A predictive validation study of criterion-referenced tests for the certification of soldiers in specialist-level military training programs

Darl McDaniels

College of William & Mary - School of Education

Follow this and additional works at: https://scholarworks.wm.edu/etd

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://dx.doi.org/doi:10.25774/w4-m029-6v54

This Dissertation is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the original text directly from the copy submitted. Thus, some dissertation copies are in typewriter face, while others may be from a computer printer.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyrighted material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is available as one exposure on a standard 35 mm slide or as a 17" × 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. 35 mm slides or 6" × 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
A predictive validation study of criterion-referenced tests for the certification of soldiers in specialist-level military training programs

McDaniels, Darl, Ed.D.
The College of William and Mary, 1988
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages
2. Colored illustrations, paper or print
3. Photographs with dark background
4. Illustrations are poor copy
5. Pages with black marks, not original copy
6. Print shows through as there is text on both sides of page
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements
9. Tightly bound copy with print lost in spine
10. Computer printout pages with indistinct print
11. Page(s) lacking when material received, and not available from school or author.
12. Page(s) seem to be missing in numbering only as text follows.
13. Two pages numbered. Text follows.
14. Curling and wrinkled pages
15. Dissertation contains pages with print at a slant, filmed as received
16. Other

_________________________________________________________________________
_________________________________________________________________________

UMI
A PREDICTIVE VALIDATION STUDY OF
CRITERION-REFERENCED TESTS FOR
THE CERTIFICATION OF SOLDIERS
IN SPECIALIST-LEVEL
MILITARY TRAINING
PROGRAMS

A Dissertation
Presented to
The Faculty of the School of Education
The College of William and Mary in Virginia

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Education

By
Darl McDaniels
February 1986
A PREDICTIVE VALIDATION STUDY OF
CRITERION-REFERENCED TESTS FOR
THE CERTIFICATION OF SOLDIERS
IN SPECIALIST-LEVEL
MILITARY TRAINING
PROGRAMS

By

Darl McDaniel

Approved February 1988 by

ROBERT MAIDMENT, Ed.D.
ARMAND J. GALFO, Ed.D.
ROBERT J. HANNY, Ed.D., Chairman
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF CHARTS</td>
<td>vii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ix</td>
</tr>
<tr>
<td>Chapter I. INTRODUCTION.</td>
<td>1</td>
</tr>
<tr>
<td>Justification for Study</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>4</td>
</tr>
<tr>
<td>Theoretical Rationale</td>
<td>5</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>8</td>
</tr>
<tr>
<td>Research Hypotheses</td>
<td>11</td>
</tr>
<tr>
<td>Intervention, Sample Description, and General Data Gathering Procedures</td>
<td>12</td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td>15</td>
</tr>
<tr>
<td>Ethical Considerations</td>
<td>16</td>
</tr>
<tr>
<td>II. REVIEW OF LITERATURE.</td>
<td>18</td>
</tr>
<tr>
<td>Summary of Rationale and Its Relationship to Problem</td>
<td>18</td>
</tr>
<tr>
<td>Summary of Relevant Research</td>
<td>19</td>
</tr>
<tr>
<td>Standardized Objective Tests</td>
<td>19</td>
</tr>
<tr>
<td>Criterion-Referenced Measurement</td>
<td>21</td>
</tr>
<tr>
<td>Test Validity and Cutoff Scores</td>
<td>26</td>
</tr>
<tr>
<td>Summary of Previous Research and Its Relationship to Problem</td>
<td>29</td>
</tr>
<tr>
<td>III. METHODOLOGY</td>
<td>35</td>
</tr>
<tr>
<td>Problem</td>
<td>35</td>
</tr>
<tr>
<td>Specific Null Hypotheses</td>
<td>36</td>
</tr>
<tr>
<td>Populations and Samples</td>
<td>37</td>
</tr>
<tr>
<td>Procedures</td>
<td>39</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>44</td>
</tr>
<tr>
<td>Research Design</td>
<td>46</td>
</tr>
<tr>
<td>Statistical Analysis Technique</td>
<td>47</td>
</tr>
<tr>
<td>Summary of Methodology</td>
<td>49</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>51</td>
</tr>
<tr>
<td>Test Content Validity</td>
<td>51</td>
</tr>
<tr>
<td>Indices of Relative Emphasis Computation</td>
<td>51</td>
</tr>
<tr>
<td>Job Relevance Assessment</td>
<td>54</td>
</tr>
<tr>
<td>Test Cutoff Score Computation</td>
<td>57</td>
</tr>
<tr>
<td>Raw Score Performance Standard</td>
<td>57</td>
</tr>
<tr>
<td>Percentage Score Performance Standard</td>
<td>58</td>
</tr>
</tbody>
</table>
Standard Errors ........................................... 60
Composite Standard Error Computation ...................... 62
Sampling error computation .................................. 62
Measurement error computation .............................. 63
Composite error ......................................... 65
Cutoff Scores and Associated Failure Rates .................. 68
Summary: Test Content Validity and Cutoff Score Computation ........................................ 71
Description of Study Samples ................................ 72
Test of Hypotheses ........................................ 72
Hypothesis 1 ............................................. 76
Intragroup variability of probability estimates ............... 76
Inter-rater reliability of probability estimates ............... 81
Hypothesis 2 ............................................. 83
Part 1 - Test of Hypothesis 2 ................................ 85
AACTS with AASRS1 ....................................... 85
AACTS with AASRS2 ....................................... 87
AACTS with AASRS3 ....................................... 88
AACTS with AASRS4 ....................................... 88
Part 2 - Test of Hypothesis 2 ................................ 89
PMCTS with PM5RS1 ....................................... 89
PMCTS with PM5RS2 ....................................... 90
PMCTS with PM5RS3 ....................................... 90
PMCTS with PM5RS4 ....................................... 91
Major Findings ........................................... 91
Hypothesis 3 ............................................. 92
Part 1 - Test of Hypothesis 3 ................................ 93
AACTS with AASQT ........................................ 93
Part 2 - Test of Hypothesis 3 ................................ 94
PMCTS with PMSQT ........................................ 94
Major Findings ........................................... 94
Hypothesis 4 ............................................. 95
Part 1 - Test of Hypothesis 4 ................................ 96
Part 2 - Test of Hypothesis 4 ................................ 97
ARCTS with ARSRS1 ....................................... 97
ARCTS with ARSRS2 ....................................... 97
ARCTS with ARSRS3 ....................................... 98
ARCTS with ARSRS4 ....................................... 99
Part 3 - Test of Hypothesis 4 ................................ 100
AASRS1 with ARSRS1 ...................................... 100
AASRS2 with ARSRS2 ...................................... 103
AASRS3 with ARSRS3 ...................................... 103
AASRS4 with ARSRS4 ...................................... 103
Major Findings ........................................... 106
Hypothesis 5 ............................................. 108
Part 1 - Test of Hypothesis 5 ................................ 109
AACTS with AASQT ........................................ 109
LIST OF TABLES

1. Curriculum and Test Emphasis Estimates .................. 52
2. Indices of Relative Emphasis on Curriculum and Test. .......... 53
3. Frequency of Judges' Average Relevance Ratings ................ 56
4. Raw Score Rating Data ........................................ 59
5. Percentage Score Rating Data .................................. 61
6. Statistics for Derived Percentage Score Standards .......... 63
7. Measurement Error and Test Statistics ......................... 65
8. Composite Standard Errors ...................................... 66
9. Measurement Error and Associated Statistics for Percentage Score Standards ........................................ 67
10. Derived and Adjusted Percentage Score Standards .......... 68
11. Student Test Scores and Derived Failure Rates, CY 1985 .......... 70
12. Part 1 of Hypothesis 1 - Two-Way Analysis of Variance of Probability Estimates by Instructors ........................... 78
15. Part 2 of Hypothesis 1 - Interrater Reliability Coefficients and Confidence Intervals ........................................ 82
16. Hypothesis 2 - Correlation Coefficients .......................... 87
17. Hypothesis 4 - Correlation Coefficients .......................... 96
18. Hypothesis 4 - T Values ........................................ 101
19. Part 3 of Hypothesis 4 - T-Test of Significant Difference in Supervisory Rating Scores Concerning Technical Skills of Angoff "Accepts" and Angoff "Rejects" .................. 102
20. Part 3 of Hypothesis 4 - T-Test of Significant Difference in Supervisory Rating Scores Concerning Job Effort of Angoff "Accepts" and Angoff "Rejects" .................. 104
21. Part 3 of Hypothesis 4 - T-Test of Significant Difference in Supervisory Rating Scores Concerning Angoff "Accepts" and Angoff "Rejects" Adherence to Job Procedures ........................ 105
22. Part 3 of Hypothesis 4 - T-Test of Significant Difference in Supervisory Rating Scores Concerning Angoff "Accepts" and Angoff "Rejects" on Equipment Maintenance ............... 107
23. Part 2 of Hypothesis 5 - T-Test of Significant Difference in Skill Qualification Test Scores of Angoff "Accepts" and Angoff "Rejects" ........................................ 112
LIST OF CHARTS

1. Description of Study Samples ........................................ 73
2. Variable Labels.......................................................... 86
DEDICATION

This study is dedicated to my wife, Lenorris, and daughter, Darlene.
ACKNOWLEDGMENTS

To Doctor Robert J. Hanny, I extend my sincere appreciation for the generous assistance and support which he provided during my graduate studies and as advisor for this project.

To Doctor Robert Maidment, I offer my thanks for his nurturance in the classroom and invaluable service on my committee.

To Doctor Armand J. Galfo, I express appreciation for his assistance in the design of the research for this study and help on my committee.

To Mr. Frank Smith, I extend special thanks for the advice and help which he rendered in designing and preparing the data gathering instruments in this study.

To Lieutenant Colonel Thomas E. Coley, a particular debt of gratitude is owed for his diligent help and support in gathering test item and job performance data from U. S. Army units in the Continental United States.

To Colonel Winifred G. Williams, I extend my thanks for his support in gathering test item and job performance data from U. S. Army units in Germany.

To Colonel Robert C. Barrett, I am especially appreciative of his continual support and the help which he provided in the conversion of computer software for the computation of cutoff scores using the Angoff score-setting method and assessment of intragroup variability and interrater reliability significance.
CHAPTER I

INTRODUCTION

Justification for Study

Since the mid-sixties educators have been concerned with a relatively new method of assessing student achievement called criterion-referenced measurement. This new method measures individual student accomplishment based on a specific criterion or standard of performance, whereas the traditional assessment method, commonly referred to as norm-referenced measurement, assesses student achievement relative to the performance of others. Harasym (1981) affirms there are "many procedures . . . available for setting [criterion-referenced measurement] standards, but little is known of the extent to which different procedures yield consistent evaluation outcomes" (p. 725). Because traditional measures for estimating validity are based primarily on variability among scores, they are appropriate for norm-referenced measurement but not for criterion-referenced testing (Gronlund, 1985). According to Popham and Husek (1969), the standard for criterion-referenced measurement is usually expressed as a performance objective. Mehrens and Lehmann (1984) have pointed out there is no guaranteed match between test items and the instructional objectives, and the test could easily lack content validity. According to Scannel and Tracy (1975), because there are no specific bases for setting the ideal level or cutoff score
for a criterion-referenced test, this level usually has been set according to the expectations of the teacher or the school. Hambleton (1980) called for more empirical investigations into the factors to be considered in selecting the proper standard-setting procedure for a particular testing situation.

The United States Army Training and Doctrine Command and its associated schools have made extensive use of criterion-referenced measurement in their assessments of student performance in specialist-level training programs. Student performance standards have been embodied in the training objectives for each subject taught. The standards have been expressed as the minimum satisfactory level of performance for each subject in the training program.

The Army's confidence in the objective-based student education system has been documented in United States Continental Army Command (USCONARC) Pamphlet No. 310-14 (1966). In this pamphlet the Army declares that

Student performance objectives offer a number of advantages to the school. . . . They give the basis for preparing a valid test of the student's ability to perform on the job. . . . Since student performance objectives are derived directly from job tasks, it is obvious that valid examinations based on performance objectives will measure the student's ability to perform (p. 6).

This confidence has prevailed, even though a valid means to differentiate between the performance standards for the beginner and the experienced jobholder is not available. In the pamphlet cited above, it is also pointed out that without the necessary means "decisions must be based upon judgment growing out of experience and reasoning" (p. 31).
This conclusion is in consonance with the assertion by Popham (1981) that much work needed to be done in determining how to derive classes of measures which satisfactorily serve to sample the behavior and content domain delimited by a well-stated objective.

The United States Army Training and Doctrine Command (TRADOC), in TRADOC Regulation 350-2 (1979), has reiterated that "the Army requires that training be based on the performance of the precise tasks which the soldier will perform on the job" (p. 15). In this document the Army has mandated that "absolute standards based on job requirements be established and that the soldier's performance be evaluated against the standards" (p. 16). TRADOC Pamphlet 350-30 (1975) states:

Within-training standards should be based on reasonable expectations after given amounts of training. They should be low enough to be attainable and high enough to be an interesting challenge to the trainees, . . . The end-of-training standard should approximate the job standard (p. 47).

Swezey (1981) has noted that the basis for the criterion-referenced measurement model is to distinguish masters from non-masters in terms of their performance on tests. He called attention to the fact that assignment of an incompetent individual to a critical job could affect seriously the performance of the organization. He noted, too, that huge wastes of money, manpower, and time could occur if a master were erroneously labeled a non-master.

Annually the United States Army has trained thousands of soldiers in specialist-level training programs. In these courses the student's performance has been assessed using criterion-referenced measurement. Overtraining in these courses would constitute a waste of critical resource and deprive the unit of the soldier's on-the-job
support. Undertraining would also be detrimental. Many of the soldiers graduating from these programs have been assigned to positions of great responsibility, where less than minimal competency performance could impact seriously on their units' readiness and mission support capabilities. A means to ensure that the graduates from these programs are trained to the proper standards is essential.

There appears to be ample guidance in existing literature to determine test validity and interpret the significance of student performance using norm-referenced measurement. There is, however, a need to develop a means to assess test validity and measure examinee performance with respect to clearly defined criteria and a domain of behavior.

Statement of the Problem

A relatively new method for assessing student achievement, called criterion-referenced measurement, was developed to assess student performance based on a specific criterion or standard of performance and domain of behavior. Traditional measures for estimating test validity are based primarily on variability among student scores and are, therefore, appropriate for norm-referenced measurements but not for criterion-referenced testing (Gronlund, 1985). Because there are no specific bases for setting the ideal level or cutoff score for a criterion-referenced test, this level has usually been set according to the expectations of the teacher or the school. The need for a procedure for setting precise cutoff scores for criterion-referenced tests, as noted by Scannel and Tracy (1975), persists today. Harasym (1981) posited that even though there are procedures available for setting cutoff scores in criterion-referenced measurement, little is known about the
consistency of these procedures. He also pointed out that "the standard-setting procedure is a significant factor in the determination of the evaluation outcome" (p. 725). This study has examined a means to deal with these matters. It also provides answers to the following questions:

1. What are the level of intragroup variability and the difference in the interrater reliability estimates of instructors and job specialists in their assessment of the minimum passing score for a criterion-referenced classroom test?

2. What is the relationship of job effectiveness (supervisory ratings and skill qualification test scores) to classroom test cutoff scores for two groups of job performers whose criterion-referenced test cutoff scores were set by an empirically based method and the present method founded on instructor consensus?

Theoretical Rationale

This study is based on psychometric theory as it relates to validity assessment of norm-referenced and criterion-referenced measurements. Because traditional measures for estimating validity are based primarily on variability among scores of examinees, they are appropriate for norm-referenced measurement but not for criterion-referenced testing (Gronlund, 1985).

Cronbach (1971) has asserted that determining criterion-referenced test content validity is a matter of matching the test and performance objectives.
Payne (1974) has agreed that

. . . the quality of an item on a criterion-referenced test is a function of the degree to which it matches the objective. . . [providing that] panels of expert judges could be employed to help make these item-objectives evaluations (p. 295).

Swerey (1981) refined the test-objective relationship, noting that a criterion-referenced test must meet the following conditions in order to be valid:

A criterion-referenced test may be presumed to be content valid if all items are carefully derived from the required performance, conditions, and standards specified in the objectives and if the sample of test items appropriately represents the objectives (p. 149).

Rovinelli and Hambleton (1976) have advocated the use of test item-objective relationship to determine the test cutoff score based on an index of item-objective congruence. They endorse the use of content specialists, noting that the specialists could complete their ratings with a high degree of reliability and validity, as evidenced by the high level of item-objective congruence achieved.

Millman and Popham (1974) limited the scope of the test-objective relationship, discerning that "the population of observations originally prescribed for criterion-referenced tests is a domain of items and the responses of a single individual to them" (p. 137).

Popham and Musek (1969) pointed out that the validation of a criterion-referenced test is "a carefully made judgment based on the test's apparent relevance to the behaviors legitimately inferable from those delimited by the criterion" (p. 6).

Nunnally (1967), on the other hand, has posited that a valid criterion test
... must stand by itself as an adequate measure of what it is supposed to measure. Validity cannot be determined by correlating the test with a criterion because the test itself is the criterion of performance (p. 80).

Swezey (1981) has asserted that setting of cutoff scores for criterion-referenced tests is a difficult matter. "Practical considerations, ... scoring feasibility, and objective criticality must all be considered in establishing reasonable cutoff scores" (p. 134).

Zieky (1973) also has noted the difficulty in setting the criterion of performance, or the cutoff score, recommending the utilization of a group of performers considered to be minimally competent to set cutoff scores; the mean score of the group would constitute the desired standard.

Klein and Kosecoff (1973) categorized the types of scores that criterion-referenced tests could provide:

1. Number or percent[age] of items correct for a given objective or set of items that encompass a few highly related objectives.
2. Mastery of a given objective or set of items where mastery is defined in terms of a certain level of performance such as 90 [percent] correct.
3. The time it takes for an individual to achieve a given performance level.
4. The time ... to perform a certain task or set of tasks related to an objective.
5. The percent[age] [of examinees] who pass each item (p. 8).

Four theorists, Nedelsky (1954), Angoff (1971), Ebel (1972), and Jaeger (1978), provided methodologies for determining content validity and establishing test cutoff scores for performance measurement using content judges. The Angoff, Ebel, and Nedelsky methods proposed
that a judge or group of judges estimate the probable score of a hypothetical examinee who was capable of minimal acceptable performance. The Jaeger method proposed that a panel of judges indicate whether a person seeking certification in a particular field should be able to answer each question on the test. These methods have produced varying levels of acceptability and success. According to Van Der Linden (1982), "the Angoff technique . . . has become one of the best known and widely used methods of standard setting" (p. 296).

**Definition of Terms**

The relevance and application of criterion-referenced measurement to training have been hindered in part due to a lack of a clear understanding of the concept of training to a criterion or standard. The terms in this study whose definitions clarify the concept of criterion-referenced measurement are set forth below:

*Angoff "Accepts."* Soldiers whose classroom test scores met or exceeded the test cutoff score set by the Angoff score-setting method (formulated during the study).

*Angoff "Rejects."* Soldiers whose classroom test scores were lower than the test cutoff score set by the Angoff score-setting method (formulated during the study).

**Content Validity.** The adequacy of items in a test to sample the domain about which inferences are to be made (Mehrens & Lehmann, 1984, p. 19).

**Criterion.** "The standard by which something is measured... In evaluation, it is the measure used to determine the adequacy of a product, process, or behavior" (TRADOC Circular 351-84-1,
1984, p. 17). "A criterion is a standard or level of performance above which one is regarded as having passed and below which one is considered to have failed" (Hills, 1981, p. 446).

**Criterion-Referenced Measure.** That measure "used to ascertain an individual's status with respect to some criterion; i.e., performance standard" (Popham & Musek, 1969, p. 2).

**Criterion-Referenced Test.** A test "that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971, p. 653).

**Critical Task.** "A task which is essential for accomplishment of the unit mission, successful individual skill performance, and survival in combat, and requires training" (TRADOC Reg 351-2, 1986, Glossary 1).

**Cutoff Score.** "The score above which people are placed into one classification and below which they are placed in another such as passing above and falling below. It is sometimes called the criterion" (Hills, 1981, p. 446).

**Domain-Referenced Testing.** "The measurement of an examinee's performance that is referenced to a defined structured set or domain of learner behavior" (Hively, Maxwell, Robehr, Sension, & Lundin, 1973, p. 15).

**Instructional Validity.** "The correspondence between the test and what has been taught" (Yallow & Popham, 1983, p. 10).

**Job Performer.** A prescribed load list (PLL) clerk whose performance is assessed for the purpose of estimating the predictive
validity of a criterion-referenced classroom test with cutoff score set by the conventional and Angoff-based methods (formulated during this study).

Job Specialist. A prescribed load list (PLL) clerk whose role is to serve as a test item judge (formulated during this study).

Master. A soldier whose job performance meets the commander's/supervisor's expectations (formulated during this study).

Military Occupational Specialty (MOS). "A term used to identify a grouping of duty positions possessing such close occupational or functional relationship that an optimal degree of interchangeability among persons so classified exists at any given level of skill" (Army Regulation (AR) 310-25, 1975, p. 169).

Minimally Competent Individual. A hypothetical individual who has the minimum amount of knowledge to complete an approved [specialist-level] training program and perform effectively in his or her chosen occupational specialty (Cross, 1984, p. 31).

Non-Master. A soldier whose job performance does not meet the commander's/supervisor's expectations (formulated during this study).

Norm-Referenced Measure. That measure "which is used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device" (Popham & Husek, 1969, p. 2).

