypothesis. This was due to the fact that the study focused on the test instruments rather than individual performance over time, therefore, requiring only a single sample from a representative test population. The descriptive and correlational methodology was also utilized for this reason. Because the intent was not to measure the effect of one variable on another, but rather to obtain comparative data with which to evaluate the test instruments, descriptive and correlational information was appropriate to test the hypothesis. Specifically, the reliability indicators, test scoring costs and the relationship between the two instruments were most readily obtained using descriptive techniques.

In addition to the overall project design, several research design factors were considered in determining an appropriate group for test administration. The decision to use an actual applicant group, rather than an incumbent group, was based on three factors. First, the data collection was less disruptive to County operations by administering the instruments to a test group which was already scheduled as opposed to an incumbent group that would be taken away from other duties to participate in the study. Secondly, using applicants eliminated the restriction of range in the sample that would be encountered had a group of incumbents who had all achieved scores sufficient for hire been used. Finally,
because the tests are designed to select entry-level employees, an applicant group was more appropriate than an incumbent group which would have been influenced by training on the job.

**INSTRUMENTS**

The indirect instrument used in this study is a component of the written test currently used to select entry level Deputy Sheriffs. The twenty-four multiple choice questions include grammar, word usage, punctuation, spelling and basic writing practices.

A direct instrument was developed for the purpose of this study. Because the most frequent use of writing on the job is to complete incident reports, the exercise utilized the descriptive mode of discourse. The writing topic selected involved describing an accident scene. Candidates were provided with a picture and then given thirty minutes to complete a written description of the scene. It should be noted that due to test security concerns, the County of San Bernardino was unable to release the test instruments for inclusion in this publication.

**DATA COLLECTION PROCEDURES**

**Sample**

Both instruments were administered to an entry level
Deputy Sheriff applicant group of 1151 candidates. From this group, 149 cases were randomly selected for inclusion in the study by selecting every seventh candidate from the applicant pool. The 149 cases were selected because this number would provide sufficient statistical power while keeping the time invested in scoring at a minimum. The sample group consisted of the following:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Cases</th>
<th>Ethnicity</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>113</td>
<td>White/Caucasian</td>
<td>88</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>Black</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hispanic</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asian</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Native American</td>
<td>3</td>
</tr>
</tbody>
</table>

Scoring

The indirect measure was administered utilizing machine scorable answer sheets, which allowed for immediate scoring following test administration. The direct measure ratings, which were ultimately also machine scored, required many additional steps before final scoring could be completed. Existing literature was reviewed to provide the model for conducting the holistic rating session, with the final evaluation plan incorporating suggestions from the work of White, Mullis and Honey. The scoring plan emphasized standardization and training to increase both inter-rater
reliability and rating speed. Therefore, four major steps were included: Rating and training material development, a benchmark session, rater orientation and training, and the actual scoring session.

Prior to scoring the writing samples, rating and training materials were developed. The rating materials were adapted from those used by the Commission on Police Officer Standards and Training (POST) (Appendix A) and consisted of a single six point rating scale describing benchmark rating criteria (Appendix B) and a companion sheet providing specific examples of characteristics that papers at the benchmark levels might include (Appendix C). Writing samples not included in the study group were screened to select papers for the benchmark rating session. This preliminary review was conducted to assure that the papers presented to the benchmark group represented the full range of the rating scale. The full range of the scale could have been achieved by selecting papers at random; however, it is likely that the benchmark panel would have had to review many more papers to reach representation and, this would have increased the time invested in the process.

The sample papers were reproduced and placed into packets of five or six papers for the benchmarking session. Each packet was rated and then discussed sequentially by the panel.
which included three Personnel Analysts and a Supervising Personnel Analyst. As a result, benchmark papers were identified and the rating materials were fine-tuned.

Six raters were obtained to rate the test papers. These individuals were all professionals who were employed in a variety of human resource areas. The diversity of the raters was intentional because it was felt that using examination analysts from the same jurisdiction would influence reliability estimates due to similarity in training and experience. Therefore, the rating panel included individuals from two counties working in the areas of employment and classification. The rating session began with an extensive training period which included an introduction to the project and the method of holistic rating, a review of the exercise given to the candidates and presentation of the rating materials. Following this overview, a packet of benchmark papers was rated by the group. Each member of the group then indicated the score they assigned to the first paper in the packet. Feedback was also given at this point on the consensus score obtained from the benchmark group. Following this, the group was encouraged to discuss the rationale for their individual ratings. Such discussion resulted in clarifying the rating scales and bringing the group to a common basis of interpretation. Each rater appeared to have individual biases regarding writing that influenced ratings
and these were addressed during the discussion period. This process was repeated for each paper in the packet and then for three more packets of training papers. At the conclusion of the training session, the raters were consistently rating training papers within one point of one another.