Objective-Based Test. A test "whose items have been constructed to measure an instructional objective" (Popham, 1981, p. 29).
Skill Qualification Test. "A major diagnostic tool for the individual training system... It contains written performance-oriented tests on a representative sample of critical tasks..." (TRADOC Reg 351-2, 1986, Glossary 2).

Specialist-Level Training Program (Advanced Individual Training). "Training given to enlisted personnel subsequent to completion of basic training, so as to render them qualified for the award of a military occupational specialty" (AR 310-25, 1975, p. 10).

Research Hypotheses

The hypothesized relationships presented in this study concern two important measures in estimating test validity and establishing a test cutoff score. These measures are intragroup variability and interrater reliability of instructors' and job specialists' probability estimates and the relationships between test cutoff scores and examinees' job effectiveness. The directional hypotheses for this study are:

Hypothesis 1 -- Intragroup variability in probability estimates by instructors and job specialists concerning validity of a criterion-referenced classroom test will differ among panel members; and the interrater reliability of the probability estimates of instructors and job specialists will not equal zero.

Hypothesis 2 -- Supervisory rating scores will be higher for job performers whose classroom test cutoff scores were set by the Angoff method than by the present method.
Hypothesis 3 -- Skill qualification test scores will be higher for job performers whose classroom test cutoff scores were set by the Angoff method than by the present method.

Hypothesis 4 -- Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) will receive higher supervisory rating scores than Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard).

Hypothesis 5 -- Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) will achieve higher skill qualification test scores than the Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard).

Intervention, Sample Description, and General Data Gathering Procedures

This study investigated the standard-setting method reported by Angoff (1971) for use by the United States Army Quartermaster School. The method was modified as it was for use in the Validation Study of the National Teacher Examinations for Certification of Entry-level Teachers in the State of Virginia (Cross, 1982). Specifically, the study examined empirical means for determining validity and reliability and setting cutoff scores for tests utilized in the Equipment Records and Parts Specialist training program. The scores represent the minimal
amount of knowledge that the soldier must possess to be awarded a particular military occupational specialty (MOS). The study was designed to accomplish the following:

Assess the correspondence between course curriculum and test content.

Assess, empirically, job relevance of the test and make reviewer estimates of the probability that a minimally competent examinee would answer each test question correctly.

Develop, empirically, cutoff scores based on probability ratings.

Assess, empirically, the significance of the modified Angoff method compared to the present method for setting cutoff scores for a specialist-level military training program.

The participants for this study comprised eight samples. The first sample consisted of test item judges drawn from the instructor staff for the Equipment Records and Parts Specialist Course, often referred to as the Prescribed Load List (PLL) Clerk Course, taught at the United States Army Quartermaster School. The second sample, which also consisted of test item judges, was drawn from PLL clerks (job specialists) in United States Army units located in Germany and the Continental United States. The third, fourth, fifth, sixth, seventh, and eighth samples were also drawn from practicing PLL clerks (job performers) for utilization in the assessment of the predictive validity of the end-of-course classroom test for the PLL Clerk Course.
Before data gathering began, a group of eight randomly selected faculty members reviewed the test items and the course content to determine if the universe of behavior sampled by the test is adequate. This assessment was expressed in terms of whether each test topic in job performance is given the same, more, or less emphasis in the course curriculum.

The data gathering procedure for this study involved instructors and job specialists in the analysis of the end-of-course criterion-referenced test for the PLL clerk training program. The analysis entailed determining the job relevance of each test item and making probability assessments that a PLL clerk would answer each question correctly. Assessments were used to establish relevance ratings and probability estimates by item and by rater. These ratings and assessments were summed to provide composite relevance and probability scores for use in setting the test cutoff score. The probability scores were also used to assess intragroup variability and interrater reliability of the first and second samples.

The data gathering procedure also included the collection of information relevant to job effectiveness of job performers who graduated from the PLL Clerk Course. These data relate to supervisory ratings and skill qualification test scores for two groups of job performers from the course, one whose classroom success was based on the cutoff score set by the Angoff method, and the other based on the present method.
Limitations of the Study

The limitations of this study concern three areas. These are the test item judges' possible lack of understanding of the concept of job performer minimal competency, the lack of control for maximum time on the job of performers whose job skills were rated, and the limited direct application of the study findings and recommendations to other courses taught at the United States Army Quartermaster School.

The concept of minimal competency was examined in a practice exercise involving the instructors and job specialists. As judges they determined test item relevance and made probability estimates prior to their involvement in the actual study. For the practice exercise, normative data indicating the number of students who answered each question correctly during the previous six months at the United States Army Quartermaster School were provided. The information assisted each reviewer in judging the stringency of the probability assessments they had rendered, based on the students' performance and the reviewer's own experiences. The practice exercise also involved a discussion among the judges wherein the stringency of their probability estimates was reviewed and a functional definition of minimal competency was derived. The study did not embody an empirical means to assess the level of understanding of the concept of minimal competency on the part of the test item judges, however.

Control by this study for the level of job experience on the part of the job specialists (test item judges) and job performers was limited to two criteria. First, each job specialist and job performer must have completed the Equipment Records and Parts Specialist Course
successfully by passing the end-of-course test for which the passing score was set using the Angoff cutoff score-setting method. Second, each job specialist and job performer must have served on the job as a PLL clerk at a satisfactory or above level of performance for at least three months prior to their involvement in the study. An upper limit concerning the length of job assignment which would have minimized variability in supervisors' ratings of job performer effectiveness attributable to job experience was not established.

The findings and recommendations of this study concerning test content validity and cutoff score cannot be generalized directly to other courses taught at the United States Army Quartermaster School. They are applicable only to the Equipment Records and Parts Specialist Course. Initially, a new cutoff score will affect the training of approximately 2,500 students annually. Ultimately the change could affect the performance of approximately 7,000 soldiers worldwide. The findings and recommendations concerning the use of the Angoff methodology can be generalized to many Army specialist-level courses which employ criterion-referenced measurement.

**Ethical Considerations**

This study was approved by the Human Subjects Research Committee of the College of William and Mary. Each study participant was provided a reasonable explanation of his/her proposed role. Informed consent, with emphasis on voluntary participation and expressed approval to disclose, anonymously, confidential data collected, was obtained from each
participant. Sample participant consent forms are provided as Appendices A and B, pages 143 and 144, respectively. Timely information concerning the study outcome was offered to each participant; no one requested information on the completed study.
A review of the literature and research related to this study is presented in this chapter. The review is divided into three summaries: rationale, relevant research, and previous research.

**Summary of Rationale and Its Relationship to Problem**

Today there are basically two measurement methods for assessing student achievement: norm-referenced measurement and criterion-referenced measurement. Norm-referenced measurement, the traditional method for evaluating student performance, assesses individual achievement based on the performance of others (Gronlund, 1985). Criterion-referenced measurement, on the other hand, assesses student performance relative to the achievement of criterion (performance standard) in a well-defined domain of behavior (Mehrens & Lehmann, 1984).

The psychometric characteristics of a norm-referenced test are based on the variability of student scores. This is not an appropriate method for determining the validity of a criterion-referenced test where test results should cluster around criterion (Gronlund, 1985).

In criterion-referenced measurement, the criterion or performance standard is usually set through teacher consensus (Scannel & Tracy, 1975). Although there are various methods which had been developed for determining test content validity and cutoff scores for
criterion-referenced tests, "little is known of the extent to which different procedures yield consistent . . . outcomes" (Harasym, 1981, p. 725).

There is also significant educator misunderstanding of the relationships of criterion and performance domains, the establishment of test cutoff scores, and the means to determine content validity of criterion-referenced tests (Mehrens & Lehmann, 1984). These conditions may have been thwarting progress in developing a means to assess, empirically, the validity and ideal cutoff scores for criterion-referenced tests.

In addition to these shortcomings, "in recent years, there has also been an increasing use of criterion-referenced exercises . . . Accountability and minimum-competency testing . . . require criterion-referenced measures" (Gronlund, 1985, p. 475). Existing literature is adequate in its treatment of norm-referenced testing and the inferences one could draw from using test norms. A serious need exists, however, to expand the psychometric treatment of test content validity as it pertains to criterion-referenced measurement.

**Summary of Relevant Research**

**Standardized Objective Tests**

The United States Army began in 1814 to use examinations to test surgeons and prospective military academy trainees. It was not until late in the nineteenth century, however, that the use of standardized objective tests in the federal bureaucracy began to grow (Zeidner & Drucker, 1983).
Psychological testing as a vital facet in developing a fighting force was officially begun in the Medical Department of the Army in 1917 (Zeidner & Drucker, 1983). Under the chairmanship of Major Robert M. Yerkes, a plan calling for the development of three kinds of tests was developed. These tests were:

... the written Army Alpha for literate recruits, the pictorial Army Beta for illiterates and for men who had failed the Alpha, and individual examinations—some version of the Binet scales—for failure on the Army Beta" (p. 11).

From the administration of these tests, scores were made available to commanding officers for use in assigning individuals to specific duties and for discharging mentally incompetent recruits from the Army. Opinions concerning the value of the tests varied within the Army. Some officials were outright hostile toward the use of tests; some treated the test personnel politely and then ignored them; and others became suspicious of the intent and launched three independent investigations of the testing program.

By January 31, 1919, Lieutenant Colonel Yerkes had presided over the administration of tests to 1,726,966 men, including 42,000 officers. More than 83,500 of the enlisted men had been given individual examinations in addition to the group test. Yerkes' other accomplishments included the development of the Army grading system to represent a coarse grouping according to ability to learn, oral trade tests to assess job knowledge, job specifications, an officers' [sic] qualification card, and other standard forms and procedures .... (p. 12).

Yerkes' final evaluation of the program was that "psychology [mental testing] had helped to win the war. ... it has incidentally established itself among the other sciences and demonstrated its right to serious consideration in ... human engineering" (Kelves, 1968, p. 581).
Interest in personnel testing increased because of World War II. Problems growing out of the sudden influx of men into the Army and the extraordinary demands to classify them for training and assignments had to be solved. New tests were needed to deal with this situation.

A new general classification test which could also be given to non-English speaking men and illiterates was introduced in October 1940. This test, the Army General Classification Test (AGCT), was quite successful in designating men for specialist training. "From 1940 to 1945 . . . an average of 4,000 persons [was] given the test daily" (Zeidner & Drucker, 1983, p. 35). Broad acceptance of test scores was in sharp contrast to the experiences with the Alpha/Beta tests during World War I. "More than nine million men took one form or another of the AGCT by the war's end" (p. 21).

Since the fifties there has been continuing improvement to the test battery for the selection and classification of military personnel, including predictive measures for successful combat performance. One of the major changes included the development of the Armed Services Vocational Battery (ASVAB), a group of twelve tests which was combined into various composites for measuring soldier aptitude and academic ability. The rapid modernization of the Army and changing doctrine and tactics have dictated continuing exploration and improvement of personnel testing procedures (Zeidner & Drucker, 1983).

Criterion-Referenced Measurement

For more than two decades educators had heard about a new approach to testing, labeled criterion-referenced measurement. The concept stemmed from the impact that improved instructional techniques were
having on traditional measurement tactics (Popham, 1981). Criterion-referenced assessment became popular along with the programmed learning movement in the fifties. It was utilized principally in the criterion test that was developed for use in the tryout and revision of programmed materials. According to Rountree (1977), its use virtually ensured mastery of the lesson by similar students in the future.

The term criterion-referenced measurement was first used in 1962 (Glaser & Klaus, 1962). It was not until 1963 that Glaser introduced the concept of norm-referenced and criterion-referenced measurements. In a provocative essay he drew a distinction between tests which compared student performance to norms based on the performance of other students and tests which measured a student's absolute performance in a specific behavioral domain.

Measurement of student performance took different forms. Haak (1960) recommended the systematic adjustment of the student's summative grade by applying an adjustment factor for deciding on the average grade. Hinley (1968) offered a handicapping formula for crediting the student with progress made, calculated as part of the student's last test score.

In the late sixties Hively, Patterson, and Page (1968), at the University of Minnesota, developed a method for defining curricular content in science and mathematics. Their scheme, called item form, consisted of a detailed set of specifications for limiting the form of items used for measuring a particular skill. Hively and others (1973) labeled their approach domain-referenced testing because the examinee's performance concerned a defined domain of learner behaviors. Goodman
(1971), on the other hand, suggested that normative assessment should be used because educators were not confident concerning the criteria for measuring student performance. In the interest of measuring student progress, Cox (1971) suggested an index of discrimination based on the relative number of students passing an item at the beginning and conclusion of a unit of instruction. Rountree (1977), however, contended that progress was sometimes not enough, that in those instances where the student must acquire a prerequisite skill, criterion-referenced measurement was necessary.

According to Popham (1981), interest in criterion-referenced measurement virtually exploded in the early seventies. Avant garde educators applied the term criterion-referenced to nearly every unescorted noun in the education-related lexicon, including, some authors say, criterion-referenced references. This small vocal group, the proponents for a new order, proclaimed: "Criterion-referenced tests ... are education's only salvation" (p. 25). Despite this enthusiasm, the reaction in general to criterion-referenced measurement was mixed. Minor controversy therefore arose. For example, Block (1971) suggested that criterion-referenced measures were well suited to the classroom; on the other hand, Ebel (1971) argued that criterion-referenced measures had limited application to day-to-day classroom activity. While authors' opinions differed, part of the controversy may have been attributable to a lack of understanding of the concept of criterion-referenced measurement.
In Glaser's 1963 article two meanings of the term criterion unfortunately were introduced. One implied definition connoted performance in a defined behavioral domain; the other noted performance at a given level of proficiency. Popham and Husek (1969) defined criterion-referenced measurement as one that identified individual achievement relative to an established standard of performance. Simon (1969) embellished the definition, stating that the standard was expressed as a behavioral objective which described anticipated student achievement under conditions and criteria in which the achievement could be demonstrated.

Glaser and Nitko (1971) defined a criterion-referenced test as "one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Popham (1981) altered his 1969 opinion that criterion meant performance standard, arguing now that

... to interpret criterion as a level of examinee proficiency yields almost no dividends over traditional testing practices. In fact, by using that concept of criterion, one could magically transform any norm-referenced test into a criterion-referenced test merely by setting a specific proficiency level for the test (p. 28).

Payne (1974) espoused the belief that the major difference between norm-referenced measurement and criterion-referenced measurement was in the interpretation of the scores. He noted that it is possible for a criterion-referenced measure to be used in a norm-referenced way, but unlikely that the converse is true.

Rountree (1977) advocated an expansion of the concepts of norm-referenced measurement and criterion-referenced measurement to include
self-referenced assessment, or how well one has done in comparison to oneself. The enlargement of the concept of testing did not detract from the relevance of norm-referenced and criterion-referenced assessments, according to Rountree. He acknowledged that criterion-referenced assessment is vital in sequential teaching, where a prerequisite skill has to be attained before the student can progress.

Popham (1971) asserted there were certain tasks which must be performed to a very high level. To the extent that the performance is critical, the test on such instruction should be criterion-referenced.

Gronlund (1985), in his discussion on the validity of criterion-referenced mastery tests, has provided a useful insight into the use of criterion-referenced test scores for drawing inferences concerning student performance. Contending that the tests were typically used to evaluate pupil performance in terms of mastery, he postulated that criterion-referenced referred to the performance domain described by instructional objectives, while criterion-related described some second measure of performance on a specified domain of achievement.

Today most measurement experts agree on the distinction between norm-referenced and criterion-referenced test scores' interpretation. Disagreement persists, however, concerning the distinctions between the tests. There is a lack of consensus among writers in the criterion-referenced field that criterion-referenced test construction permits inferences from test question results to the entire domain. Some authors contend that test results represent a specified content parameter (Mehrens & Lehmann, 1984). Most authors agree that a change in name would help, that domain-referenced is a more appropriate term than
criterion-referenced, since it carries no implication of a cutoff score or standard (Popham, 1981). The term criterion-referenced persists, however.

Test Validity and Cutoff Scores

Many authors (Nedelsky, 1954; Angoff, 1971; Ebel, 1972; Zieky, 1973; Payne, 1974; Buck, 1975; Rovinelli & Hambleton, 1976; and Popham, 1981) called for the use of expert judgment in validating test items and setting cutoff scores.

Most of the standard-setting methods recommended for criterion-referenced measurement have been designed in relation to minimal competency, or the minimum acceptable level of proficiency on the part of the examinee (Andrews & Hecht, 1976). Glass (1978) rejected the concept of minimal competence. "The idea of minimal competence," he asserted, "is bad logic and even worse psychology" (p. 251). "Those who claim the ability to make the determination of mastery or competence in statistical or psychological ways . . . can't" (p. 237), he posited. Contemporary educators do not have a way to specify the minimum acceptance level of performance for a criterion-referenced test. "They cannot determine criterion levels or standards other than arbitrarily" (p. 237). Popham (1978) responded that arbitrary need not mean capricious. Glass (1978) also asserted that "setting performance standards on tests and exercises by known methods is a waste of time" (p. 259). Burton (1978) agreed that the practice of setting performance standards had little social decision-making value. Otherwise, the arguments against standard setting presented by Glass have gained little support.
According to Saunders, Ryan, and Huynh (1981), the three best known content-based procedures for determining passing scores are those proposed by Nedelsky, Angoff, and Ebel. These procedures require a judge or group of judges to estimate the probable score of a hypothetical examinee who was capable of minimum acceptable performance. A summary of each of these procedures, along with a fourth procedure developed by Jaeger (1978), follows.

Nedelsky (1954) presented a technique for determining the minimum passing score for a multiple-choice type test. He posited that:

The passing score is to be based on the instructor's judgment of what constitutes an adequate achievement on the part of the student and not on the performance by the student relative to his class or to any other particular group of students (p. 3).

In the Nedelsky (1954) score-setting process each panelist judges each distractor of each test item. The panelist eliminates those distractors which a student who is performing at the minimum passing level should be able to reject as being incorrect. The estimated probability that an examinee at this level would answer the question correctly is the reciprocal of the remaining choices. The sum of these estimated probabilities is the panelist's estimated passing score for the test. The mean of the panelists' estimates constitutes the recommended cutoff score (p. 5).

Angoff (1971) also recommended a procedure which calls for the use of content specialists for setting the test standard:

This procedure . . . ask[s] each judge to state the probability that the "minimally acceptable person" would answer each question correctly. In effect, the judges would think
of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score (p. 515).

This standard-setting method can be used for multiple-choice and other types of written tests.

Ebel (1972) proposed a method for setting the cutoff score which used certified practitioners:

It requires the certification test to be given to a large and representative sample of professionals who have been approved to practice and who are actually practicing. A decision is made by the certifying authority that any applicant who scores above the lowest quarter, lowest fifth, or lowest tenth of those actually practicing the profession deserves to be certified to practice it also (p. 495).

The passing score set by the practitioner tended to be set based on the supply of applicants.

Jaeger (1973) also proposed a method for determining test content validity and setting test cutoff scores using a large panel of judges representing several interest groups. The judgments rendered provide a yes or no answer to this question: Should every person seeking certification in this field be able to answer this question correctly? Response data are recorded on separate multiple indices for the yes/no responses to test item relevance and instructional validity.

The judgment process calls for each panelist to review each test item three times. On the first occasion the panelist provides an appropriate yes or no response without the benefit of any normative data. On the second review the panelist is provided the performance standard computed for each panelist on the first review and information on the percentage of examinees who answered each test item correctly on a previous
administration of the test. For the third review panelists are taught how to estimate the percentage of examinees who would fail to achieve the performance standard they had set previously.

The ultimate test performance standard is computed as the average of the yes answers provided by the reviewers.

Summary of Previous Research and Its Relationship to Problem

Previous research on setting minimum standards for tests concerned principally the methods developed by Angoff, Ebel, Jaeger, and Nedelsky. The majority of the past research was performed by the Educational Testing Service (ETS) in determining the validity of the standards for the National Teacher Examinations, using the Angoff method. The summary of research identified in this study pertains to the application of the procedures developed by each of the authors.

Andrews and Hecht (1976) investigated the Ebel and Nedelsky standard-setting methods. The purpose was to

... estimate their comparability in terms of the pass-fail level that would be generated by each, as well as their stability in terms of each procedure's capacity to generate similar pass-fail levels for the same test content when used by different groups of judges (p. 46).

The methods were compared using 180 multiple-choice questions from a nationally administered health professions' examination. Two groups of judges were drawn from committees that had provided test items for the certifying examination. The Ebel method was used with the even numbered items; the Nedelsky method was utilized with the odd numbered questions. Both groups reviewed the odd and the even halves of the test, but the review order was reversed for each group.
From the study it has been concluded that the different groups of judges set similar standards for the same test, but the two procedures yielded significantly different recommended passing scores. Using the Ebel procedure, the group A standard called for the student to answer correctly 68.8 percent of the questions; group B standard called for 68.0 percent. Using the Nedelsky procedure, the group A standard was 50.3 percent; the group B standard was 53.7 percent. The authors concluded that the difference in passing scores derived for the two standard-setting procedures "suggests that the specific techniques employed in setting examination standard may be a more powerful determinant of that standard than another variable" (p. 49).

Foggio, Glasnap, and Eros (1981) compared the Angoff, Ebel, and Nedelsky standard-setting methods to examine the psychometric properties of each, including the reliabilities and validities of judges' ratings. Comparison of these methods was made in setting performance standards for ten reading and mathematics tests for grade levels 2, 4, 6, 8, and 11 in the state of Kansas public schools. The number of judges per panel was 24 for the Ebel method and 31 for the Angoff procedure. The standard-setting methods were randomly assigned across the Kansas school districts, with the following reported findings:

The performance score standards ranked in order (high to low) are: Ebel, Angoff, and Nedelsky (p. 17).

The Angoff method produced greater variability in ratings among the tests, with reliability coefficients ranging from .96 to .99; the Ebel composite ratings ranged from .97 to .99; the Nedelsky composite ratings also ranged from .97 to .99 (p. 24).

Using the means as the standard, the Ebel method consistently resulted in the highest score (except in two cases) (p. 23).
Using item difficulties based on student performance as external criteria, the correlations with item ratings for each procedure produced moderate coefficients in the .40 to .70 range (p. 17).

The Nedelsky minimum passing score over tests [results] in the highest agreement coefficient... The lowest coefficient results with the Ebel method (p. 14).

Cross, Impara, Frary, and Jaeger (1984) describe a pilot study conducted in 1982 to compare the Jaeger, Nedelsky, and Angoff standards-setting methods. The study was designed to determine the validity of using scores from the National Teacher Examinations (NTE) for initial certification decisions in Virginia and to establish minimum standards for those tests determined to be valid (p. 113).

The methods were compared using two of the NTE area examinations: mathematics and elementary education. Thirty faculty members, drawn from seven teacher-training institutions in Virginia, participated as judges. These reviewers were randomly assigned to one of the three panels. The number of members from the same institutions was balanced across panels. The following conclusions were drawn from the study:

The empirical failure rates [due to the resulting minimum passing score] associated with the Jaeger method are sufficiently high that it is difficult to defend the use of this method for setting standards on the NTE area examination... The psychometric characteristics of the Nedelsky judgments argue against using this method. The judgments tended to be substantially less reliable, and have lower correlations with item difficulty values and item relevance ratings than the others... The psychometric properties of the Angoff judgments were unsurpassed (p. 126). [The judges] using the Angoff method expressed greater confidence in their knowledge estimates and greater confidence in the standards resulting from these judgments, on the average, than those using the other methods (p. 127).

Cross (1982) reported on a two-phase study which was conducted to validate the National Teacher Examinations (NTE) for certification
of entry-level teachers in the state of Virginia. There were two major goals for the study:

. . . to determine the validity of using the NTE examinations,
. . . [and] to establish cut-scores that represent minimal amounts of knowledge prospective teachers must demonstrate for initial certification (p. 1).

Phase I of the study concerned seventeen of the eighteen area examinations pertaining to the various teaching specialties; Phase II focused on the core battery of the NTE, plus a newly revised area examination on speech/communication.

The validation was executed using the Angoff standard-setting method, modified to include the use of normative data to help the judges evaluate the stringency of the standards they were setting. Two panels of judges were utilized for the area examinations, consisting of teacher educators and practicing teachers. Only teacher educators judged the core battery tests. Each judge rendered test content and instructional validity estimates, test item relevance ratings, and probability estimates that a minimally competent person would know the answer to each question without resorting to guessing. The judges also rendered several holistic judgments relative to the validity of using the tests to certify entry-level teachers in Virginia.

Validity estimates for the tests were characterized as either strong, substantial, moderate, or weak. Instructional validity estimates relative to the core battery tests ranged from strong to substantial; instructional validity estimates for the area examinations ranged from strong to moderate. Derived standards were computed for each test. The standards were stringent.
A downward adjustment to the derived standards was recommended to minimize the risk that a competent examinee would be declared incompetent. The recommended adjustment was equivalent to two standard errors, the composite error that takes into account measurement instability and reviewer sampling error.

From this study it could be concluded that tests with cutoff scores set by the Angoff method could be used to measure accurately and reliably minimal competency of the examinee.

Jaeger and Busch (1984) reported on the use of the Jaeger standard-setting method for determining prerequisite test scores for prospective teacher education students. The objective of the study was to determine the validity of and appropriate standards for Core Battery Tests I (General Knowledge) and II (Communication Skills) of the National Teacher Examinations (NTE) "for screening applicants for admission to North Carolina's teacher education program" (p. 3). The students were required to pass the tests prior to the end of their junior year of college to be accepted into an approved teacher education program.

Two groups of judges were drawn from the population of North Carolina public school teachers and college and university faculty members. Two types of validity evidence were sought. The first type concerned the relevance of the skills and knowledges assessed by the tests. The second type concerned the probability that most prospective teachers would have had a reasonable opportunity to acquire the skills and knowledges involved.
From this study it was concluded that the level of relevance of the Social Studies and Literature and Fine Arts subtests of the General Knowledge Test was acceptable; the Mathematics and Science subtests levels were satisfactory and marginal, respectively. The level of relevance of the four subtests (Reading, Writing, Listening, and Essay Writing) of the Communication Skills Test was also acceptable.

The instructional validity judgments concerning the General Knowledge Test were mixed, with additional investigation of the Social Science subtest deemed essential. The instructional validity of the Communication Skills Test was judged to be satisfactory.

The median passing scores recommended by the judges, with adjustments for error measurement, were proposed for implementation. The study determined, however, that "continued use of the tests demands an ongoing program of review and research" (p. 22).

From these studies it has been concluded that reliable examination standards could be set by employing the judgments of practitioners. These judgments are subjective, and the evaluation outcome is dependent upon the standard-setting procedure used. To minimize subjectivity, standard-setting studies must be designed so that the judgments are informed and reasoned.
CHAPTER III

METODOLOGY

Problem

A relatively new method for assessing student achievement, called criterion-referenced measurement, was developed to assess student performance based on a specific criterion or standard of performance and domain of behavior. Traditional measures for estimating test validity are based primarily on variability among student scores and are, therefore, appropriate for norm-referenced measurement but not for criterion-referenced testing (Gronlund, 1985). Because there are no specific bases for setting the ideal level of cutoff score for a criterion-referenced test, this level has usually been set according to the expectations of the teacher or the school (Scannel & Tracy, 1975).

This study investigated the use of the Angoff method for setting cutoff scores for criterion-referenced tests used in specialist-level military training programs. It assessed test instructional and content validity and used job relevance and probability estimates in setting the test cutoff score. The study further assessed, empirically, the predictive validity of the classroom test with cutoff scores based on the Angoff and traditional score-setting methods by comparing student classroom test scores with actual job performance. This chapter presents a
discussion of the study methodology, to include specific null hypotheses, populations and samples, procedures, instrumentation, research design, statistical analysis technique, and summary of methodology.

Specific Null Hypotheses

The following null hypotheses, which concerned the significance of intragroup variability of probability estimates, interrater reliability of probability judgments, and classroom test predictive validity, have been tested in this study:

Hypothesis 1 — There will be no significant difference in intragroup variability of the probability estimates by instructors and job specialists in their assessment of the minimum passing score for a criterion-referenced classroom test; and the interrater reliability of instructors' and job specialists' probability judgments will not be statistically significant.

Hypothesis 2 — There will be no difference in the correlation between classroom test scores and supervisory rating scores of job performers whose classroom test cutoff scores were set by the Angoff method and the present method.

Hypothesis 3 — There will be no difference in the correlation between classroom test scores and skill qualification test scores of job performers whose classroom test cutoff scores were set by the Angoff method and the present method.
Hypothesis 4 -- There will be no difference in the correlation between the classroom test scores and supervisory rating scores received by Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) and Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard); and there will be no significant difference in the means of supervisory rating scores received by Angoff "accepts" and Angoff "rejects."

Hypothesis 5 -- There will be no difference in the correlation between the classroom test scores and skill qualification test scores received by Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) and Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard); and there will be no significant difference in the means of skill qualification test scores received by Angoff "accepts" and Angoff "rejects."

Populations and Samples

There were three populations from which the participants for this study were drawn. The first one was the instructor staff for the Equipment Records and Parts Specialist Course, often referred to as the Prescribed Load List (PLL) Clerk Course, taught at the United States
Army Quartermaster School. These personnel were noncommissioned officers who prepare the instructional materials for and teach the PLL Clerk Course. The second population was the PLL clerks from United States Army units located in Germany and the Continental United States. These personnel were specialist-level job performers who completed the PLL Clerk Course at the United States Army Quartermaster School and whose job performance was at least satisfactory. These soldiers had been assigned to a PLL clerk position for a minimum of three months when chosen to participate in this study. The third population was a group of Angoff "accepts" and "rejects," job performers whose classroom test scores exceeded, met, or fell below the Angoff-derived standard. These soldiers had also been assigned to the PLL clerk positions for a minimum of three months when chosen to participate.

Total population of PLL Clerk Course instructors at the United States Army Quartermaster School was 70. The approximate population of PLL clerks in three Army divisions was 240.

The study samples from these populations consisted of 36 instructors (test item judges); 36 PLL clerks (job specialists/test item judges); 64 PLL clerks (job performers) for the comparison of students' classroom test scores and supervisory rating scores for the Angoff score-setting method, and 100 PLL clerks (job performers) for the present method; 49 PLL clerks (job performers) for the comparison of students' classroom test scores and skill qualification test scores received for the Angoff score-setting method, and 66 PLL clerks (job performers) for the present method; 64 Angoff "accepts" [same as 64 PLL clerks above], and 36 Angoff "rejects" for the comparison of students'
classroom test scores and supervisory rating scores; and 49 Angoff "accepts" (same as 49 PLL clerks above), and 17 Angoff "rejects" for the comparison of students' classroom test scores and skill qualification test scores. The participants were selected randomly. The sample sizes were based on the following guidance and precedents:

Test Item Judges:

Poggio, Clasnap, and Eros (1981) used 41 judges for the Angoff standard-setting method; Cross (1984) employed 40 reviewers per test for the core battery and 17 reviewers per test for area tests of the National Teacher Examinations (NTE); and Cross, Impara, Frary, and Jaeger (1984) employed 15 judges per test to review elementary education and mathematics tests of the NTE.

Job Performers and Supervisors:

Borg and Gall (1983) noted that "in causal-comparative and experimental research, it is desirable to have a minimum of 15 cases in each group to be compared" (p. 257).

 Procedures

The data gathering methods for this study concerned the collection of information for four purposes. These were the assessment of test content validity, the estimation of test item relevance and probability values, the computation of test cutoff scores, and the assessment of the predictive validity of the classroom test with cutoff scores set by the Angoff method and the present method.
Assessment of content validity of the criterion-referenced test began with an evaluation of the test's sampling validity. This was accomplished by a group of eight randomly selected faculty members.

The first step in the process was to determine if the universe of behavior sampled by the test was adequate. The correspondence between course content and test content was determined by specifying the test topics to be measured and judging the emphasis given to each in the training program. This assessment was made by determining whether the emphasis given to each test topic was the same, more, or less in the course curriculum. The proportion of the respondents selecting each option was computed for each test topic (Helmstadter, 1964). These proportions were multiplied by the percent of the items in the test devoted to each task. By summing the products, an index rating from 0 to 100 was provided (Cross, 1982).

Prior to involving instructors and job specialists in setting test cutoff scores, the correct answer to each question was recorded on the test to minimize ego involvement among the reviewers. Normative data indicating the number of examinees who answered each question correctly during the previous six months at the United States Army Quartermaster School were also provided. This information assisted each reviewer in judging the stringency of the probability assessments being considered, based upon the performance of a large group of students in addition to the reviewer's own experiences.

A practice exercise in determining test item relevance and making probability assessments was conducted prior to involving the judges in the actual study. This exercise emphasized test review procedures.
It also involved a discussion wherein the judges reviewed the stringency of their probability assessments and came to an agreement on a functional definition of minimal competency. This procedure has been corroborated by Festinger (1954) and Schachter (1959). A copy of the test rating instructions, based on Cross (1982), given to each test review coordinator is provided at Appendix C, page 145.

Setting of test cutoff scores began by involving classroom instructors and job specialists in judging the job relevance of each test item. Relevance assessment was made by judging the subject matter content of each test item relative to the domain of knowledge that entry-level job performers should possess. A 1-to-4 point rating was used based on the following values: 1 = not relevant; 2 = questionable; 3 = important; and 4 = crucial. Each judge rendered a job relevance rating for each test item on a machine-readable form. Any item receiving an average relevance rating below 2.25 was, in turn, given a value of zero in the probability assessment. A value of 2.25 rather than 2.0 was used to prevent a single reviewer from tipping the scale (Cross, 1982).

The rendering of probability estimates also involved classroom instructors and job specialists. In this exercise each participant judged the probability that a minimally competent person would answer each test item correctly. A rating was recorded for each test item by each judge. These estimates were assessed using a 1-to-10 scale probability, as shown on the response scale: a probability rating of 1 = .0 through .10; 2 = .11 through .20; 3 = .21 through .30; ... 10 = .91
through 1.0. Because the probability ratings concerned criterion-referenced measurement, the judges were instructed to equate minimal competence with the high point of the intervals shown on the response scale. The judges' probability ratings used in computing the cutoff score therefore correspond to the upper limit of the intervals. A raw score performance standard, or cutoff score, was computed for each judge by summing the probability values assigned. The cutoff score for the test was determined by computing the mean of the cutoff score across individual judges.

The raw score performance standard was converted to a percentile standard. In turn, the potential failure rate predicated on the new cutoff score was computed. This calculation was based on test scores of course graduates from the Equipment Records and Parts Specialist Course during the year prior to the setting of the new performance standard.

A new cutoff score was not adopted without taking into consideration two potential sources of error: sampling error associated with selection of panelists and measurement error associated with examinee performance. Standard error scores were computed for each panel of judges. A composite standard error was calculated and used to determine potential failure rates. Allowance was made for one, two, or three standard errors in computing a recommended new cutoff score.

Estimating interrater reliability entailed the assessment of intragroup differences in panelists' judgments. The assessments provided variance estimates among reviewers from the instructor and job
specialist panels. The reliability ratings were calculated using the procedure described by Ebel (1951).

The predictive validity of the classroom test with cutoff scores set by the Angoff and the present cutoff score-setting methods was determined in two ways. The first assessment was made through statistical comparisons of supervisory rating scores, skill qualification test scores, and soldiers' classroom performance, based on the standards established by the two methods. The supervisory ratings were made using a behaviorally anchored rating scale. The skill qualification test scores were derived from formal Army records of soldier achievement on the test. As prescribed by Army Regulation 350-37 (1986), the skill qualification test was a performance-based set of multiple-choice questions used to evaluate soldier proficiency in a representative sample of critical job tasks. The test is administered to the individual job performer annually. Reliability of the skill qualification test for the P.I.I. clerk has not been assessed. The second predictive validity assessment was made by comparing Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) and Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard) with supervisory rating scores and skill qualification test scores.

The procedures employed in determining test validity and job performer effectiveness embodied stringent ethical safeguards. Test item relevance and probability estimates rendered by instructors and job specialists have been kept confidential. Information concerning the
method employed in setting the cutoff score used to measure the classroom success of each job specialist was withheld from the rating supervisor. The soldier's classroom test score also was withheld. The skill qualification test scores, obtained from Department of the Army records, were not relatable to an individual job performer by anyone other than the originator of this research.

**Instrumentation**

A Likert-like 1-to-4 scale has been used by instructors and job specialists to assess test item content relevance to the domain of knowledge that the entry-level job performer should possess. This machine-readable form is a 4-point scale used to assess these judgments of each rater: 1 = not relevant; 2 = questionable; 3 = important; and 4 = crucial. A relevance value of 2.25 was set for the purpose of dropping from the cutoff score computation any test item with a mean relevance rating of less than 2.25 assessed by a majority of the judges. This value was based on the works of Cross (1982, 1984) where "the value of 2.25 was used rather than 2.0 to ensure that a single rater could not tip the balance" (1982, p. 47).

A machine-readable, 10-point probability scale has been used to represent each judge's estimate of the chance that a minimally competent person would answer each test item correctly. A functional definition of a minimally competent person was derived through panelist discussion and consensus. A probability estimate was rendered for each item by each judge, except for those test items declared not relevant by the majority of the judges. Each 1-to-10 probability scale estimate was converted to a probability value, using an interval scale in which a
1 rating = .0 through .10; 2 = .11 through .20; . . . 10 = .91 through 1.0. The potential range of probability values was therefore .1 to 1.0. The sum of the rater's probability values for all test items represented the rater's expectation of performance of a minimally competent person on the test. The mean of the probability values across individual judges, corrected for standard error, constituted the recommended test cutoff score or performance standard. The individual probability judgments were also utilized to estimate intragroup variability and inter-rater reliability for judgments across items. The rating scale developed by Lawrence H. Cross, Virginia Polytechnic Institute and State University, for use in the Validation Study of the National Teacher Examinations for Certification of Entry-Level Teachers in the State of Virginia, was adapted to use in this study. A copy of the letter of authorization to utilize the rating scale is provided at Appendix D, page 152. A semblance of the rating scale used to record relevance and probability estimates by each judge in this study is provided at Appendix E, page 153.

A behaviorally anchored 7-point rating scale has been used to assess supervisor estimates of soldier job effectiveness. The Army-Wide Performance Rating Scales, Form 6A, provided for an assessment of ten specific categories and one general category of performance for first-term soldiers. These categories of performance are technical knowledge/skill, effort, following regulations and orders, integrity, leadership, maintaining assigned equipment, military appearance, physical fitness,
self-development, self-control, and overall effectiveness. The categories of performance utilized in this study covered technical knowledge/skill, effort, following regulations and orders, and maintaining assigned equipment. The content validity of this instrument has been assessed through workshops and instrument refinements involving persons regarded as experts in the performance areas to be measured. Reliability of this rating scale was established through a series of interrater reliability assessments concerning supervisor and peer ratings. Interrater reliability assessments of supervisor ratings involved approximately 1,400 personnel in eight military occupational specialties. The reliability coefficients for these assessments were: .76, .67, .59, .58, .55, .54, .42, and .42. Permission to use the Army-Wide Performance Rating Scales, Form 6A, in this study was granted by the United States Army Research Institute. Permission to publish the instrument in any form in this study was withheld, however.

Research Design

This study utilized the quasi-experimental design (Campbell & Stanley, 1963). The design accommodated the test for intragroup, across items variability and interrater reliability through the calculation of intragroup variance estimates. The tests for intragroup variability and interrater reliability concerned two groups of judges involved in making independent assessments of test item content relevance and probability estimates (knowledge assessments). In this study separate tests for significance of the correlation between supervisory rating scores and skill qualification test scores of two groups of job performers have been made. Also, the significance of the correlation
between the classroom test scores of Angoff "accepts" and "rejects" and their supervisory rating scores and skill qualification test scores was computed. In addition, t-tests were performed to determine if the differences between the means of supervisory rating scores and skill qualification test scores of the "accepts" and "rejects" populations were significant. Angoff "accepts" and "rejects" samples were drawn from the population of job performers whose levels of classroom performance were assessed using cutoff scores set by the Angoff method and the present method. Supervisory rating scores have been calculated using a behaviorally anchored rating scale. Skill qualification test scores were collected from Department of the Army records. It was not helpful to use the codes and graphics presentation utilized by Campbell and Stanley to describe the research design.

**Statistical Analysis Technique**

A two-way analysis of variance (items by raters) was used to assess intragroup variability concerning test probability estimates. A correlation coefficient (r) has been employed to assess interrater reliability of the judges' probability estimates using a procedure described by Ebel (1951). The independent variable was the test item probability assessment. The dependent variable was the probability score. The .05 level of significance was selected for these analyses.

A Pearson Product-Moment correlation coefficient (r) was used to analyze the relationship between soldier classroom test scores achieved under the Angoff and present cutoff score-setting methods and the supervisory rating scores received by these job performers. The analysis estimated the relationship across select job performance categories of
the supervisory rating scale. The measure of job performance relative to the Angoff method concerned the Angoff-derived test cutoff score through 100 percentage points; the measure of performance relative to the present method concerned the present test cutoff score through 100 percentage points. The independent variables were the cutoff score-setting methods and the cutoff scores which they yielded. The dependent variables were the supervisory rating scores of the two groups of job specialists. The .05 level of significance was selected for this analysis.

A Pearson Product-Moment correlation coefficient \( r \) was used to analyze the relationship between soldier classroom test scores achieved under the Angoff and present score-setting methods and the skill qualification test scores received by these job performers. The measure of performance relative to the Angoff method concerned the Angoff-derived test cutoff score through 100 percentage points; the measure of performance relative to the present method concerned the present test cutoff score through 100 percentage points. The independent variables were the cutoff score-setting methods and the cutoff scores which they yielded. The dependent variables were the skill qualification test scores of the two groups of job performers. The .05 level of significance was selected for this analysis.

The Pearson Product-Moment correlation coefficient \( r \) was used to analyze the relationship between the classroom test scores achieved by Angoff "accepts" and Angoff "rejects" and the supervisory rating score rendered on each performer. In addition, a \( t \)-test was performed to determine if the difference between the means of the supervisory
rating scores of the "accepts" and "rej ects" samples was significant. The independent variable was the job performer's classroom test score. The dependent variable was the supervisory rating score. The .05 level of significance was selected for this analysis.

The Pearson Product-Moment correlation coefficient (r) was used to analyze the relationship between the classroom test scores achieved by the Angoff "accepts" and Angoff "rej ects" and the skill qualification test score achieved by each performer. In addition, a t-test was performed to determine if the difference between the means of the skill qualification test scores of the "accepts" and "rej ects" samples was significant. The independent variable was the job performer's classroom test score. The dependent variable was the skill qualification test score. The .05 level of significance was selected for this analysis.

Summary of Methodology

This study has assessed a means to determine content validity and set cutoff scores for criterion-referenced tests using the empirically based Angoff cutoff score-setting method. It has tested the effectiveness of the Angoff method and the present method in setting performance standards for measuring classroom performance relative to a clearly defined criterion and domain of behavior. The study has exam-ined the predictive validity of the classroom test using cutoff scores set by the Angoff and present methods. It also has assessed the empirical relationship between examinee test scores achieved and job performer effectiveness.