For the actual rating session, the raters were divided into three panels of two raters each and given approximately fifty papers to score. The panels were closely monitored with any paper receiving ratings more than one scale point apart being referred to a third rater from one of the other rating panels. Of the 149 papers scored, 20 required an additional rating. In fourteen of those twenty cases, the third rating was identical to one of the original two ratings. In the other six cases, the following decision rules were developed to determine which of the two original ratings to retain:

1. If the original raters gave scores of three and five and the third rating is a four, the ratings with scores of four and five should be retained. The rationale for this rule is that two of the three raters gave the paper a passing score as defined by the rating scales. (See Appendix A) This decision rule was applied in two cases.

2. If the original raters gave scores of two and four and the third rating is a three, the ratings with scores of
two and three should be retained. This rule is consistent with decision rule number one in that a consensus approach is taken. In this case, two of the three raters viewed the paper as inadequate. This decision rule was applied in one case.

3. For all other ratings that are two score points apart, if the third rating falls in between the two original ratings, the third rating and the highest original rating should be retained. The rationale for this rule is that it is consistent with decision rule #1 and it gives the candidate the benefit of the doubt. This decision rule was applied in two cases.

4. If the original ratings are two or more points apart and the third rating does not fall between them, but rather, on either end of the scale, the rating nearest the third rating should be retained. This rule again utilizes the consensus approach in that the two closest ratings are retained. This rule was applied in one case.

DATA ANALYSIS

Summary Statistics

Scoring information is presented for both test measures in Table 1 below.
Table 1
Summary Test Statistics

<table>
<thead>
<tr>
<th></th>
<th>Indirect Measure</th>
<th>Direct Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>14.93</td>
<td>3.91</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.85</td>
<td>.96</td>
</tr>
<tr>
<td>Minimum Score</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Maximum Score</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>Maximum Possible Score</td>
<td>24</td>
<td>6</td>
</tr>
</tbody>
</table>

Reliability

The reliability of the indirect measure was estimated at .53 using the Kuder-Richardson 20 method (Ebel & Frisbie, 1986). The reliability of the direct measure, inter-rater reliability, was estimated by computing the correlation between ratings for each rating panel and then computing the average of these correlations using Fisher's Z Transformation (Glass & Stanley, 1970). This data is presented in Table 2.

Table 2
Reliability Estimates

<table>
<thead>
<tr>
<th></th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect Measure</td>
<td>.53</td>
</tr>
<tr>
<td>Direct Measure:</td>
<td></td>
</tr>
<tr>
<td>Rating Panel 1</td>
<td>.75</td>
</tr>
<tr>
<td>Rating Panel 2</td>
<td>.59</td>
</tr>
<tr>
<td>Rating Panel 3</td>
<td>.79</td>
</tr>
<tr>
<td>Rating Panel 4 (3rd Ratings)</td>
<td>.89</td>
</tr>
<tr>
<td>Average Reliability: Direct Measure</td>
<td>.78</td>
</tr>
</tbody>
</table>
This data rejected the first hypothesis regarding test reliability. However, this finding alone does not indicate that the use of the indirect measure should be discontinued. Although, it does demonstrate that the current test should be modified to improve its reliability. Fourteen of the twenty-four test questions were outside of the moderate difficulty range suggested by Ebel (Ebel & Frisbie, 1986), being either too easy or too difficult. Replacing these items with items of a more appropriate difficulty level should improve test reliability. In addition, increasing the total number of test items should also improve reliability. With these modifications, it is likely that the reliability of the indirect instrument can be improved to a level that is at least comparable to that achieved with the direct measure, if not greater.

Correlation

The uncorrected correlation between performance on the two measures was .427. This is significant at the .01 level and supports the second hypothesis. When correction is made for attenuation (unreliability) in the indirect measure, the correlation is .667. Given that there is a significant relationship between the scores achieved on the tests, for the purposes of employment screening, where time and cost are critical elements, it would be reasonable to utilize only one of the two test instruments.
The best indicator of the cost to administer the tests is the time invested in the scoring process. For the indirect measure, this includes computer scanning of the test answer sheets which required approximately one second of clerical time per candidate. The direct measure ratings were also scanned by the computer and, because there were two ratings per candidate, the scanning time was approximately two seconds. However, compared to the rest of the scoring process, this difference is negligible. The key difference is in the time invested to develop rating and training material, train the raters and, finally, rate the papers.