There were three target populations: classroom instructors who assessed job relevance and probability estimates for use in setting test
cutoff scores, job specialists who also made job relevance and probability estimates, and job performers whose job effectiveness was rated relative to the cutoff score-setting methods employed. Scores derived from the Army Skill Qualification Test Program and the use of a behaviorally anchored rating scale, the Army-Wide Performance Rating Scales, Form 6A, were used to assess soldier job effectiveness.

A two-way analysis of variance and a correlation procedure by Ebel were employed to assess intragroup variability and interrater reliability of probability estimates; Pearson Product-Moment correlations and t-tests were employed to test the significance of the relationships among the two cutoff score-setting methods employed, examinee test scores, skill qualification test scores, and soldier performance.
CHAPTER IV

RESULTS

This chapter presents an assessment of content validity, intragroup variability calculation, interrater reliability estimation, cutoff score computation, and predictive validity measurement relative to the end-of-course test for the Equipment Records and Parts Specialist Course. The computation of the test cutoff score employs the Angoff score-setting method. The calculation of intragroup variability uses analysis of variance of test item judges' probability estimates. The estimation of interrater reliability makes use of a method proposed by Ebel (1951) which assesses reliability as a function of the variability across test items and among judges. The assessment of test predictive validity concerns a classroom test whose cutoff scores were based on the traditional and Angoff score-setting methods.

Test Content Validity

The assessment of the test's content validity encompassed two functions: the computation of an indices of relative emphasis for each test content topic in addition to the overall test and the estimate of the job relevance of each test item.

Indices of Relative Emphasis Computation

The relative emphasis given to each test content topic was computed through an estimate of the test's sampling validity by a group
of eight randomly selected faculty members. These reviewers assessed the level of correspondence between the curriculum and seven content topics on which the test was based. The reviewers specified whether the emphasis in the curriculum is the "same," "more," or "less" than it was in the test. Overall, the estimates indicated there is unanimous agreement among the reviewers that more curriculum emphasis is given to content topic number 1, Introduction and Army Master Data File Retrieval Microform System. The estimates also indicated that a similar situation exists concerning test topic number 5, Shop Stock Procedures (Manual). The reviewers' emphasis estimates are presented in Table 1.

TABLE 1

CURRICULUM AND TEST EMPHASIS ESTIMATES

<table>
<thead>
<tr>
<th>Content Topic</th>
<th>% of Test</th>
<th>Same</th>
<th>More</th>
<th>Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction and Army Master Data File Retrieval Microform System</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Prescribed Load List Procedures (Manual)</td>
<td>44</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Prescribed Load List Procedures (Automated)</td>
<td>13</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4. The Army Maintenance Management System (TAMHS) Procedures</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5. Shop Stock Procedures (Manual)</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>6. Shop Stock Procedures (Automated)</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>7. Shop Clerk Procedures</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
The proportion of all reviewers who selected each option was calculated for each test content topic. These proportions were, in turn, multiplied by the percent of the items on the test devoted to each topic. The products were summed, providing a weighted index in the range of 0 to 100. The separate indices, computed for each emphasis category, show the relative emphasis given to each content topic and the overall test. These indices are presented in Table 2.

**TABLE 2**

**INDICES OF RELATIVE EMPHASIS ON CURRICULUM AND TEST**

<table>
<thead>
<tr>
<th>Content Topic</th>
<th>% of Test</th>
<th>Same</th>
<th>More</th>
<th>Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction and Army Master Data File Retrieval Microform System</td>
<td>1</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Prescribed Load List Procedures (Manual)</td>
<td>44</td>
<td>44.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Prescribed Load List Procedures (Automated)</td>
<td>13</td>
<td>11.38</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td>4. The Army Maintenance Management System (TAMMS) Procedures</td>
<td>16</td>
<td>8.00</td>
<td>8.00</td>
<td></td>
</tr>
<tr>
<td>5. Shop Stock Procedures (Manual)</td>
<td>1</td>
<td>.13</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>6. Shop Stock Procedures (Automated)</td>
<td>7</td>
<td>3.50</td>
<td>3.50</td>
<td></td>
</tr>
<tr>
<td>7. Shop Clerk Procedures</td>
<td>18</td>
<td>9.00</td>
<td>2.25</td>
<td>6.75</td>
</tr>
</tbody>
</table>

Indices: 76.01, 9.24, 14.75
The index computed for the overall test in the "same" category is 76.01. The "more" index is 9.24. The "less" index is 14.75. A high index in the "more" category would imply that the test does not cover adequately the material taught in the course, thereby raising some doubt concerning the test's content validity. A high "less" index would suggest that the test contains material that was not taught in the course, thereby threatening the instructional validity of the curriculum. According to Cross (1982), a [norm-referenced] test is identified for further study "if the 'less' index exceeds 25, or, . . . if the 'less' index exceeds the 'more' index by 5 points or more" (p. 73). Even though the United States Army Quartermaster School's standard for criterion-referenced tests is more stringent, calling for a zero difference between the curriculum and the test, this analysis revealed a 5.51 difference between the "more" and "less" indices of relative emphasis on the test discussed in this study. This difference indicates there is a possible weakness in the test's content validity. Since curricular validity and instructional validity are two kinds of evidence of test content validity, a need exists to enhance the correspondence between the curriculum and the test.

Job Relevancce Assessment

The second function in assessing the test's content validity concerned the estimation of the job relevance of each test item. This involved a group of 72 test item judges, comprised of 36 instructors from the faculty of the United States Army Quartermaster School and 36 job specialists from Army units in Germany and the Continental United States. Each judge estimated the relevance of the subject matter
content of each test item of the 380-item, criterion-referenced, end-of-course comprehensive test to the domain of knowledge that entry-level job performers should possess.

Using a 4-point rating scale, each judge assessed whether the job relevance of each test item was equal to 1 (not relevant), 2 (questionable), 3 (important), or 4 (crucial). Average job relevance ratings were computed for each test item by the instructor, job specialist, and all rater categories of judges. The frequency of the judges' ratings, portrayed in 1-point intervals, is presented in Table 3. For instructors the average relevance estimates ranged from 3.194 to 3.806, a spread of .612 across all items. Relevance estimates by job specialists ranged from 2.788 to 3.944, a spread of 1.156 across all items. The range of estimates for job specialists was approximately 1.888 times greater than the range for instructors. The range of relevance estimates for all raters across all items was 3.042 to 3.833, a spread of .791. Based on the average relevance ratings of all raters, 368 of 380, or 96.84 percent of all test items, fell into the 3.300 to 3.900

---

1 Relevance ratings were computed using a software program that was developed at the U.S. Army Logistics Management Center, Fort Lee, Virginia. The program was based on a computer code developed at the Virginia Polytechnic Institute and State University, Blacksburg, Virginia in 1982 for a project directed by Dr. Lawrence H. Cross. The University project was conducted for the Virginia State Board of Education and was titled "Validation Study of the National Teachers Examination for the Certification of Entry-Level Teachers." The Logistics Management Center-developed program incorporates the Ebel (1951) intra-class formula to obtain a measure of test reliability utilizing the judgments of two groups of job experts. It was developed in Borland TURBO PASCAL version 3.0 and runs on any IBM or IBM-compatible personal computer using MS-DOS version 2.0 or higher.
<table>
<thead>
<tr>
<th>Range of Ratings</th>
<th>Instructors</th>
<th>Job Specialists</th>
<th>All Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.700 - 2.799</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2.800 - 2.899</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2.900 - 2.999</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3.000 - 3.099</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.100 - 3.199</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3.200 - 3.299</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>3.300 - 3.399</td>
<td>40</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>3.400 - 3.499</td>
<td>76</td>
<td>23</td>
<td>47</td>
</tr>
<tr>
<td>3.500 - 3.599</td>
<td>107</td>
<td>75</td>
<td>107</td>
</tr>
<tr>
<td>3.600 - 3.699</td>
<td>121</td>
<td>122</td>
<td>123</td>
</tr>
<tr>
<td>3.700 - 3.799</td>
<td>24</td>
<td>76</td>
<td>59</td>
</tr>
<tr>
<td>3.800 - 3.899</td>
<td>1</td>
<td>54</td>
<td>13</td>
</tr>
<tr>
<td>3.900 - 3.999</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>4.000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
relevance range, showing, overall, a strong relationship between the
test and the PLL clerk's job.

All test item average relevance ratings exceeded the 2.25 relevance
floor established for the elimination of any item from the proba-
bility assessment. All items were therefore included in the computation
of probability estimates and the test cutoff score.

**Test Cutoff Score Computation**

Concurrent with the assessment of the job relevance of each test
item, the judges estimated the probability that a minimally competent
individual would be able to answer each test question correctly without
guessing. Using a 1-to-10 scale, where the probability values corre-
pond to the upper limit of the intervals on the response scale: 1 = .0
through .10; 2 = .11 through .20; 3 = .21 through .30; ... 10 = .91
through 1.0, each judge rendered a probability estimate for each of the
380 items on the test. The probability values were summed and averaged,
which provided a raw score performance standard. This score was, in
turn, converted to a percentage score performance standard.  

---

1Probability estimates and standards computed using U. S. Army
Logistics Management Center software program, based on a computer code
developed at the Virginia Polytechnic Institute and State University.
standards. The average of the sum of the judges' probability values by each judge represents the new test standard, the number of test items a minimally competent student would be expected to answer correctly.

Raw score mean probability values (performance standard), standard deviations, and interrater reliability estimates across items and among judges were computed for instructors, job specialists, and all raters. The instructors' mean probability value was 345.77; the standard deviation was 17.99; and the interrater reliability was .037 based on individual judge's ratings on individual items, and .582 on the judges' average rating on individual items. The mean probability value for job specialists was 352.72, a difference of only 6.95 points between instructors and job specialists across 380 test items; the standard deviation was 16.38; and the interrater reliability based on job specialists' ratings on individual items was .057, and on item average ratings, .685. For all raters the raw score mean probability value was 349.25; the standard deviation was 17.44; and the interrater reliability values were .047 and .780. These raw score ratings are presented in Table 4.

Percentage Score Performance Standard

Because the U. S. Army Quartermaster School's grading system is based on percentage scores rather than raw scores, the derived all rater raw score standard was converted to a percentage score. To compare panel ratings, a percentage score standard was computed for each panel of judges. The computation was made by converting the mean probability values of instructor, job specialist, and all rater panels to a percentage score performance standard, the raw score of 380 being equal to 100
TABLE 4

RAW SCORE RATING DATA

<table>
<thead>
<tr>
<th>Raters</th>
<th>No. of Items</th>
<th>No. of Raters</th>
<th>Mean Score</th>
<th>$S_x$</th>
<th>Ind Rel</th>
<th>Avg Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructors</td>
<td>380</td>
<td>36</td>
<td>345.77</td>
<td>17.99</td>
<td>.037</td>
<td>.582</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>380</td>
<td>36</td>
<td>352.72</td>
<td>16.38</td>
<td>.057</td>
<td>.685</td>
</tr>
<tr>
<td>All Raters</td>
<td>380</td>
<td>72</td>
<td>349.25</td>
<td>17.44</td>
<td>.047</td>
<td>.780</td>
</tr>
</tbody>
</table>

Mean Score = Raw score mean probability value
$S_x$ = Standard deviation of rater panel probability value
Ind Rel = Interrater reliability of raters' ratings on indiv items
Avg Rel = Interrater reliability of raters' average rating for indiv items

percent. Because the difference in average raw score estimates of the instructor and job specialist panels was only 6.95 points, little difference in the percentage score standards of the raters was found. The conversion of mean raw scores to percentage scores (mean raw score divided by 3.8 = percentage score) produced the following standards:

<table>
<thead>
<tr>
<th>Raters</th>
<th>Mean Raw Score</th>
<th>Percentage Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructors</td>
<td>345.77</td>
<td>90.99</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>352.72</td>
<td>92.82</td>
</tr>
<tr>
<td>All Raters</td>
<td>349.25</td>
<td>91.91</td>
</tr>
</tbody>
</table>
The previous computational procedure was also utilized to convert raw score standard deviations (SD) to percentage score standard deviations. The conversion produced the following measures for each rater category:

<table>
<thead>
<tr>
<th>Raters</th>
<th>Raw Score SD</th>
<th>Percentage Score SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructors</td>
<td>17.99</td>
<td>4.73</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>16.38</td>
<td>4.31</td>
</tr>
<tr>
<td>All Raters</td>
<td>17.44</td>
<td>4.59</td>
</tr>
</tbody>
</table>

The significance of the interrater reliability ratings for the percentage score standards, shown in Table 5, is presented in the discussion of the study results relative to Hypothesis 1, page 81.

The percentage score for all raters constitutes the Angoff-derived test cutoff score, or new performance standard, for the end-of-course test for the Equipment Records and Parts Specialist Course. This new standard of 91.91, rounded to 92 percent, is more stringent than the present cutoff score of 85 percent which is based on instructor and school preferences.

Using the traditionally based cutoff score of 85 percent and the Angoff-based score of 92 percent, the effect of the change in test standard on the students' failure rate was assessed and is discussed on page 68 of this study.

**Standard Errors**

Before the relationship between the test standards and the soldiers' on-the-job performance was measured, the potential effect of chance error on the derived score, attributable to panelist and examinee performance inconsistency, was calculated. As indicated by Cross
(1982), a derived standard is subject to error from two potential sources: sampling error associated with the selection of test item judges and measurement error associated with the performance of the examinee. He noted that selection of a different set of test item judges could produce higher or lower standards based on the judges' probability estimates. He noted, too, that examinees' test scores may fluctuate if they are retested. The potential sampling error, measurement error, and composite of these errors were therefore computed.
Composite Standard Error Computation

A composite standard error ($SE_{comp}$) which incorporates sampling error of judges and measurement error of examinees was calculated. The computation was made by deriving the square root of the squared standard error of the mean of the raters plus the squared standard error of measurement of the examinees as shown in the following formula:

$$SE_{comp} = \sqrt{S_{\bar{x}}^2 + S_{meas}^2}$$

Sampling errors were computed for each panel of judges and the composite panel, labeled all raters. The measurement error was calculated using the scores of 725 of the students who completed the Equipment Records and Parts Specialist Course in 1986.

Sampling error computation

Computation of the sampling error of judges involved the calculation of $S_{\bar{x}}$, the standard error of the mean of the raters' derived percentage score for this study. Using the standard deviation for the rater panel's derived percentage scores ($S_X$), shown in Table 5, page 61, the $S_{\bar{x}}$ was computed for instructors, job specialists, and all raters. Each computation was made by dividing $S_X$ by the square root of the number of raters in each panel, using the following formula:

$$S_{\bar{x}} = S_X / \sqrt{n}$$

The statistics for the derived percentage score standards are presented in Table 6.
### TABLE 6

**STATISTICS FOR DERIVED PERCENTAGE SCORE STANDARDS**

<table>
<thead>
<tr>
<th>Raters</th>
<th>Characteristics of Derived Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td>Instructors</td>
<td>36</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>36</td>
</tr>
<tr>
<td>All Raters</td>
<td>72</td>
</tr>
</tbody>
</table>

- **n**: Number of raters in rater category
- **$S_x$**: Standard deviation of rater panel derived percentage scores
- **$S_{\bar{x}}$**: Standard error of the mean of raters' derived percentage scores

**Measurement error computation**

Computation of the measurement error concerned the calculation of an estimate of the reliability of the test involved in this study. The split-half method was used to obtain two test scores on each student.

The Rulon (1939) shortcut method for finding split-half reliability was utilized. In turn, the Spearman-Brown formula was applied to estimate the reliability of a test doubled in length to compensate for the half-test reliability which resulted from the application of the split-half reliability measurement procedure.
The estimate of the split-half reliability utilized the test scores of 725 of the students who completed the Equipment Records and Parts Specialist Course in 1986. The test was split using the student scores on the odd and even items of the test. The variance of the difference between each student's half-test scores ($\sigma_d^2$) and the variance of the sum of each student's half-test scores ($\sigma_n^2$) were calculated. These variances were 111.675 (differences) and 423.5 (sum). Utilizing these variances in the Rulon shortcut method, the following estimated reliability coefficient was obtained. The (e) denotes student scores.

$$r(e) = 1 - \frac{\sigma_d^2}{\sigma_n^2} = 1 - \frac{111.675}{423.5} = 1 - .2637 = .7363$$

Since the split-half method of computing the correlation (reliability) coefficient of the test produced only a half-test reliability, the simplified Spearman-Brown formula was used to double the length of the test. Where $r$ is the estimated coefficient, $r(e)$ is the obtained coefficient, and 2 denotes the doubling of the test length, the formula is: $r = \frac{2r(e)}{1+r(e)}$. The effect of doubling the length of the test is shown here:

$$r = \frac{2r(e)}{1+r(e)} = \frac{2 \times .7363}{1 + .7363} = \frac{1.4726}{1.7363} = .848$$

The validity of the test reliability coefficient was measured at the 95 percent confidence level. Accordingly, as shown in Table 7, the test's true reliability lies within the .80 to .88 range. Based on the derived test reliability coefficient (.848), repeated test results would fall within the established interval 85 percent of the time.
The error of measurement of examinee performance \( S_{\text{meas}} \) was calculated using the formula \( \sigma_{se} = \sigma_t \sqrt{1-r} \) (Hopkins and Antes, 1978). 

\( \sigma_{se} \) = standard error of measurement, \( \sigma_t \) = standard deviation of the student sample test scores, and \( r \) = the reliability coefficient of the test. The \( S_{\text{meas}} \) computation follows:

\[
\sigma_{se} = \sigma_t \sqrt{1-r} = 4.865 \sqrt{1 - .848} = 4.865 \sqrt{.152} = 4.865 \times .390 = 1.897
\]

The computer-produced standard error of measurement and associated statistics for the test used in this study are also presented in Table 7.

**TABLE 7**

**MEASUREMENT ERROR AND TEST STATISTICS**

<table>
<thead>
<tr>
<th>Test</th>
<th>n</th>
<th>( r )</th>
<th>( S(x) )</th>
<th>( S_{\text{meas}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLL Clerk Course</td>
<td>725</td>
<td>.848</td>
<td>4.865</td>
<td>1.897</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.80-0.88)</td>
</tr>
</tbody>
</table>

\( n \) = Number of examinees in test reliability sample  
\( r \) = Test reliability with confidence interval  
\( S(x) \) = Standard deviation of test reliability sample  
\( S_{\text{meas}} \) = Standard error of measurement

**Composite error**

The composite standard error \( S_{\text{comp}} \) was calculated utilizing the standard error of the mean of raters' derived percentage score, \( S_k \), shown in Table 6, page 63, and the student measurement error \( S_{\text{meas}} \).
error of measurement \( (S_{\text{meas}}) \) shown in Table 7, page 65. As noted previously, the composite standard error quantifies the chance error associated with performance inconsistency on the part of the test item judges and the students. The composite standard error was calculated for instructor and job specialist rater categories for informational purposes only. The formula used is \( SE_{\text{comp}} = \sqrt{\frac{S_x^2}{n} + S_{\text{meas}}^2} \). The composite standard error for each category of rater is presented in Table 8.

**TABLE 8**

**COMPOSITE STANDARD ERRORS**

<table>
<thead>
<tr>
<th>Raters</th>
<th>( SE_{\text{comp}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructors</td>
<td>2.05</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>2.03</td>
</tr>
<tr>
<td>All Raters</td>
<td>1.97</td>
</tr>
</tbody>
</table>

A summary of the measurement error and statistics associated with the percentage score standard is presented in Table 9. The summary also includes the characteristics of the test and the derived test standards.

The derived score standards and the adjusted scores for one, two, and three standard errors are shown in Table 10, page 68.
### TABLE 9

**MEASUREMENT ERROR AND ASSOCIATED STATISTICS FOR PERCENTAGE SCORE STANDARDS**

#### Test Characteristics

<table>
<thead>
<tr>
<th>Test</th>
<th>n</th>
<th>r</th>
<th>$S(x)$</th>
<th>$S_{\text{meas}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLL Clerk Course</td>
<td>725</td>
<td>.848</td>
<td>4.865</td>
<td>1.897</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.80-0.88)</td>
</tr>
</tbody>
</table>

#### Characteristics of Derived Standards

<table>
<thead>
<tr>
<th>Raters</th>
<th>Instructors</th>
<th>Job Specialists</th>
<th>All Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n^1$</td>
<td>$S_x$</td>
<td>$S_{\bar{x}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>4.73</td>
<td>.79</td>
</tr>
</tbody>
</table>

$n^1$ = Number of examinees in test reliability sample  
$r$ = Test reliability with confidence interval  
$S(x)$ = Standard deviation of test reliability sample  
$S_{\text{meas}}$ = Standard error of measurement  

$S_x$ = Standard deviation of rater panel derived percentage scores  
$S_{\bar{x}}$ = Standard error of the mean of raters' derived percentage scores  

$SE_{\text{comp}}$ = Composite standard error
### TABLE 10

**Derived and Adjusted Percentage Score Standards**

<table>
<thead>
<tr>
<th>Raters</th>
<th>Derived</th>
<th>Minus 1</th>
<th>Minus 2</th>
<th>Minus 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Std. Error</td>
<td>Std. Errors</td>
<td>Std. Errors</td>
</tr>
<tr>
<td>Instructors</td>
<td>90.99</td>
<td>88.94</td>
<td>86.89</td>
<td>84.84</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>92.82</td>
<td>90.79</td>
<td>88.76</td>
<td>86.73</td>
</tr>
<tr>
<td>All Raters</td>
<td>91.91</td>
<td>89.94</td>
<td>87.97</td>
<td>86.00</td>
</tr>
</tbody>
</table>

**Cutoff Scores and Associated Failure Rates**

Potential failure rates were calculated based on the new Angoff-derived test standard of 92 percent for the end-of-course comprehensive test. Failure rates were also computed for standards that had been reduced to compensate for estimated standard error. The rates were computed using the test scores achieved by 2,406 students who completed the Equipment Records and Parts Specialist Course during calendar year 1985.