The most time intensive activity in developing the rating and training materials was the benchmark session. This session included four individuals and lasted two and one-half hours for a total of ten work hours. For the actual rating session, three hours were spent training the six raters and then two and one-half hours were spent scoring the papers. In addition, two administrators conducted the training and monitored the rating. Combined, the time of these eight individuals totals 44 hours. The total time investment in scoring the direct test for the sample group was fifty-four hours, or 21.75 minutes per candidate. This time estimate does not include the administrative overhead required to arrange the benchmark and rating sessions and to prepare the
materials for these sessions. A breakdown of scoring time required for the sample group and the total test group is provided in Table 3.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Sample Group</th>
<th>Total Test Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark Session</td>
<td>10 work hours</td>
<td>10 work hours</td>
</tr>
<tr>
<td>Rater Training Session</td>
<td>24 work hours</td>
<td>24 work hours</td>
</tr>
<tr>
<td>Rating Time</td>
<td>20 work hours</td>
<td>153 work hours*</td>
</tr>
<tr>
<td>Total Hours</td>
<td>54</td>
<td>187</td>
</tr>
<tr>
<td>Time per candidate</td>
<td>21.75 minutes</td>
<td>9.75 minutes</td>
</tr>
</tbody>
</table>

* Estimate based on rating time for Sample Group

In converting the time invested in scoring to actual costs, assumptions were made about the dollar value per hour of staff time. Clerical staff time was estimated at fifteen dollars per hour, while professional staff time was estimated at thirty dollars per hour. The indirect measure required only clerical time for scanning the test answer sheets. The direct measure involved professional time in the benchmark, training and rating sessions in addition to the clerical time for computer scanning. A breakdown of the estimated costs to score each measure is provided in Table 4.
As is illustrated by Table 4, the difference between scoring costs for the indirect and direct measures is extremely large. The total cost of scoring the entire test group using the indirect measure is less than scoring only one candidate using the direct measure. This cost for the total test group makes the cost ratio of using the direct measure more than a thousand times greater than using the indirect measure, and thus, supports the third hypothesis regarding costs.

**Adverse Impact**

Because the literature did not provide a foundation to develop a hypothesis regarding adverse impact, this area was not included in the research objectives. However, the tests were evaluated to determine if test performance differed by
gender or ethnicity. In both cases, the differences between the mean scores for the various groups were not significant. Data from this analysis is presented in Table 5.

Table 5

<table>
<thead>
<tr>
<th>Indirect Measure</th>
<th>N</th>
<th>X</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>35</td>
<td>14.71</td>
<td>2.84</td>
</tr>
<tr>
<td>Black</td>
<td>20</td>
<td>15.35</td>
<td>2.67</td>
</tr>
<tr>
<td>Caucasian</td>
<td>88</td>
<td>15.07</td>
<td>2.81</td>
</tr>
<tr>
<td>Male</td>
<td>113</td>
<td>14.67</td>
<td>2.66</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>16.03</td>
<td>2.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Direct Measure</th>
<th>N</th>
<th>X</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>35</td>
<td>3.64</td>
<td>1.03</td>
</tr>
<tr>
<td>Black</td>
<td>20</td>
<td>3.90</td>
<td>.94</td>
</tr>
<tr>
<td>Caucasian</td>
<td>88</td>
<td>3.93</td>
<td>.91</td>
</tr>
<tr>
<td>Male</td>
<td>113</td>
<td>3.83</td>
<td>1.09</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>4.15</td>
<td>.92</td>
</tr>
</tbody>
</table>

DISCUSSION

The purpose of this study was to compare indirect and direct measures of writing ability to determine the most cost effective approach. The goal was to evaluate these two testing alternatives so that the Employment Division might maximize its own resource utilization while providing optimum service to the Sheriff's Department. Because, at this point,
both measures utilize a content validation strategy, the study focused on test reliability, the relationship between performance on the two instruments, and costs.

The research demonstrated a significant relationship between performance on the two measures, and a clear cost advantage in utilizing the indirect measure. Since there are obvious, readily applicable means of improving the reliability of the indirect measure, the County at this time will revise and continue using the indirect measure of writing ability to assess entry level Deputy Sheriff candidates. It is possible that in the future, criterion information will be collected that will provide the opportunity to consider validity data and, thus re-evaluate this decision.