The scores were obtained from U. S. Army Quartermaster School records. For this analysis, scores were deleted for 220 of 233 students who received the minimum passing score of 85 percent on the test. Because these students did not pass the test initially, they were given remedial training and retested at least once. For this reason their scores were not considered to be germane to the analysis of failure rates associated with the higher test standard. The analysis was, therefore, based on a student population of 2,186.
The number of students who received test scores ranging from 85 through 100 percent are arrayed by month and year in Table 11. Failure rates based on the new test standard, together with failure rates for standards that have been reduced by one, two, and three standard errors, are also shown. Specifically, the attrition which would have resulted from the establishment of the higher test standard of 92 percent is 800 students, or 36.60 percent of the adjusted annual enrollment. Lowering the standard by one standard error, or two percentage points, would have reduced the attrition rate to 447 students, or 20.45 percent of the adjusted annual enrollment. Adjustment of the passing score by two and three standard errors would have reduced the attrition to 196 students, or 8.97 percent of the enrollment, and 13 students, amounting to 0.59 percent of the enrollment, respectively.

Whereas other studies located in the review of existing literature made allowance for possible misclassification of competent individuals as incompetent, this study is concerned primarily with the possible misclassification of incompetent soldiers as competent. Criterion-referenced measurement is utilized in specialist-level training programs at the U.S. Army Quartermaster School to enhance the compatibility of the examinees' acquired skills and knowledge and the needs of the Army units to which the course graduates are assigned. Most course graduates occupy critical positions where substandard performance would affect the unit's capability to accomplish its mission. Any adjustment to the derived performance standard to compensate for possible chance error should be made only after very careful consideration.
### TABLE 11

STUDENT TEST SCORES AND DERIVED FAILURE RATES, CY 1985

<table>
<thead>
<tr>
<th>MONTH</th>
<th>85</th>
<th>86</th>
<th>87</th>
<th>88</th>
<th>89</th>
<th>90</th>
<th>91</th>
<th>93</th>
<th>94/95</th>
<th>96/97</th>
<th>98-100</th>
<th>85-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>______</td>
<td>4</td>
</tr>
<tr>
<td>Feb</td>
<td>13</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>13</td>
<td>16</td>
<td>49</td>
<td>67</td>
<td>48</td>
<td>13</td>
<td>251</td>
</tr>
<tr>
<td>Mar</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>13</td>
<td>14</td>
<td>23</td>
<td>25</td>
<td>30</td>
<td>______</td>
<td>7</td>
</tr>
<tr>
<td>Apr</td>
<td>13</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>14</td>
<td>15</td>
<td>23</td>
<td>55</td>
<td>43</td>
<td>40</td>
<td>12</td>
<td>242</td>
</tr>
<tr>
<td>May</td>
<td>15</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>11</td>
<td>15</td>
<td>36</td>
<td>37</td>
<td>32</td>
<td>______</td>
<td>7</td>
</tr>
<tr>
<td>Jun</td>
<td>23</td>
<td>5</td>
<td>11</td>
<td>14</td>
<td>13</td>
<td>21</td>
<td>18</td>
<td>46</td>
<td>29</td>
<td>27</td>
<td>______</td>
<td>7</td>
</tr>
<tr>
<td>Jul</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>24</td>
<td>25</td>
<td>29</td>
<td>19</td>
<td>______</td>
<td>142</td>
</tr>
<tr>
<td>Aug</td>
<td>13</td>
<td>5</td>
<td>10</td>
<td>16</td>
<td>12</td>
<td>16</td>
<td>15</td>
<td>48</td>
<td>55</td>
<td>54</td>
<td>22</td>
<td>266</td>
</tr>
<tr>
<td>Sep</td>
<td>20</td>
<td>15</td>
<td>4</td>
<td>12</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>45</td>
<td>47</td>
<td>37</td>
<td>14</td>
<td>235</td>
</tr>
<tr>
<td>Oct</td>
<td>31</td>
<td>11</td>
<td>16</td>
<td>11</td>
<td>16</td>
<td>18</td>
<td>24</td>
<td>36</td>
<td>27</td>
<td>14</td>
<td>6</td>
<td>210</td>
</tr>
<tr>
<td>Nov</td>
<td>35</td>
<td>9</td>
<td>6</td>
<td>11</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>31</td>
<td>32</td>
<td>32</td>
<td>25</td>
<td>207</td>
</tr>
<tr>
<td>Dec</td>
<td>42</td>
<td>10</td>
<td>17</td>
<td>19</td>
<td>18</td>
<td>23</td>
<td>17</td>
<td>27</td>
<td>16</td>
<td>22</td>
<td>18</td>
<td>229</td>
</tr>
<tr>
<td>TOTAL</td>
<td>220/13</td>
<td>77</td>
<td>106</td>
<td>122</td>
<td>129</td>
<td>170</td>
<td>183</td>
<td>436</td>
<td>419</td>
<td>377</td>
<td>154</td>
<td>2406/2186</td>
</tr>
</tbody>
</table>

**FAILURE RATE**

<table>
<thead>
<tr>
<th></th>
<th>800 students (36.60%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less 1 Std Err</td>
<td>447 students (20.45%)</td>
</tr>
<tr>
<td>Less 2 Std Err</td>
<td>196 students (8.97%)</td>
</tr>
<tr>
<td>Less 3 Std Err</td>
<td>13 students (.59%)</td>
</tr>
</tbody>
</table>
Summary: Test Content Validity and Cutoff Score Computation

The results of the study of test content validity and setting a cutoff score are as follows:

Assessment of the test's sampling validity through the development of an indices of relative emphasis showed there is a lack of proper correspondence between two of the test's content topics and the curriculum. This difference indicated a possible weakness in the test's content validity.

Through the assessment of the job relevance of each test item by a panel of instructors and job specialists, the conclusion was drawn that a strong relationship exists between each test item and the domain of knowledge that the entry-level soldier should possess. The estimated job relevance of each test item exceeded the relevance floor of 2.25 established for the elimination of the item from cutoff score computation. All test items fell into the 3.300 to 3.900 range, showing a strong relationship between the test and the soldier-specialist's job.

The computation of the test cutoff score using the Angoff score-setting method produced an average raw score rating of 349.25 out of a possible maximum raw score of 380. The raw score standard of 349.25 was converted to a percentage score standard of 91.91 across all raters. The derived test standard, rounded to 92 percent, is seven points higher than the present cutoff score of 85 set by instructor and school preference.

The potential effect of chance error due to panelist and examinee performance inconsistency was calculated. Sampling error (\(S_{X}\))
associated with the selection of test item judges was calculated at .54 for the all rater category of judges. Measurement error (\(S_{\text{meas}}\)) was calculated at 1.897 for 725 students whose grades on the end-of-course comprehensive test were subjected to the split-half reliability test, with a derived reliability (correlation) coefficient of .848 assessed at the 95 percent confidence level. Utilizing the sampling and measurement errors, a composite standard error of 1.97 was derived for the all rater category of test item judges.

Associated attrition rates for calendar year 1985, based on the derived percentage score standard of 92 percent, together with the failure rates for standards that were reduced by one, two, and three standard errors, were computed. The failure rates ranged from 800 students, or 36.60 percent of the annual student enrollment for the higher test standard, to 13 students, or .59 percent of the annual enrollment for the test standard that had been reduced by three standard errors.

**Description of Study Samples**

Eight samples were drawn for use in this study. The role, source, and size of each sample for each of the five hypotheses (Ho1 through Ho5) are summarized in Chart 1.

**Test of Hypotheses**

There are five hypotheses tested in this study. These postulations concern intragroup variability in the probability estimates rendered by test item judges and the interrater reliability among the judges. They also relate to the tests for strengths of linear relationships between soldier classroom test scores and supervisory rating
<table>
<thead>
<tr>
<th>Sample No:</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Role:</strong></td>
<td>Test item judges. Participants rendered test item relevance and probability estimates.</td>
</tr>
<tr>
<td><strong>Source:</strong></td>
<td>U. S. Army Quartermaster School instructor staff.</td>
</tr>
<tr>
<td><strong>Size:</strong></td>
<td>36 instructors.</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong></td>
<td>Ho1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample No:</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Role:</strong></td>
<td>Test item judges. Participants rendered test item relevance and probability estimates.</td>
</tr>
<tr>
<td><strong>Source:</strong></td>
<td>PLL clerks (job specialists) in Army units.</td>
</tr>
<tr>
<td><strong>Size:</strong></td>
<td>36 job specialists.</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong></td>
<td>Ho1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample No:</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Role:</strong></td>
<td>PLL clerks. Job performers' classroom test scores, ranging from 92 through 100 percent, based on Angoff score-setting method, were correlated with supervisory rating scores.</td>
</tr>
<tr>
<td><strong>Source:</strong></td>
<td>PLL clerks in U. S. Army units in the Continental United States, Germany, and Korea.</td>
</tr>
<tr>
<td><strong>Size:</strong></td>
<td>64 PLL clerks (job performers/Angoff &quot;accepts&quot;).</td>
</tr>
<tr>
<td><strong>Hypotheses:</strong></td>
<td>Ho2, Ho4</td>
</tr>
<tr>
<td>Sample No:</td>
<td>4</td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
</tr>
<tr>
<td>Role: PLL clerks. Job performers' classroom test scores, ranging from 85 through 100 percent, based on the present score-setting method, were correlated with supervisory rating scores.</td>
<td></td>
</tr>
<tr>
<td>Source: PLL clerks in U. S. Army units in the Continental United States, Germany, and Korea.</td>
<td></td>
</tr>
<tr>
<td>Size: 100 PLL clerks (job performers).</td>
<td></td>
</tr>
<tr>
<td>Hypothesis: Ho2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample No:</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role: PLL clerks. Job performers' classroom test scores, ranging from 92 through 100 percent, based on Angoff score-setting method, were correlated with soldiers' skill qualification test scores.</td>
<td></td>
</tr>
<tr>
<td>Source: PLL clerks in U. S. Army units in the Continental United States, Germany, and Korea.</td>
<td></td>
</tr>
<tr>
<td>Size: 49 PLL clerks (job performers/Angoff &quot;accepts&quot;).</td>
<td></td>
</tr>
<tr>
<td>Hypotheses: Ho3, Ho5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample No:</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role: PLL clerks. Job performers' classroom test scores, ranging from 85 through 100 percent, based on present cutoff score-setting method, were correlated with soldiers' skill qualification test scores.</td>
<td></td>
</tr>
<tr>
<td>Source: PLL clerks in U. S. Army units in the Continental United States, Germany, and Korea.</td>
<td></td>
</tr>
<tr>
<td>Size: 66 PLL clerks (job performers).</td>
<td></td>
</tr>
<tr>
<td>Hypothesis: Ho3</td>
<td></td>
</tr>
</tbody>
</table>
Sample No: 7
Role: PLL clerks. Job performers' classroom test scores, ranging from 85 through 91 percent for Angoff "rejects," were correlated with supervisory rating scores.
Source: PLL clerks in U. S. Army units in the Continental United States, Germany, and Korea.
Size: 36 PLL clerks (Angoff "rejects").
Hypothesis: Ho4

Sample No: 8
Role: PLL clerks. Job performers' classroom test scores, ranging from 85 through 91 percent for Angoff "rejects," were correlated with skill qualification test scores.
Source: PLL clerks in U. S. Army units in the Continental United States, Germany, and Korea.
Size: 17 PLL clerks (Angoff "rejects").
Hypothesis: Ho5
scores and classroom test scores and skill qualification test scores.
In addition, the hypotheses involve tests for significant differences in
means of supervisory rating scores and skill qualification test scores
for two groups of soldiers labeled Angoff "accepts," soldiers whose
classroom test scores ranged from 92 (Angoff-derived cutoff score)
through 100 percent, and Angoff "rejects," soldiers whose classroom test
scores ranged from 85 (present cutoff score) through 91 percent.

Hypothesis 1

Hypothesis 1 states there will be no significant difference in
intragroup variability in the probability estimates by instructors and
job specialists in their assessment of the minimum passing score for a
criterion-referenced classroom test; and the interrater reliability of
instructors' and job specialists' probability judgments will not be
statistically significant. The test of this hypothesis was conducted in
two parts. Part 1 concerned the assessment of the significance of
difference in intragroup variability of two groups of test item judges
engaged in assessing the probability that a minimally competent person
would answer each question correctly without guessing. Part 2 assessed
the interrater reliability of the two groups of judges engaged in making
the probability estimates.

Intragroup variability of probability estimates

In Part 1, the test for the significance of difference in intra-
group variability (p < .05) was made by performing a two-way analysis of
variance (items by raters) which treated the probability assessment as
the independent variable and the probability score as the dependent
variable. The two-way analysis of variance of the estimates of probability concerned 36 instructors, 36 job specialists, and the combined groups (all raters). The analysis of the instructor estimates resulted in an F ratio of 2.393 with 379 and 35 degrees of freedom. The difference in the estimates was significant at $p < .00001$. The results are presented in Table 12.

The analysis of variance of the estimates of probability rendered by the job specialists yielded an F ratio of 3.176 with 379 and 35 degrees of freedom. The difference in these estimates was also significant at $p < .00001$. The results of this analysis are presented in Table 13, page 79.

The analysis of variance of the probability estimates assessed by the all rater category of test item judges resulted in an F ratio of 4.542 with 379 and 71 degrees of freedom. The outcome revealed a significant difference in intragroup variability of the probability estimates by the all rater category of judges. The difference was significant at $p < .00001$. This indicates that the backgrounds and experiences of the two groups of judges influenced their estimates concerning the capability of a minimally competent job performer. The results are presented in Table 14, page 80. On the basis of this significant difference, Part 1 of Hypothesis 1, which states that there will be no significant difference in intragroup variability of the probability estimates by instructors and job specialists in their assessment of the minimum passing score for a criterion-referenced classroom test, was rejected.
### TABLE 12

**PART 1 OF HYPOTHESIS 1 - TWO-WAY ANALYSIS OF VARIANCE OF PROBABILITY ESTIMATES BY INSTRUCTORS**

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM OF SQUARES</th>
<th>DEGREES OF FREEDOM</th>
<th>MEAN SQUARE</th>
<th>F</th>
<th>SIGNIFICANCE OF F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Items</td>
<td>4.878</td>
<td>379.000</td>
<td>0.013</td>
<td>2.393</td>
<td>0.00000*</td>
</tr>
<tr>
<td>Raters</td>
<td>29.814</td>
<td>35.000</td>
<td>0.852</td>
<td>158.373</td>
<td>0.00000</td>
</tr>
<tr>
<td>Error</td>
<td>71.348</td>
<td>13265.000</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>106.040</td>
<td>13679.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interrater Reliability of Individual Ratings: 0.037  95% C.I. = 0.02-0.05
Interrater Reliability of Average Rating: 0.582  95% C.I. = 0.47-0.67

*Significant at \( p < 0.00001 \)
### TABLE 13

**PART 1 OF HYPOTHESIS 1 - TWO-WAY ANALYSIS OF VARIANCE OF PROBABILITY ESTIMATES BY JOB SPECIALISTS**

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM OF SQUARES</th>
<th>DEGREES OF FREEDOM</th>
<th>MEAN SQUARE</th>
<th>F</th>
<th>SIGNIFICANCE OF F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Items</td>
<td>9.679</td>
<td>379.000</td>
<td>0.026</td>
<td>3.176</td>
<td>0.00000*</td>
</tr>
<tr>
<td>Raters</td>
<td>24.704</td>
<td>35.000</td>
<td>0.706</td>
<td>87.775</td>
<td>0.00000</td>
</tr>
<tr>
<td>Error</td>
<td>106.670</td>
<td>13265.000</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>141.053</td>
<td>13679.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interrater Reliability of Individual Ratings: 0.057
95% C.I. = 0.04-0.08

Interrater Reliability of Average Rating: 0.685
95% C.I. = 0.60-0.75

*Significant at p < 0.00001
TABLE 14

PART 1 OF HYPOTHESIS I - TWO-WAY ANALYSIS OF VARIANCE OF PROBABILITY ESTIMATES BY ALL RATERS

<table>
<thead>
<tr>
<th>Raters:</th>
<th>All Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum of Squares</td>
</tr>
<tr>
<td>Test Items</td>
<td>11.577</td>
</tr>
<tr>
<td>Raters</td>
<td>56.807</td>
</tr>
<tr>
<td>Error</td>
<td>180.997</td>
</tr>
<tr>
<td>Total</td>
<td>249.381</td>
</tr>
</tbody>
</table>

Interrater Reliability of Individual Ratings: 0.047  95% C.I. = 0.03-0.06
Interrater Reliability of Average Rating: 0.780  95% C.I. = 0.72-0.83

*Significant at p < .00001
**Interrater reliability of probability estimates**

In Part 2 of the test of Hypothesis 1, the interrater reliability of the judges' probability ratings (p < .05) was assessed using a procedure described by Ebel (1951) for estimating the reliability of average ratings. The confidence interval for the reliability estimates was set at 95 percent. The estimate of reliability of the judges' average ratings used the variance components from the two-way analysis of variance of the probability estimates by the instructors and job specialists. These variance components were entered into the following formula, (Ebel (1951)), where i denotes items and r denotes reviewers:

\[
\hat{r}_{xx} = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \hat{\sigma}_{ir/nr}^2}
\]

The reliability estimates for each panel of judges do not reflect individual differences or between reviewer variance. Instead, they reflect only the extent to which the judges' probability estimates covary (correlate) across items and among judges.

The test for significance of interrater reliability concerned instructor and job specialist average probability ratings for 380 individual test items. For comparison purposes a reliability coefficient was computed for the 36-member instructor panel, the 36-member job specialist panel, and the all rater panel of 72 judges. The test of the hypothesis was based on the correlation coefficient computed for the all rater panel.

The interrater reliability coefficient of correlation for the instructor panel was .582, significant at p < .01. The 95 percent confidence limits obtained for the \(r\) value for the instructor panel were
.47 and .67. The interrater reliability coefficient of correlation for the job specialist panel was .685, significant at $p < .01$. The 95 percent confidence limits for the $r$ value for the job specialist panel were .60 and .75. Interrater reliability coefficient of correlation for the all rater panel was .780, significant at $p < .01$. The 95 percent confidence limits for the all rater panel $r$ value were .72 and .83. These reliability coefficients and confidence limits for each rater panel are presented in Table 15.

**Table 15**

**PART 2 OF HYPOTHESIS 1 - Interrater Reliability Coefficients and Confidence Intervals**

<table>
<thead>
<tr>
<th>Panel</th>
<th>$r$ Value</th>
<th>Significance of $r$</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructors</td>
<td>.582</td>
<td>.01</td>
<td>.47 - .67</td>
</tr>
<tr>
<td>Job Specialists</td>
<td>.685</td>
<td>.01</td>
<td>.60 - .75</td>
</tr>
<tr>
<td>All Raters</td>
<td>.780</td>
<td>.01</td>
<td>.72 - .83</td>
</tr>
</tbody>
</table>
The confidence intervals obtained for the $r$ values for each panel enable the researcher to assert with 95 percent confidence that the population value for each of the correlation coefficients lies within the obtained limits. At 95 percent confidence, one can conclude there is a 1-in-20 chance that the all rater panel results with a correlation coefficient of .780 would fall outside the confidence limits.

The correlation coefficient for the all rater panel of judges, like the $r$ value for the instructor and job specialist panels, indicated there is a statistically significant interrater reliability of the test item judges:

$$r = .780$$
$$n = 378$$
$$p < .01$$

Based on the strength of this correlation coefficient, Part 2 of Hypothesis 1, which states that the interrater reliability of instructors' and job specialists' probability judgments will not be statistically significant, was rejected.

Hypothesis 2

Hypothesis 2 states there will be no difference in the correlation between classroom test scores and supervisory rating test scores of job performers whose classroom test cutoff scores were set by the Angoff method and the present method. The test of this hypothesis was conducted in two parts. Part 1 assessed the strength of the linear relationship between the classroom test scores and supervisory rating scores of 64 job performers who received a grade in the range of 92 through 100 percent, based on the Angoff-derived test cutoff score of 97.
Part 2 assessed the strength of the linear relationship between classroom test scores and supervisory rating scores of 100 job performers who received a test score in the range of 85 through 100 percent, based on the present test cutoff score of 85. The same test score data were used for grade range 92 through 100 percent in the Angoff-derived and present cutoff score groups.

This hypothesis was not tested statistically in the classical sense. The Pearson Product-Moment correlation, significant at \( p < .05 \), was used to assess the strength of the linear relationship between the independent variable (classroom test scores) and the dependent variable (supervisory rating scores) for two groups of soldiers whose classroom test cutoff scores were set by the Angoff and conventional methods.\(^1\)

The test of the hypothesis assessed the similarity of the correlations by comparing the relative sizes of the two \( r \)'s. The magnitude of the difference between the correlation coefficients was used to assess the significance of the difference between the two samples.

The test of this hypothesis involved the computation of correlation coefficients between classroom test scores and four supervisory

---

\(^1\) The Pearson Product-Moment correlation was used to assess the relationship between the classroom test scores and supervisory rating scores of two groups of soldiers whose classroom test cutoff scores were set by the conventional method and the Angoff method. The similarity of the correlations for the two samples was assessed by comparing the relative sizes of the two coefficients. Dr. Lawrence H. Cross, Associate Professor, Educational Research, Virginia Polytechnic Institute and State University, in his letter, dated June 18, 1987, recommended against the use of either the ANCOVA or ANOVA statistic for the test of Hypotheses 2 and 3. He noted that "the two groups are anything but randomly equivalent . . . . They differ systematically by definition of group membership." A copy of Dr. Cross' letter is provided at Appendix F, page 154.
rating scores for the two score-setting methods. The four job performance areas for which supervisory ratings of soldier performance were rendered included technical knowledge/skill, effort, following regulations and orders, and maintaining assigned equipment. A description of each of the variable labels used in the tests of Hypotheses 2 through 5 is presented in Chart 2. A summary of the coefficients of correlation between soldier classroom test scores and supervisory rating scores on soldier job performance based on the Angoff score-setting and present methods is presented in Table 16, page 87.

Part 1 - Test of Hypothesis 2

AACTS with AASRSl

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores on soldier technical knowledge/skill (AACTS with AASRSl) for the Angoff cutoff score-setting method. This computation resulted in a coefficient of correlation of .8724, significant at $p < .001$ level of confidence:

$r = .8724$

$n = 64$

$p < .001$

This correlation coefficient indicated there was a statistically significant relationship between soldier classroom test scores and supervisory rating scores of soldier performance in technical knowledge/skill with test cutoff score set by the Angoff method.
### Chart 2

#### Variable Labels

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACTS</td>
<td>Angoff Accepts - Classroom test scores</td>
</tr>
<tr>
<td>AASRS1</td>
<td>Angoff Accepts - Supv rating scores/technical skills</td>
</tr>
<tr>
<td>AASRS2</td>
<td>Angoff Accepts - Supv rating scores/effort</td>
</tr>
<tr>
<td>AASRS3</td>
<td>Angoff Accepts - Supv rating scores/follow procedures</td>
</tr>
<tr>
<td>AASRS4</td>
<td>Angoff Accepts - Supv rating scores/maintain equipment</td>
</tr>
<tr>
<td>PMCTS</td>
<td>Present Method - Classroom test scores</td>
</tr>
<tr>
<td>PMSRS1</td>
<td>Present Method - Supv rating scores/technical skills</td>
</tr>
<tr>
<td>PMSRS2</td>
<td>Present Method - Supv rating scores/effort</td>
</tr>
<tr>
<td>PMSRS3</td>
<td>Present Method - Supv rating scores/follow procedures</td>
</tr>
<tr>
<td>PMSRS4</td>
<td>Present Method - Supv rating scores/maintain equipment</td>
</tr>
<tr>
<td>AASQT</td>
<td>Angoff Accepts - Skill qualification test scores</td>
</tr>
<tr>
<td>PMSQT</td>
<td>Present Method - Skill qualification test scores</td>
</tr>
<tr>
<td>ARCTS</td>
<td>Angoff Rejects - Classroom test scores</td>
</tr>
<tr>
<td>ARSRS1</td>
<td>Angoff Rejects - Supv rating scores/technical skills</td>
</tr>
<tr>
<td>ARSRS2</td>
<td>Angoff Rejects - Supv rating scores/effort</td>
</tr>
<tr>
<td>ARSRS3</td>
<td>Angoff Rejects - Supv rating scores/follow procedures</td>
</tr>
<tr>
<td>ARSRS4</td>
<td>Angoff Rejects - Supv rating scores/maintain equipment</td>
</tr>
<tr>
<td>ARSQT</td>
<td>Angoff Rejects - Skill qualification test scores</td>
</tr>
</tbody>
</table>
### Table 16

**Hypothesis 2 - Correlation Coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Angoff Method</th>
<th>Present Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AACTS</td>
<td>PMCTS</td>
</tr>
<tr>
<td>with AASRS1</td>
<td>( r = .8724 )</td>
<td>( r = .519! )</td>
</tr>
<tr>
<td>( n = 64 )</td>
<td>( p &lt; .001 )</td>
<td>( n = 100 )</td>
</tr>
<tr>
<td></td>
<td>AACTS</td>
<td>PMCTS</td>
</tr>
<tr>
<td>with AASRS2</td>
<td>( r = .6841 )</td>
<td>( r = .3980 )</td>
</tr>
<tr>
<td>( n = 64 )</td>
<td>( p &lt; .001 )</td>
<td>( n = 100 )</td>
</tr>
<tr>
<td></td>
<td>AACTS</td>
<td>PMCTS</td>
</tr>
<tr>
<td>with AASRS3</td>
<td>( r = .6728 )</td>
<td>( r = .3542 )</td>
</tr>
<tr>
<td>( n = 64 )</td>
<td>( p &lt; .001 )</td>
<td>( n = 100 )</td>
</tr>
<tr>
<td></td>
<td>AACTS</td>
<td>PMCTS</td>
</tr>
<tr>
<td>with AASRS4</td>
<td>( r = .6543 )</td>
<td>( r = .2476 )</td>
</tr>
</tbody>
</table>
| \( n = 64 \)    | \( p < .001 \) | \( n = 100 \)  |}

**AACTS with AASRS2**

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier job effort (AACTS with AASRS2) for the Angoff cutoff score-setting method. This computation resulted in a coefficient of correlation of \( .6841 \), significant at \( p < .001 \) level of confidence:

\[
\begin{align*}
r &= .6841 \\
p &= 64 \\
p &< .001
\end{align*}
\]
This correlation coefficient indicated there was a statistically significant relationship between soldier classroom test scores and supervisory rating scores concerning soldier job effort with test cutoff score set by the Angoff method.

AACTS with AASRS3

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier performance in following established regulations and orders (AACTS with AASRS3) for the Angoff cutoff score-setting method. This computation resulted in a coefficient of correlation of \( r = 0.672 \), significant at \( p < 0.001 \) level of confidence:

\[
\begin{align*}
    r &= 0.672 \\
    n &= 64 \\
    p &< 0.001
\end{align*}
\]

This correlation coefficient indicated there was a statistically significant relationship between soldier classroom test scores and supervisory rating scores concerning soldier performance in following established regulations and orders with test cutoff score set by the Angoff method.

AACTS with AASRS4

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier performance in maintaining assigned equipment (AACTS with AASRS4) for the Angoff score-setting method. This computation resulted in a coefficient of correlation of \( r = 0.6543 \), significant at \( p < 0.001 \) level of confidence:
This correlation coefficient indicated there was a statistically significant relationship between soldier classroom test scores and supervisory rating scores concerning soldier performance in maintaining assigned equipment with test cutoff score set by the Angoff method.

**Part 2 - Test of Hypothesis 2**

**PMCTS with PMSRS1**

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores on soldier technical knowledge/skill (PMCTS with PMSRS1) for the present cutoff score-setting method. This computation resulted in a coefficient of correlation of .5191 which was significant at \( p < .001 \) level of confidence:

\[
\begin{align*}
\text{r} &= .5191 \\
\text{n} &= 100 \\
\text{p} &< .001
\end{align*}
\]

This correlation coefficient indicated there was a statistically significant relationship between PMCTS and PMSRS1 for the present score-setting method. In comparison, the correlation coefficient based on the Angoff-derived test cutoff score (AACTS with AASRS1) showed a stronger relationship than the present score-setting method. The Angoff-based coefficient was .8724.
PMCTS with PMSRS2

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier job effort (PMCTS with PMSRS2) for the present cutoff score-setting method. This computation resulted in a coefficient of correlation of .3980 which was significant at \( p < .001 \) level of confidence:

\[
\begin{align*}
\text{r} & = .3980 \\
\text{n} & = 100 \\
\text{p} & < .001
\end{align*}
\]

This correlation coefficient indicated there was a statistically significant relationship between PMCTS and PMSRS2 for the present cutoff score-setting method. In comparison, the correlation coefficient based on the Angoff-derived test cutoff score (AACTS with AASRS2) showed a stronger relationship than the present score-setting method. The Angoff-based coefficient was .6841.

PMCTS with PMSRS3

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores on soldier performance in following established regulations and orders (PMCTS with PMSRS3) for the present cutoff score-setting method. This computation resulted in a coefficient of correlation of .3542 which was significant at \( p < .001 \) level of confidence:

\[
\begin{align*}
\text{r} & = .3542 \\
\text{n} & = 100 \\
\text{p} & < .001
\end{align*}
\]
This correlation coefficient indicated there was a statistically significant relationship between PMCTS and PMSRS3 for the present cutoff score-setting method. In comparison, the correlation coefficient based on the Angoff-derived test cutoff score (AACTS with AASRS3) showed a stronger relationship than the present score-setting method. The Angoff-based coefficient was .6728.

PMCTS with PMSRS4

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier performance in maintaining assigned equipment (PMCTS with PMSRS4) for the present cutoff score-setting method. This computation resulted in a coefficient of correlation of .2476 which was significant at \( p < .006 \) level of confidence:

\[
    r = .2476 \\
    n = 100 \\
    p < .006
\]

This correlation coefficient indicated there was a statistically significant relationship between PMCTS and PMSRS4 for the present cutoff score-setting method. In comparison, the correlation coefficient based on the Angoff-derived test cutoff score (AACTS with AASRS4) showed a stronger relationship than the present score-setting method. The Angoff-based coefficient was .6543.

Major Findings

In Hypothesis 2 the similarity of the correlation coefficients of classroom test scores and supervisory rating scores for the Angoff and present score-setting methods was assessed by comparing the relative
sizes of the two r's for each of four sets of correlations. In each comparison the strength of the relationship of the scores under the Angoff method exceeded the strength of the relationship under the present method. On the basis of these differences, the hypothesis that there would be no difference in the correlations between classroom test scores and supervisory rating scores of job performers whose classroom test cutoff scores were set by the Angoff method and present method was rejected.

Hypothesis 3

Hypothesis 3 states there will be no difference in the correlation between classroom test scores and skill qualification test scores of job performers whose classroom test cutoff scores were set by the Angoff method and the present method. The test of this hypothesis assessed the strength of the linear relationship between the two sets of scores of 49 job performers who received a test score in the range of 92 through 100 percent, based on the Angoff-derived test cutoff score of 92. It also assessed the strength of the linear relationship between the classroom test scores and skill qualification test scores of 66 job performers who received a grade in the range of 85 through 100 percent, based on the present test cutoff score of 85. The same test score data were used for grade range 92 through 100 percent in the Angoff-derived and present cutoff score groups.

This hypothesis was not tested statistically. A Pearson Product-Moment correlation, significant at $p < .05$, was used to assess the relationship between the independent variable (classroom test score) and the dependent variable (skill qualification test score) for two
groups of soldiers whose classroom test cutoff scores were set by the Angoff and conventional methods. The test of the hypothesis assessed the similarity of the correlations by comparing the relative sizes of the two r's. The null hypothesis was not accepted or rejected in the classical sense. Instead, the magnitude of the difference between the correlation coefficients was used to assess the significance of the difference between the two samples.

Part I - Test of Hypothesis

AACTS with AASQT

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and soldier skill qualification test scores (AACTS with AASQT) for the Angoff cutoff score-setting method. This resulted in a correlation coefficient of .3686, significant at \( p < .005 \) level of confidence:

\[
\begin{align*}
r &= .3686 \\
n &= 49 \\
p &= .005
\end{align*}
\]

This correlation coefficient indicated there was a statistically significant relationship between AACTS and AASQT for the Angoff cutoff score-setting method.

---

1The Pearson Product-Moment correlation was used based on the recommendation of Dr. Lawrence H. Cross, Virginia Polytechnic Institute and State University, in his letter, dated June 18, 1987, not to use either ANCOVA or ANOVA because the two groups were not randomly equivalent.
Part 2 - Test of Hypothesis 3

PMCTS with PMSQT

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and skill qualification test scores (PMCTS with PMSQT) for the present cutoff score-setting method. This computation resulted in a correlation coefficient of .5282, significant at $p < .001$ level of confidence:

$$r = .5282$$
$$n = 66$$
$$p < .001$$

This correlation coefficient indicated there was a statistically significant relationship between PMCTS and PMSQT for the present method.

Major Findings

Hypothesis 3 was not tested statistically in the classical sense. The similarity of the correlation coefficients of classroom test scores and skill qualification test scores for the Angoff and present score-setting methods was assessed by comparing the relative sizes of the two $r$'s. This comparison showed that the strength of the correlation coefficient for the present method ($r = .5282$, $p < .001$) exceeded the coefficient for the Angoff method ($r = .3686$, $p < .005$). On the basis of this difference, the hypothesis that there would be no difference in the correlations between classroom test scores and skill qualification test scores of job performers, based on the Angoff method and present score-setting methods, was rejected.
Hypothesis 4

Hypothesis 4 states there will be no difference in the correlation between classroom test scores and supervisory rating scores received by Angoff "accepts" and Angoff "rejects"; and there will be no significant difference in the means of the supervisory rating scores received by these two groups of job performers. The test of Hypothesis 4 was conducted in three parts. Part 1 assessed the strength of the linear relationship between classroom test scores and supervisory rating scores of 64 job performers referred to as Angoff "accepts," soldiers who achieved classroom test scores from 92 through 100 percent. Part 2 assessed the strength of the linear relationship between classroom test scores and supervisory rating scores of 36 job performers referred to as Angoff "rejects," soldiers who achieved classroom test scores from 85 through 91 percent. Part 3 assessed the difference in means of supervisory rating scores of Angoff "accepts" and Angoff "rejects."

Parts 1 and 2 of Hypothesis 4 were not tested in the classical sense. The Pearson Product-Moment correlation, significant at $p < .05$, was used in Parts 1 and 2 of the test of this hypothesis to assess the strength of the linear relationship between the independent variable (classroom test scores) and the dependent variable (supervisory rating scores) on soldier performance in four job areas: technical knowledge/skill, effort, following regulations and orders, and maintaining assigned equipment. The $t$-test, significant at $p < .05$, was applied to assess the statistical significance of the difference between the means of the supervisory rating scores in the four job performance areas for Angoff "accepts" and Angoff "rejects." A summary of the coefficients of
correlation between soldier classroom test scores and supervisory rating scores on soldier job performance for Angoff "accepts" and Angoff "rejects" is presented in Table 17.

TABLE 17

HYPOTHESIS 4 - CORRELATION COEFFICIENTS

<table>
<thead>
<tr>
<th>Angoff &quot;Accepts&quot;</th>
<th>Angoff &quot;Rejects&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACTS</td>
<td>AACTS</td>
</tr>
<tr>
<td>with</td>
<td>with</td>
</tr>
<tr>
<td>AASRS1</td>
<td>ARCT5S1</td>
</tr>
<tr>
<td>P &lt; .001</td>
<td>P &lt; .485</td>
</tr>
<tr>
<td>AACTS</td>
<td>ARCT5S2</td>
</tr>
<tr>
<td>with</td>
<td>with</td>
</tr>
<tr>
<td>AASRS2</td>
<td>ARCT5S2</td>
</tr>
<tr>
<td>P &lt; .001</td>
<td>P &lt; .109</td>
</tr>
<tr>
<td>AACTS</td>
<td>ARCT5S3</td>
</tr>
<tr>
<td>with</td>
<td>with</td>
</tr>
<tr>
<td>AASRS3</td>
<td>ARCT5S3</td>
</tr>
<tr>
<td>P &lt; .001</td>
<td>P &lt; .148</td>
</tr>
<tr>
<td>AACTS</td>
<td>ARCT5S4</td>
</tr>
<tr>
<td>with</td>
<td>with</td>
</tr>
<tr>
<td>AASRS4</td>
<td>ARCT5S4</td>
</tr>
<tr>
<td>P &lt; .001</td>
<td>P &lt; .009</td>
</tr>
</tbody>
</table>

Part 1 - Test of Hypothesis 4

In Part 1 of the test of Hypothesis 4, the correlations for the Angoff "accepts" included AACTS with AASRS1, AACTS with AASRS2, AACTS with AASRS3, and AACTS with AASRS4. This study sample and the correlation coefficients presented are the same as those reflected in Part 1 of the test of Hypothesis 2 (pages 85 through 89) and are, therefore, not
discussed here. The sizes of these correlation coefficients were compared to the \( r \) values for soldier classroom test scores and supervisory rating scores in Part 2 of the test of this hypothesis for the Angoff "rejects," to assess significance of the difference in the two samples.

**Part 2 - Test of Hypothesis 4**

**ARCTS with ARSRS1**

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores of soldier technical knowledge/skill (ARCTS with ARSRS1) for Angoff "rejects." This computation produced an \( r \) value of \( .0067 \), significant at \( p < .485 \) level of confidence:

\[
\begin{align*}
\text{r} &= .0067 \\
n &= 36 \\
p &= .485
\end{align*}
\]

This correlation coefficient indicated that the relationship between ARCTS and ARSRS1 was not statistically significant for Angoff "rejects." The relative strength of this coefficient for Angoff "rejects" and the \( r \) value for classroom test scores and supervisory ratings of soldier job performance for Angoff "accepts" (\( r = .8724, p < .001 \)) was not assessed. Such a comparison is regarded as invalid.

**ARCTS with ARSRS2**

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier job effort (ARCTS with ARSRS2) for Angoff "rejects."
This computation produced a negative coefficient of correlation of 
\[-.2106\] at \(p < .109\) level of confidence:

\[
\begin{align*}
\text{r} &= -.2106 \\
n &= 36 \\
p &< .109
\end{align*}
\]

This correlation coefficient indicated that the relationship 
between ARCTS and ARSRS2 was not statistically significant for Angoff
"rejects." The relative strength of this coefficient for Angoff "re­
jects" and the \(r\) value for classroom test scores and supervisory ratings 
of soldier job performance for Angoff "accepts" (\(r = .6841, p < .001\)) was 
not assessed. Such a comparison is regarded as invalid.

ARCTS with ARSRS3

A Pearson Product-Moment coefficient of correlation was computed 
for soldier classroom test scores and supervisory rating scores on sol­
dier performance in following established regulations and orders (ARCTS 
with ARSRS3) for Angoff "rejects." This computation produced a negative 
coefficient of correlation of \[-.1788\], at \(p < .148\) level of confidence:

\[
\begin{align*}
\text{r} &= -.1788 \\
n &= 36 \\
p &< .148
\end{align*}
\]

This correlation coefficient indicated that the relationship 
between ARCTS and ARSRS3 was not statistically significant for Angoff
"rejects." The relative strength of this coefficient for Angoff "re­
jects" and the \(r\) value for classroom test scores and supervisory ratings 
of soldier job performance for Angoff "accepts" (\(r = .6728, p < .001\)) was 
not assessed. Such a comparison is regarded as invalid.
A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and supervisory rating scores concerning soldier performance in maintaining assigned equipment (ARCTS with ARSR54) for Angoff "rejects." This computation produced a negative coefficient of correlation of \(-.3928\), significant at \(p < .009\) level of confidence:

\[
r = -.3928 \\
n = 36 \\
p < .009
\]

This correlation coefficient indicated there was a statistically significant negative relationship between ARCTS and ARSR54 for Angoff "rejects." In comparison, the correlation coefficient based on the Angoff-derived test cutoff score (AACTS with AASRS4) showed a stronger relationship than the present score-setting method. The Angoff-based coefficient was \(.6543\).

The similarity of the correlation coefficients of classroom test scores and supervisory rating scores for Angoff "accepts" and Angoff "rejects" was not assessed for the three areas of job performance (technical knowledge/skill, effort, and following regulations and orders) in which coefficients for Angoff "rejects" were not significant. The sizes of the two r's concerning soldier job performance in maintaining assigned equipment were compared. In this comparison the strength of the relationship of classroom test scores and supervisory rating scores for Angoff "accepts" exceeded the strength of the relationship of the scores for Angoff "rejects." On the basis of the strength of the r
values obtained, one can assert there is a difference in the correlations between classroom test scores and supervisory rating scores for Angoff "accepts" and Angoff "rejects."

**Part 3 - Test of Hypothesis 4**

Part 3 of the test of Hypothesis 4 involved the use of the $t$-test ($p < .05$) to determine the significance of the difference in means of the supervisory rating scores received by Angoff "accepts," soldiers who achieved classroom test scores from 92 through 100 percent, and Angoff "rejects," soldiers who achieved classroom test scores from 85 through 91 percent. The $t$-test was applied to scores on four areas of job performance: AASRS1 with ARSRS1 (technical knowledge/skill), AASRS2 with ARSRS2 (effort), AASRS3 with ARSRS3 (follow procedures), and AASRS4 with ARSRS4 (maintenance). A summary of the $t$ values for Angoff "accepts" and Angoff "rejects" is provided in Table 18.

**AASRS1 with ARSRS1**

A $t$-test was used to determine if there was a statistically significant difference in the supervisory rating scores on soldier technical knowledge/skill received by Angoff "accepts" and Angoff "rejects" (AASRS1 with ARSRS1). Angoff "accepts" achieved a .9271 mean gain, from 4.4167 to 5.3438, representing a $t$ value of 3.61, significant at $p < .001$. The results are presented in Table 19, page 102.