Although the direct instrument will not be implemented as part of the test battery, its development and administration provided valuable information to the Employment Division. First, it provided the comparative basis to support the current test plan for entry level Deputy Sheriff. In addition, it demonstrated that the holistic method of writing assessment is a very viable alternative which can be utilized confidently for other job classifications where examinations developed by the County are administered to a more reasonable number of applicants. The study demonstrated that the traditional concerns about direct writing assessment, high
cost and low reliability, can be effectively mitigated with this approach. Finally, the project allowed the Employment Division the opportunity to demonstrate the benefits of including additional information relating to cost effectiveness in the examination planning and decision making process. By using skills and knowledge already available to examination analysts, examination units may provide tangible evidence supporting their procedures and more effectively address organizational concerns. It is this type of approach that will enable those in selection research to contribute more directly to their organizations and become a more valuable asset to top level management.
Appendix A

POST SCORING GUIDE

Candidates should be rewarded for what they do well. They are asked, first, to narrate or describe an event or situation from personal experience. In the last part of the prompt, they are directed to provide some sort of analysis of the experience. Although the assignment calls for a two-part response, one part may be implicit in the other.

RANGE OF SCORES

6  The "6" essay will be fluent, well developed, and well organized. It will show clear command of language and will be relatively free of errors in sentence structure, grammar, and mechanics.

5  The "5" paper, despite occasional faults, will be generally well written and well organized. It will be less fluent and less detailed than the "6" paper, but will demonstrate greater facility than the "4" paper.

4  The "4" paper will demonstrate basic writing competence, though it may have some problems in sentence structure, diction, or mechanics or have limited development.
3 The "3" paper may not provide adequate development, may lack detail and specificity, or may be poorly organized. It usually has problems in diction, grammar, and mechanics.

2 The "2" paper may lack coherence or adequate development. Generally, it will be marred by multiple errors in sentence structure, grammar, and mechanics. It suggests incompetence.

1 The "1" paper will show clear incompetence.

Non-response papers and off-topic papers should be given to the chief reader.
6 A Superior Report. The "6" report narrative will be fluent, well developed, and well organized. Relevant information will be clearly and logically presented. It will show clear command of the language and will be relatively free of errors in sentence structure, grammar, and mechanics.

5 A Proficient Report. The "5" report, despite occasional technical writing faults, is generally well written and well organized. It may not be as logically sound and well developed as the "6" report.

4 A Basically Competent Report. The "4" report will demonstrate basic writing competence, though it may have some problems in sentence structure, diction, or mechanics. It will contain the essential information and elements related to a given incident, but there may be only marginally acceptable logical development. Parts of the report may have to be rewritten to ensure proper interpretation.

3 An Inadequate Report. The "3" report may not provide adequate development, may lack detail and specificity,
or may be poorly organized. Important information or elements may be omitted, and logically development may be inadequate. A "3" report usually has problems in diction, grammar, and mechanics. Reports written at this level will typically require revision and rewriting.

2 An Unacceptable Report. The "2" report, may lack coherence or adequate development. Such a report often omits much necessary information. Generally, it will be marred by multiple errors in sentence structure, grammar, and mechanics. It suggests incompetence. Reports at this level require complete rewriting. If the content is sufficiently weak, the report may not even be suitable for revision.

1 An Incompetent Report. The "1" report will show clear incompetence. It will not be suitable for revision.

* Adapted from the POST Scoring Guide by Richard Honey.
Appendix C

SCORING GUIDE*

Score of 6: SUPERIOR
- Addresses the assignment fully.
- Shows substantial depth, fullness, and complexity of thought.
- Demonstrates clear, focused, unified, and coherent organization.
- Is fully developed and detailed.
- Evidences superior control of diction, syntactic variety, and transition; may have a few minor flaws.

Score of 5: STRONG
- Clearly addresses the assignment.
- Shows some depth and complexity of thought.
- If effectively organized.
- Demonstrates control of diction, syntactic variety, and transition; may have a few flaws.

Score of 4: COMPETENT
- Adequately addresses the assignment.
- Shows clarity of thought but may lack complexity.
- Is organized.
- Is adequately developed, with some detail.
- Demonstrates competent writing, may have some flaws.
Score of 3: FLAWED IN SOME IMPORTANT WAY
- May be simplistic or stereotyped in thought.
- May demonstrate problems in organization.
- May show patterns of flaws in language, syntax, or mechanics.
- May be undeveloped.

Score of 2: INADEQUATE
- Will demonstrate serious inadequacy in one or more of the areas specified for the 3 paper.

Score of 1: INCOMPETENT
- Failed attempts to begin discussing the topic.
- Papers so incompletely developed as to suggest or demonstrate incompetence.
- Papers wholly incompetent mechanically.

* Adapted from the POST Scoring Guide by Richard Honey.
REFERENCES