This $t$ value indicated there was a statistically significant difference in the supervisory rating scores concerning soldier technical knowledge/skill received by Angoff "accepts" and Angoff "rejects."
### Table 18

**Hypothesis 4 - T Values**

<table>
<thead>
<tr>
<th>AREA</th>
<th>T VALUE</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASRS1</td>
<td>t = 3.61</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>with ARSRS1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AASRS2</td>
<td>t = 3.21</td>
<td>&lt; .002</td>
</tr>
<tr>
<td>with ARSRS2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AASRS3</td>
<td>t = 2.53</td>
<td>&lt; .013</td>
</tr>
<tr>
<td>with ARSRS3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AASRS4</td>
<td>t = 1.75</td>
<td>&lt; .084</td>
</tr>
<tr>
<td>with ARSRS4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 19

PART 3 OF HYPOTHESIS 4 - T-TEST OF SIGNIFICANT DIFFERENCE IN SUPERVISORY RATING
SCORES CONCERNING TECHNICAL SKILLS OF ANGOFF "ACCEPTS" AND ANGOFF "REJECTS"

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>NUMBER OF CASES</th>
<th>MEAN</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>F VALUE</th>
<th>2-TAIL PROB</th>
<th>POOLED VARIANCE ESTIMATE</th>
<th>T VALUE</th>
<th>DEGREES OF FREEDOM</th>
<th>2-TAIL PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASRS1 AND ARSRS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GROUP 1</td>
<td>64</td>
<td>5.3438</td>
<td>1.042</td>
<td>0.130</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td>3.61</td>
<td>98</td>
</tr>
<tr>
<td>GROUP 2</td>
<td>36</td>
<td>4.4167</td>
<td>1.519</td>
<td>0.253</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant at p < .001
AASRS2 with ARSRS2

A t-test was used to determine if there was a statistically significant difference in the supervisory rating scores on job effort received by Angoff "accepts" and Angoff "rejects" (AASRS2 with ARSRS2). Angoff "accepts" achieved a .8646 mean gain, from 4.4167 to 5.2813, representing a t value of 3.21, significant at p < .002. The results are presented in Table 20.

This t value indicated there was a statistically significant difference in the supervisory rating scores on job effort received by Angoff "accepts" and Angoff "rejects."

AASRS3 with ARSRS3

A t-test was used to determine if there was a statistically significant difference in the supervisory rating scores on soldier performance in following established regulations and orders received by Angoff "accepts" and Angoff "rejects" (AASRS3 with ARSRS3). Angoff "accepts" achieved a .6250 mean gain, from 4.7500 to 5.3750, representing a t value of 2.53, significant at p < .013. The results are presented in Table 21, page 105.

This t value indicated there was a statistically significant difference in the supervisory rating scores on soldier performance in following established regulations and orders received by Angoff "accepts" and Angoff "rejects."

AASRS4 with ARSRS4

A t-test was used to determine if there was a statistically significant difference in the supervisory rating scores on soldier performance in maintaining assigned equipment received by Angoff "accepts" and
### TABLE 20

PART 3 OF HYPOTHESIS 4 — T-TEST OF SIGNIFICANT DIFFERENCE IN SUPERVISORY RATING SCORES CONCERNING JOB EFFORT OF ANGOFF "ACCEPTS" AND ANGOFF "REJECTS"

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>NUMBER OF CASES</th>
<th>MEAN</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>F VALUE</th>
<th>2-TAIL PROB.</th>
<th>POOLED VARIANCE ESTIMATE</th>
<th>T VALUE</th>
<th>DEGREES OF FREEDOM</th>
<th>2-TAIL PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASRS2 AND ARSRS2 EFFORT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GROUP 1</td>
<td>64</td>
<td>5.2813</td>
<td>1.147</td>
<td>0.143</td>
<td>1.75</td>
<td>0.052</td>
<td></td>
<td>3.21</td>
<td>98</td>
<td>0.002</td>
</tr>
<tr>
<td>GROUP 2</td>
<td>36</td>
<td>4.4167</td>
<td>1.519</td>
<td>0.253</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**TABLE 21**

PART 3 OF HYPOTHESIS 4 - T-TEST OF SIGNIFICANT DIFFERENCE IN SUPERVISORY RATING SCORES CONCERNING ANGOFF "ACCEPTS" AND ANGOFF "REJECTS" ADHERENCE TO JOB PROCEDURES

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>GROUP 1 - CTS</th>
<th>GROUP 2 - CTS</th>
<th>MEAN</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>F VALUE</th>
<th>2-TAIL PROB.</th>
<th>POOLED VARIANCE ESTIMATE</th>
<th>T VALUE</th>
<th>DEGREES OF FREEDOM</th>
<th>2-TAIL PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASRS3 AND ARSRS3 FOLLOW PROCEDURES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Angoff "rejects" (AASRS4 with ARSRS4). Angoff "accepts" achieved a .5139 mean gain, from 4.6111 to 5.1250, representing a $t$ value of 1.75, at $p < .084$ level of confidence. The results are presented in Table 22.

This $t$ value indicated that the supervisory rating scores received by Angoff "accepts" and Angoff "rejects" concerning maintenance of equipment were not significantly different at $p < .05$ established for this study.

Major Findings

Hypothesis 4 was tested for the strength of the linear relationship of classroom test scores and supervisory rating scores received by Angoff "accepts" and Angoff "rejects." This assessment compared the sizes of the two $r$'s for one of four sets of correlations relative to four areas of soldier job performance. This comparison concerned the maintenance of assigned equipment where the $r$'s were statistically significant. Each correlation coefficient for Angoff "accepts" was significant at $p < .001$; one correlation coefficient for Angoff "rejects," with negative direction (-.3928 for ARCTS with ARSRS4), was significant at $p < .009$ level of confidence.

On the basis of these differences in $r$ values, it was determined there was a difference in the correlations between classroom test scores and supervisory rating scores of Angoff "accepts" and Angoff "rejects."

Hypothesis 4 was also tested for significant differences in the means of supervisory rating scores received by Angoff "accepts" and Angoff "rejects." From this four-part test, it was concluded that the differences were statistically significant in three of four areas of job
TABLE 22

PART 3 OF HYPOTHESIS 4 - T-TEST OF SIGNIFICANT DIFFERENCE IN SUPERVISORY RATING SCORES CONCERNING ANGOFF "ACCEPTS" AND ANGOFF "REJECTS" ON EQUIPMENT MAINTENANCE

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>NUMBER OF CASES</th>
<th>MEAN</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>F 2-TAIL</th>
<th>T VALUE</th>
<th>DEGREES OF FREEDOM</th>
<th>2-TAIL PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>POOLED VARIANCE ESTIMATE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AASRS4 AND ARSRS4 MAINTENANCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GROUP 1</td>
<td>64</td>
<td>5.1250</td>
<td>1.303</td>
<td>0.163</td>
<td>1.49</td>
<td>0.168</td>
<td>1.75</td>
<td>98</td>
</tr>
<tr>
<td>GROUP 2</td>
<td>36</td>
<td>4.6111</td>
<td>1.591</td>
<td>0.265</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
performance: AASRS1 with ARSRS1, $r = 3.61$, $p < .001$ (technical knowledge/skill); AASRS2 with ARSRS2, $r = 3.21$, $p < .002$ (effort); and AASRS3 with ARSRS3, $r = 2.53$, $p < .013$ (following regulations and orders).

On the basis of the strengths of the $r$ values obtained, the hypothesis that there would be no difference in the correlation between the classroom test scores and the supervisory rating scores received by Angoff "accepts" and Angoff "rejects" was rejected. On the strength of the $r$ values obtained, the hypothesis that there would be no significant difference in the mean of supervisory rating scores received by Angoff "accepts" and Angoff "rejects" was also rejected.

Hypothesis 5

Hypothesis 5 states there will be no difference in the correlation between classroom test scores and skill qualification test scores received by Angoff "accepts" and Angoff "rejects"; and there will be no significant difference in the means of the skill qualification test scores received by Angoff "accepts" and Angoff "rejects."

The test of Hypothesis 5 was conducted in three parts. Part 1 assessed the strength of the linear relationship between soldier classroom test scores and skill qualification test scores of 49 job performers referred to as Angoff "accepts," soldiers who achieved classroom test scores of 92 through 100 percent. Part 2 assessed the strength of the linear relationship between soldier classroom test scores and skill qualification test scores of 17 job performers referred to as Angoff "rejects," soldiers who achieved classroom test scores from 85 through 91 percent. Part 3 assessed the difference in means of skill qualification test scores of Angoff "accepts" and Angoff "rejects."
The Pearson Product-Moment correlation, significant at \( p < .05 \), was used in the test of this hypothesis to assess the strength of the linear relationship between classroom test scores and skill qualification test scores. The t-test, significant at \( p < .05 \), was used to assess the statistical significance of the differences in the means of the skill qualification test scores of Angoff "accepts" and Angoff "rejects."

**Part 1 - Test of Hypothesis 5**

**AACTS with AASQT**

The Pearson Product-Moment coefficient of correlation computed for soldier classroom test scores and skill qualification test scores (AACTS with AASQT) for Angoff "accepts" makes use of the same study sample and the comparison that was made in Part 1 of Hypothesis 3, page 93. This computation produced an \( r \) value of .3686, significant at \( p < .005 \) level of confidence:

\[
\begin{align*}
\text{r} &= .3686 \\
\text{n} &= 49 \\
\text{p} &< .005
\end{align*}
\]

This correlation coefficient indicated that the relationship of AACTS and AASQT was statistically significant for Angoff "accepts."

**Part 2 - Test of Hypothesis 5**

**ARCTS with ARSQT**

A Pearson Product-Moment coefficient of correlation was computed for soldier classroom test scores and skill qualification test scores (ARCTS with ARSQT) for Angoff "rejects." This computation produced an \( r \) value of .3148, at \( p < .109 \) level of confidence:
This correlation coefficient indicated that the relationship between ARCTS and ARSQT was not statistically significant for Angoff "rejects."

For Parts 1 and 2 of Hypothesis 5, the strength of the relationship of the r values for Angoff "accepts" and Angoff "rejects" was not assessed. The correlation coefficient of .3686 for Angoff "accepts" was statistically significant at \( p < .005 \), while the r value for Angoff "rejects," \( r = .3148, p < .109 \), was not statistically significant. A comparison of these r's is therefore considered invalid. Based on the sizes of the two r's, it was concluded there is a difference in correlations between classroom test scores and skill qualification test scores of Angoff "accepts" and Angoff "rejects."

Part 3 - Test of Hypothesis 5

Part 3 of the test of Hypothesis 5 involved the use of the t-test (\( p < .05 \)) to determine the significance of the difference between the means of skill qualification test scores received by Angoff "accepts," soldiers who achieved classroom test scores from 92 through 100 percent, and Angoff "rejects," soldiers who achieved classroom test scores from 85 through 91 percent.

A t-test was used to determine if there was a statistically significant difference in the means of the skill qualification test scores received by Angoff "accepts" and Angoff "rejects." Angoff "accepts"
achieved a 9.382 mean gain, from 71.353 to 80.735, representing a $t$ value of 3.73, significant at $p < .001$. The results are presented in Table 23.

This $t$ value indicated there was a statistically significant difference in the skill qualification test scores received by Angoff "accepts" and Angoff "rejects."

**Major Findings**

The test of Hypothesis 5 assessed the strength of the linear relationship of classroom test scores and skill qualification test scores received by Angoff "accepts" and Angoff "rejects." The correlation coefficient for Angoff "accepts" was $r = .3686$, significant at $p < .005$. In contrast, the comparable correlation coefficient for Angoff "rejects" was $r = .3148$, at $p < .109$ level of confidence. The strength of the relative sizes of the $r$'s for the two sets of job performers was therefore not assessed. A comparison of these $r$'s was considered invalid. On the basis of the differences in $r$ values, it was concluded there was a difference in the correlations between classroom test scores and skill qualification test scores of Angoff "accepts" and Angoff "rejects."

The test of Hypothesis 5 also assessed the significant differences in the means of skill qualification test scores received by Angoff "accepts" and Angoff "rejects." From this test, which produced a $t$ value of 3.73, significant at $p < .001$, it was concluded that the differences were statistically significant.

On the basis of the strengths of the $r$ values and $t$ values obtained, the hypothesis that there would be no significant differences
TABLE 23

PART 2 OF HYPOTHESIS 5 - T-TEST OF SIGNIFICANT DIFFERENCE IN
SKILL QUALIFICATION TEST SCORES OF ANGOFF "ACCEPTS" AND ANGOFF "REJECTS"

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>NUMBER OF CASES</th>
<th>MEAN</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
<th>F VALUE</th>
<th>2-TAIL PROB.</th>
<th>POOLED VARIANCE ESTIMATE</th>
<th>T VALUE</th>
<th>DEGREES OF FREEDOM</th>
<th>2-TAIL PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASQT AND ARSQT PRESENT SKILL TEST SCORES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GROUP 1</td>
<td>49</td>
<td>80.7347</td>
<td>8.149</td>
<td>1.164</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>3.73</td>
<td>64</td>
</tr>
<tr>
<td>GROUP 2</td>
<td>17</td>
<td>71.3529</td>
<td>10.926</td>
<td>2.650</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant at \( p < .001 \)
in the supervisory rating scores received by Angoff "accepts" and Angoff "rejects" was rejected.

Summary: Probability Estimates and Tests of Hypotheses

For Hypothesis 1, intragroup variability and interrater reliability of probability estimates were assessed. Through a two-way analysis of variance, it was determined there was a significant difference in probability estimates of the judges ($F = 4.542, p < .00001$). The null hypothesis that the probability estimates by the test item judges would not be significantly different was rejected.

Interrater reliability of the probability estimates of the judges' average rating for each test item was assessed using the variance components from the analysis of variance of the probability estimates. This assessment measured the extent to which the estimates covaried (correlated) across items and among judges. From the analysis it was determined that interrater reliability was significant ($r = .780, p < .01$), and the null hypothesis was rejected.

For Hypothesis 2, the strengths of the linear relationship between classroom test scores and supervisory rating scores for four categories of soldier job performance (technical knowledge/skill, effort, follow procedures, and maintenance), based on the Angoff score-setting method, were determined to be significant ($r = .8724, r = .6841$, $r = .6728$, and $r = .6543$, each at $p < .001$ level of confidence). The assessments were based on test score range of 92 through 100 percent.

The strengths of the linear relationship between classroom test scores and supervisory rating scores for the four categories of soldier job performance were also assessed relative to the present score-setting
method. The assessments were based on the test score range of 85 through 100 percent. From the analysis it was determined that a significant relationship existed for each of the four categories of soldier job performance ($r = .5191$, $r = .3980$, and $r = .3542$, significant at $p < .001$; and $r = .2476$ at $p < .006$).

A comparison of the relative sizes of the two sets of $r$'s for each of the four sets of correlations showed a stronger relationship of the scores under the Angoff method. The null hypothesis that there would be no difference in the correlation coefficients for the two score-setting methods was rejected.

For Hypothesis 3, the strength of the linear relationship between student classroom test scores and skill qualification test scores based on the Angoff and present score-setting methods was assessed. From this analysis it was determined that a statistically significant relationship existed between the sets of scores for both methods, but the strength of the coefficient for the present method ($r = .5282$, $p < .001$) was greater than the coefficient for the Angoff method ($r = .3686$, $p < .005$). The null hypothesis that there would be no difference in the correlation coefficients for the two score-setting methods was rejected.

For Hypothesis 4, the strengths of the linear relationship between classroom test scores and supervisory rating scores for four categories of soldier job performance (technical knowledge/skill, effort, follow procedures, and maintenance) for Angoff "accepts" and Angoff "rejects" were assessed; also, the test of the significance of the difference in the means of the supervisory rating scores of the two groups
of job performers was made. The assessments used classroom test score ranges of 92 through 100 percent for Angoff "accepts" and 85 through 91 percent for Angoff "rejects." From the analysis, it was determined that each correlation coefficient for Angoff "accepts" was significant (same as for Hypothesis 2), but only one correlation coefficient for Angoff "rejects," with negative direction \( r = -0.3928, p < 0.009 \), was significant.

The test for significance of difference in the means of the supervisory rating scores on each of the four categories of soldier job performance received by Angoff "accepts" and Angoff "rejects" was made. From the application of the \( t \)-test it was determined there were statistically significant differences in the scores for three of four categories of soldier job performance: \( t = 3.61, p < 0.001 \) (technical knowledge/skill); \( t = 3.21, p < 0.002 \) (effort); and \( t = 2.53, p < 0.013 \) (following regulations and orders).

On the basis of the \( r \) values and \( t \) values obtained, the null hypothesis was rejected. The study concluded there was a difference in the correlation coefficients for classroom test scores and supervisory rating scores as well as a significant difference in the means of supervisory rating scores received by Angoff "accepts" and Angoff "rejects."

For Hypothesis 5, the strength of the linear relationship between classroom test scores and skill qualification test scores for Angoff "accepts" and Angoff "rejects" was assessed; also, the test of the significance of the difference in the means of the skill qualification test scores of the two groups of job performers was made. This
assessment was based on classroom test score ranges of 92 through 100 percent for Angoff "accepts" and 85 through 91 percent for Angoff "rejects." From the analysis it was determined that the correlation coefficient for the Angoff "accepts" was statistically significant ($r = .3686, p < .005$), but the $r$ value for the Angoff "rejects" was not significant ($r = .3148, p < .109$).

The test for significance of difference in the means of the skill qualification test scores received by Angoff "accepts" and Angoff "rejects" was made. From the application of the $t$-test it was determined there was a significant difference in the means of the skill qualification test scores for the two categories of job performers ($t = 3.73, p < .001$).

On the basis of the $r$ values and $t$ values obtained, the null hypothesis was rejected. The study concluded there was a difference in the correlation coefficients for classroom test scores and skill qualification test scores as well as a significant difference in the means of skill qualification test scores received by Angoff "accepts" and Angoff "rejects."
CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This chapter discusses the lack of suitable criteria for validating criterion-referenced tests; the theory base relative to criterion-referenced test validation and test cutoff score computation; the methodology for determining test content validity and setting cutoff scores for criterion-referenced tests using the empirically based Angoff cutoff score-setting method; and the predictive validity of a criterion-referenced classroom test with cutoff scores set by the Angoff method and present method, based on the empirical relationship between examinee test scores achieved and their job performance effectiveness; and makes recommendations for future research. The following limitations should be taken into account in the interpretation of the results of this study: test item judges' possible lack of understanding of the concept of job performer minimal competency, lack of control for maximum job time of performers whose job skills were rated, and the limited direct application of the study findings and recommendations to other courses taught at the United States Army Quartermaster School.

Problem

A relatively new method of assessing student achievement, called criterion-referenced measurement, was developed to assess student performance based on a specific criterion or standard of performance and
domain of behavior. Traditional measures for estimating test validity are based primarily on variability among student scores and are, therefore, appropriate for norm-referenced measurement but not for criterion-referenced testing (Gronlund, 1985). Because there are no specific bases for setting the ideal level or cutoff score for a criterion-referenced test, this level has usually been set according to the expectations of the teacher or the school (Scannel & Tracy, 1975). Swezey (1981) noted that the basis for the criterion-referenced measurement model is to distinguish masters from non-masters on their performance on tests, to avoid assigning an incompetent individual to a critical job and to preclude huge wastes of money, manpower, and time that could occur if students were erroneously labeled. Scannel and Tracy (1975) stressed the need for a procedure for setting precise cutoff scores for criterion-referenced tests. Harasym (1981) noted there are procedures available for setting these standards, but little is known about their consistency. Hambleton (1980) called for more empirical investigations into the factors to be considered in selecting the proper standard-setting procedure for testing. The present study was conducted to develop a means to measure examinee performance relative to a clearly defined criterion (cutoff score) and domain of behavior.

Review of Literature

Over the years authors have advocated the use of different methods for measurement of student performance. Haak (1960) recommended a systematic adjustment of the student's summative grade by applying an adjustment factor for deciding the average grade. Hinely (1968) offered a handicapping formula for crediting the student with
progress made, calculated as part of the student's last test score. Cox (1971) recommended an index of discrimination based on the relative number of students passing an item at the beginning and conclusion of a unit of instruction. Popham (1971) asserted there were certain tasks which must be performed to a very high level, and where such tasks are critical the test should be criterion-referenced. Glaser and Nitko (1971) noted that a criterion-referenced test is "one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Andrews and Hecht (1976) pointed out that most standard-setting methods for criterion-referenced measurement are designed in relation to minimum acceptable competency on the part of the examinee.

Many authors (Nedelsky, 1954; Angoff, 1971; Ebel, 1972; Zieky, 1973; Payne, 1974; Buck, 1975; Rovinelli & Hambleton, 1976; and Popham, 1981) have called for the use of expert judgment in validating test items and setting cutoff scores. Four theorists, Nedelsky (1954), Angoff (1971), Ebel (1972), and Jaeger (1978), provided methodologies for determining content validity and establishing test cutoff scores using content judges.

Andrews and Hecht (1976) compared the Ebel and Nedelsky methods, using multiple-choice questions from a nationally administered health professions' examination. From the study it was concluded that the different groups of judges set similar standards for the same test but the two procedures yielded significantly different passing scores. Poggio, Glasnap, and Eros (1981) compared the Angoff, Ebel, and Nedelsky methods, including reliabilities and validities of the judges' ratings.
The comparisons, which involved 10 reading and mathematics tests for grade levels 2, 4, 6, 8, and 11, produced greater variability in ratings among the tests for the Angoff method. The Ebel method produced the highest score in most cases using means as the standard. The Nedelsky method produced the highest agreement coefficient between passing scores and test results. Cross, Impara, Frary, and Jaeger (1984) described a 1982 pilot study which compared the Jaeger, Nedelsky, and Angoff methods in determining the validity of using the National Teacher Examinations for initial certification of teachers in the state of Virginia. From the study it was concluded that the Angoff judgments surpassed those produced by the other methods. Cross (1982, 1984) reported on a two-phase study of the Angoff method for use in determining the validity of using the National Teacher Examinations for certifying entry-level teachers in Virginia. The study reported validity estimates for core battery tests ranging from strong to substantial, while the estimates for the area examinations ranged from strong to moderate. The study concluded that the test cutoff scores could be used to measure accurately and reliably the minimum competency of the examinee. Jaeger and Busch (1984) reported on the use of the Jaeger method for determining the validity and appropriate standards for Core Battery Tests I and II of the National Teacher Examinations "for screening applicants for admission to North Carolina's teacher education program" (p. 3). The study concluded that the levels of relevance for the various tests ranged from marginal to acceptable, that "continued use of the tests demands an ongoing program of review and research" (p. 22).
Based on the studies conducted, it has been concluded that reliable examination standards could be set by employing the judgments of practitioners. The judgments are subjective and the evaluation outcome is dependent upon the standard-setting procedure used. According to Van Der Linden (1982), "the Angoff ... technique has become one of the best known and widely used method of standard setting" (p. 296).

Methodology

Study Questions

The present procedure for validating a classroom test was developed for use with norm-referenced testing. This study has examined an empirical means to set performance standards, or test cutoff scores, for criterion-referenced classroom tests, utilizing the Angoff cutoff score-setting method. It has also addressed the following research questions:

1. What are the level of intragroup variability and the difference in the interrater reliability estimates of instructors and job specialists in their assessment of the minimum passing score for a criterion-referenced classroom test?

2. What is the relationship of job effectiveness (supervisory rating and skill qualification test scores) to classroom test cutoff scores for two groups of job performers whose criterion-referenced test cutoff scores were set by an empirically based method and the present method founded on instructor consensus?
Relative emphasis assessment

The first assessment of test content validity was made by computing an indices of relative emphasis for each test content topic. Eight randomly selected faculty members assessed the correspondence between the course curriculum and the test content topics, specifying whether the emphasis in the curriculum was the "same," "more," or "less" than the emphasis in the test. The proportion of the reviewers who selected each option was calculated for each test topic. These proportions were, in turn, multiplied by the percent of the items on the test devoted to each topic. The products were summed, providing a weighted index in the range of 0 to 100 which shows the relative emphasis given to each topic and the overall test.

The emphasis estimates for the individual test topics revealed that two of the test topics are given significantly more emphasis in the curriculum than in the test, as shown in Table 1, page 52. The indices for the overall test revealed a relative emphasis of 76.01 for the "same" category, 9.24 for the "more" category, and 14.75 for the "less" category, as shown in Table 2, page 53. Analysis of these results revealed a difference of 5.51 between the "more" and "less" indices for the overall tests. According to Cross (1982), a [norm-referenced] test is identified for further study "if the 'less' index exceeds 25, or . . . if the 'less' index exceeds the 'more' index by 5 points or more" (p. 73). The U. S. Army Quartermaster School's standard for criterion-referenced tests calls for a perfect correspondence between the curriculum and the test. Based on the indices computed, there is a strong
indication that the course embodies some material that is not covered by the test, raising some doubt concerning the test's content validity. The results also indicate that the test contains some material that was not taught in the course, thereby threatening the instructional validity of the course content. Since curricular validity and instructional validity are two kinds of evidence of test content validity, these findings indicate there is a need to enhance the correspondence between the curriculum and the test.

**Job relevance assessment**

The second step in determining test content validity concerned the assessment of the job relevance of each test item. For this assessment, a group of 72 test item judges, comprised of 36 instructors from the faculty of the U.S. Army Quartermaster School and 36 job specialists from Army units in Germany and the Continental United States, estimated the job relevance of each test item of the 380-item, criterion-referenced, end-of-course comprehensive test for the Equipment Records and Parts Specialist Course. Using a 4-point rating scale, each judge assessed whether the item was equal to 1 (not relevant), 2 (questionable), 3 (important), or 4 (crucial). Any question with an average relevance rating of 2.25 or below received a value of zero in the minimum passing score computation.

The average relevance estimate for each item exceeded the minimum acceptable relevance value of 2.25. The range of average relevance values for all raters across all test items was 3.042 to 3.833, a spread of .791. Based on the average relevance values, 368 of 380 test items,
or 96.84 percent of the test, fell into the 3,300 to 3,900 relevance range, as shown in Table 3, page 56. Overall, a strong relationship exists between the test items and the domain of knowledge that the entry-level Equipment Records and Parts Specialist should possess. These results confirm the validity of the test items involved in the high "less" index in the indices of relative emphasis estimates for the course and the test. From this finding it is concluded that the higher level of correspondence between course content and test should be achieved by revising the course content, not the test.

Test Cutoff Score Computation

A raw score performance standard was computed using the judges \((n = 72)\) probability estimates that a minimally competent individual would be able to answer each question correctly without guessing. Using a 1-to-10 scale, each judge rendered a probability estimate for each of the 380 items on the test. The judges' probability estimates were converted to probability values and summed and averaged, yielding a raw score performance standard of 349.25, with a standard deviation of 17.44 and an intrarater reliability correlation coefficient of .780 based on the judges' average ratings on the individual test items. The raw score standard was, in turn, converted to a percentage score standard because the U. S. Army Quartermaster School's grading system is based on percentage scores. The percentage score standard calculated was 91.91, rounded to 92 percent, with a standard deviation of 4.59 and an interrater reliability correlation coefficient of .780 for the judges' average ratings on individual items, the same as the coefficient for the raw score standard computation.
The percentage score standard was adjusted for possible chance error in the score computation to preclude the possible labeling of a master as a non-master. The chance error, labeled composite standard error, consisted of sampling error associated with the selection of the test item judges and measurement error which concerned the performance of the examinee. The sampling error was calculated at 0.54 percent. The measurement error was 1.897 percent. The composite standard error was therefore 1.97 percent. The adjusted percentage score standards, based on one, two, and three standard errors, are 89.94, 87.97, and 86.00, rounded to 90, 88, and 86 percent, respectively.

From this study it was concluded there is a significant difference in the test cutoff scores set by the Angoff method (92 percent) and present cutoff score-setting method (85 percent). The possibility therefore exists that some course graduates have been mislabeled concerning the level of job entry skills which they possess.

Student attrition rates for calendar year 1985, based on the derived percentage score standard of 92 percent and adjustments for one, two, and three standard errors, as shown in Table 11, page 70, follow: 800 students, or 36.60 percent of the annual enrollment, would have failed the course based on the derived percentage score standard of 92; 447 students, or 20.45 percent of the annual enrollment, would have failed the course if the derived passing score were adjusted for one standard error; 196 students, or 8.97 percent of the annual enrollment, would have failed the course if the derived passing score were adjusted for two standard errors; and 13 students, or 0.59 percent of the annual
enrollment, would have failed the course if the derived passing score were adjusted for three standard errors. Application of the new performance standard and any adjustment to the standard to compensate for possible chance error should be made only after careful consideration of the predictive validity measurement of the test. This subject is presented in the discussion of Hypotheses 1 through 5 which follows.

Null Hypotheses

This study examined the means to assess empirically the validity of a criterion-referenced test. Traditional measures for estimating test validity are based primarily on variability among student scores, and they are therefore inappropriate for criterion-referenced testing (Gronlund, 1985). Because there are no specific bases for setting the cutoff score for a criterion-referenced test, the standard has usually been set according to the expectations of the teacher or the school (Scannel & Tracy, 1975). The following null hypotheses concern intragroup variability of probability estimates of test item judges, interrater reliability of the judges' estimates, and the predictive validity of a criterion-referenced classroom test with cutoff scores set by the Angoff method and the present method. Each hypothesis was tested for significance at p<.05.

Hypothesis 1

This two-part hypothesis states there will be no significant difference in intragroup variability of the probability estimates by instructors and job specialists in their assessment of the minimum passing score for a criterion-referenced classroom test; and the interrater
reliability of instructors' and job specialists' probability judgments will not be statistically significant. From the statistical analysis of the performance of the test item judges, Parts 1 and 2 of this hypothesis were rejected.

1. A two-way analysis of variance indicated there was significant intragroup variability in the probability estimates rendered by the judges. The analysis resulted in an $F$ ratio of 4.542 with 379 and 71 degrees of freedom. The difference was significant at $p < .0001$. The study concluded that the backgrounds and experiences of the all rater category of test item judges affected significantly their estimates concerning the capability of a minimally competent job performer. This finding indicated there was a good cross-section of personnel in terms of job experience involved in this study and because of the broad spectrum of experience valid samples were evident.

2. The interrater reliability coefficient of correlation obtained for the performance of the all rater category of test item judges, which reflects the extent to which the probability estimates covary (correlate) across items and among reviewers, was .780, significant at $p < .01$. The 95 percent confidence limits obtained for the $r$ value of .780 were .72 and .83. It can be asserted with 95 percent confidence that the population value lies within these limits. This study concluded there is a high degree of interrater reliability among the instructor and job specialist panel members in their assessment of the job skills of a minimally competent job performer. Specifically, there is a chance that 19 out of 20 panel results with a correlation coefficient of .780 would fall within the calculated confidence limits.
Based on the interrater reliability coefficient obtained, it was concluded that the concept of minimal competence was understood by the panel members, that job specialists were technically competent to serve as test item judges, and that the results have a high degree of credibility.

Hypothesis 2

This two-part hypothesis states there will be no difference in the correlation between classroom test scores and supervisory rating scores of job performers whose classroom test cutoff scores were set by the Angoff method and the present method. Part 1 assessed the strength of the linear relationship between the classroom test scores and the supervisory rating scores of job performers who received a test score in the range of 92 through 100 percent, based on the Angoff score-setting method. Part 2 assessed the linear relationship between the classroom test scores and supervisory rating scores of job performers who received a test score in the range of 85 through 100 percent, based on the present score-setting method. This hypothesis tested the relationship between classroom test scores and supervisory rating scores concerning four areas of job performance: technical knowledge/skill, effort, following regulations and orders, and maintaining assigned equipment. A description of the variable labels for the two score-setting methods and the four areas of job performance is presented in Chart 2, page 86.

This hypothesis was not tested statistically in the classical sense. Instead, it assessed the similarity of the correlations of test scores and supervisory rating scores based on the two score-setting methods by comparing the relative sizes of the sets of r's. Based on
the magnitude of the difference between the sets of correlation coefficients for each area of job performance, this hypothesis was rejected.

1. The Pearson Product-Moment correlation coefficients obtained for Part 1 of this hypothesis indicated that the linear relationships of the Angoff-derived scores and the supervisory rating scores for the four areas of job performance were statistically significant at $p < .001$.

The $r$ values obtained were .8724 (AACTS with AASRS1), .6841 (AACTS with AASRS2), .6728 (AACTS with AASRS3), and .6543 (AACTS with AASRS4).

2. The Pearson Product-Moment correlation coefficients obtained for Part 2 of the test of this hypothesis indicated that the linear relationships of the classroom test scores and the supervisory rating scores based on the present cutoff score-setting method were statistically significant. The relationships for the first three sets of scores were significant at $p < .001$. The relationship for the fourth set was significant at $p < .006$. The $r$ values obtained were .5191 (PMCTS with PMSRS1), .3980 (PMCTS with PMSRS2), .3542 (PMCTS with PMSRS3), and .2476 (PMCTS with PMSRS4).

The strength of the difference in $r$ values for each of the four sets of correlations was high. In each case the strength of the relationship of the scores relative to the Angoff method exceeded the strength of the relationship of the scores under the present method. On the basis of these differences the study concluded that student achievement on the classroom test with the cutoff score set by the Angoff method is a better predictor of job performance effectiveness than is performance on the test with the cutoff score set by the present method.
This study also concluded that soldiers whose classroom achievement was assessed based on the Angoff-derived cutoff score performed their job tasks more effectively, more closely paralleling supervisors' expectations, than did the soldiers whose achievement was based on the test cutoff score set by the conventional method.

Hypothesis 3

This two-part hypothesis states there will be no difference in the correlation between classroom test scores and skill qualification test scores of job performers whose classroom test cutoff scores were set by the Angoff method and the present method. Part 1 of the test of this hypothesis assessed the strength of the linear relationship between the classroom test scores and skill qualification test scores of job performers who received a classroom test score in the range of 92 through 100 percent, based on the Angoff-derived cutoff score of 92. Part 2 assessed the strength of the linear relationship between the classroom test scores and skill qualification test scores of job performers who received a classroom test score in the range of 85 through 100 percent, based on the present test cutoff score of 85.

This hypothesis was not tested statistically in the classical sense. Instead, it assessed the similarity of the correlation of classroom test scores and skill qualification test scores derived by the two score-setting methods by comparing the relative sizes of the two r's. Based on the magnitude of the difference between the correlation coefficients for the two samples, this hypothesis was rejected.

1. The Pearson Product-Moment correlation coefficient obtained for Part 1 of the test of this hypothesis indicated that the strength of
the linear relationship between the classroom test scores and skill qualification test scores based on the Angoff score-setting method was statistically significant, at $p < .005$ level of confidence. The $r$ value was .3686 (AACTS with AASQT).

2. The Pearson Product-Moment correlation coefficient obtained for Part 2 of the test of this hypothesis indicated that the strength of the linear relationship between the classroom test scores and skill qualification test scores based on the present score-setting method was statistically significant, at $p < .001$ level of confidence. The $r$ value was .5282 (PMCTS with PMSQT).

The difference in the $r$ values for the two correlations was not very strong. The $r$ value for the relationship of the scores relative to the present score-setting method was .5282 compared to .3686 for the scores under the Angoff score-setting method. One should exercise caution in drawing the conclusion that student scores on the classroom test with cutoff score set by the present method are a better predictor of job performer achievement on the skill qualification test. This precaution is supported by the fact noted under Procedures, page 39, that reliability of the skill qualification test for the PLL clerk was not assessed in the test developmental process. In further support of this precaution, it was determined in the data gathering process for this study that skill qualification test scores were not available for many eligible PLL clerks. For this reason the study sample used for the test of this hypothesis may not have been a valid one.
Hypothesis 4

This three-part hypothesis states there will be no difference in the correlation between classroom test scores and supervisory rating scores received by Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) and Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard); and there will be no significant difference in the means of supervisory rating scores received by Angoff "accepts" and Angoff "rejects." Part 1 assessed the strength of the linear relationship between the classroom test scores and supervisory rating scores of job performers referred to as Angoff "accepts," soldiers who achieved classroom test scores in the range of 92 through 100 percent. Part 2 assessed the strength of the linear relationship between classroom test scores and supervisory rating scores of job performers referred to as Angoff "rejects," soldiers who achieved classroom test scores in the range of 85 through 91 percent. Part 3 assessed the difference in means of supervisory rating scores of Angoff "accepts" and Angoff "rejects."

This hypothesis tested the relationship of classroom test scores and supervisory rating scores concerning four areas of job performance: technical knowledge/skill, effort, following regulations and orders, and maintaining assigned equipment.

Parts 1 and 2 of the test of this hypothesis were not tested statistically in the classical sense. Instead, the test of these parts of the hypothesis assessed the similarity of the correlations of the classroom test scores and supervisory rating scores of the two samples of job performers, Angoff "accepts" and Angoff "rejects," by comparing
the relative sizes of the sets of statistically significant r's. Part 3 was tested for statistical significance by applying the r-test to assess the difference in means in supervisory rating scores in the four areas of job performance for Angoff "accepts" and Angoff "rejects." Based on the magnitude of the difference between the sets of correlation coefficients and the difference in means of the supervisory rating scores for each area of job performance, this hypothesis was rejected.

1. The Pearson Product-Moment correlation coefficients obtained for Part 1 of the test of this hypothesis indicated that the linear relationships between the classroom test scores and supervisory rating scores of Angoff "accepts" for each of the four areas of job performance were statistically significant at \( p < .001 \). The r values obtained were .8724 (AACS with AASRS1), .6841 (AACS with AASRS2), .6728 (AACS with AASRS3), and .6543 (AACS with AASRS4). This study sample and correlation coefficients presented are the same as those reflected in Part 1 of the test of Hypothesis 2 (pages 85 through 89).

2. The Pearson Product-Moment correlation coefficients obtained for Part 2 of the test of this hypothesis indicated that the linear relationships of the classroom test scores and supervisory rating scores of Angoff "rejects" were not statistically significant in three of the four areas of soldier job performance examined. The r values for the three areas in which the relationships were not significant are .0067, \( p < .485 \) (ARCTS with ARSRS1); -.2106, \( p < .109 \) (ARCTS with ARSRS2); and -.1728, \( p < .148 \) (ARCTS with ARSRS3). The r value for the one area of soldier job performance in which the relationship was significant, in
this case a negative coefficient of correlation, was \(-.3928, p < .009\) (ARCTS with ARSRS4).

3. The \(t\)-test used in Part 3 of the test of this hypothesis indicated that Angoff "accepts" achieved the following mean gains on supervisory rating scores: .9271 (from 4.4167 to 5.3438 for AASRS1 with ARSRS1); .8646 (from 4.4167 to 5.2813 for AASRS2 with ARSRS2); .6250 (from 4.7500 to 5.3750 for AASRS3 with ARSRS3); and .5139 (from 4.6111 to 5.1250 for AASRS4 with ARSRS4). There was a statistically significant difference in the means of supervisory rating scores for Angoff "accepts" and Angoff "rejects" in three of the four areas of soldier job performance examined. The \(t\) values for the three areas in which there were statistically significant differences in the means of supervisory rating scores were 3.61, \(p < .001\) (AASRS1 with ARSRS1); 3.21, \(p < .002\) (AASRS2 with ARSRS2); and 2.53, \(p < .013\) (AASRS3 with ARSRS3). The \(t\) value for the area of soldier job performance in which the difference was not significant was 1.75, \(p < .084\) (AASRS4 with ARSRS4).

4. The relative strengths of the differences in \(r\) values were disregarded for the three areas of job performance (technical knowledge/skill, effort, and following regulations and orders) in which the results for Angoff "rejects" were not significant. The relative strength of the difference in \(r\) values pertaining to maintenance of assigned equipment was .6543 compared to \(-.3928\). Based on the \(r\) values obtained for Angoff "rejects" and the \(r\) value obtained on the difference in means in supervisory rating scores concerning soldier performance on maintenance of assigned equipment, there is a strong indication that the
supervisors were not capable of detecting the differences in job performance of soldiers who received classroom test scores in the restricted grade ranges of 85 through 91 percent. There is also the likelihood that the soldier who received a classroom test score of 85 percent performed as effectively on the job as did the soldier who received a score of 91 percent. The study therefore concluded that classroom achievement of Angoff "accepts" is a better predictor of soldier job performance effectiveness than is the classroom achievement of Angoff "rejects."

Hypothesis 5

This three-part hypothesis states there will be no difference in the correlation between classroom test scores and skill qualification test scores received by Angoff "accepts" (job performers whose classroom test scores met or exceeded the Angoff-derived standard) and Angoff "rejects" (job performers whose classroom test scores fell below the Angoff-derived standard); and there will be no significant difference in the means of skill qualification test scores received by Angoff "accepts" and Angoff "rejects." Part 1 assessed the strength of the linear relationship between soldier classroom test scores and skill qualification test scores of job performers referred to as Angoff "accepts," soldiers who achieved classroom test scores in the range of 92 through 100 percent. Part 2 assessed the strength of the linear relationship between soldier classroom test scores and skill qualification test scores of job performers referred to as Angoff "rejects," soldiers who achieved classroom test scores in the range of 85 through 91 percent. Part 3 assessed the difference in means of skill qualification test scores of Angoff "accepts" and Angoff "rejects."
Parts 1 and 2 of the test of this hypothesis were not tested statistically in the classical sense. Instead, the test of these parts of the hypothesis assessed the similarity of the correlations of the classroom test scores and skill qualification test scores of the two samples of job performers, Angoff "accepts" and Angoff "rejects."

Part 3 was tested for statistical significance by applying the $t$-test to assess the difference in means in skill qualification test scores of Angoff "accepts" and Angoff "rejects."

Based on the magnitude of the difference between the correlation coefficients and the difference in means of the skill qualification test scores, this hypothesis was rejected.

1. The Pearson Product-Moment correlation coefficient obtained for Part 1 of the test of this hypothesis indicated that the linear relationship between the classroom test scores and skill qualification test scores of Angoff "accepts" (AACS with AASQ) was statistically significant, with an $r$ value of .3686 at $p < .005$ level of confidence. This study sample and correlation coefficient presented are the same as those reflected in Part I of the test of Hypothesis 3, page 93.

2. The Pearson-Product Moment correlation coefficient obtained for Part 2 of the test of this hypothesis indicated that the linear relationship between the classroom test scores and skill qualification test scores of Angoff "rejects" (ARCT with ARSQ) was not statistically significant. The $r$ value obtained was .3148, $p < .109$.

3. The $t$-test used in Part 3 of the test of this hypothesis indicated that Angoff "accepts" achieved a 9.382 mean difference (80.735 - 71.353 for AASQ with ARSQ) on skill qualification test scores. The
t value for this test of difference in means was 3.73, significant at \( p < .001 \).

4. On the basis of the difference in t values and the strength of the t value obtained, the study concluded that classroom achievement of Angoff "accepts" is a better predictor of soldier performance on the skill qualification test than is the classroom achievement of Angoff "rejects." As previously noted in the discussion relative to the test of Hypothesis 3, the validity of the sample drawn for the test of the relationship between classroom test scores and skill qualification test scores is questionable. One should therefore exercise caution in drawing the conclusion that the findings for the test of this hypothesis are reliable.

Discussion

The implications of the findings of this study concern the value of the Angoff score-setting method for use in criterion-referenced measurement, the value of the criterion-referenced measurement model for assessing examinee performance with respect to clearly defined criteria and a domain of behavior, and the effect of empirically based test cutoff scores on the management of criterion-referenced, specialist-level training. The contributions of this study to the body of knowledge pertaining to criterion-referenced measurement include the findings which concern test content validity assessment, cutoff score setting, and predictive validity determination.

Implications for the Angoff Score-Setting Method

The continuing growth in the application of teacher accountability, the use of minimal competency testing, and the need to discriminate
between masters and non-masters in terms of their performance on tests are bringing about an increase in the use of criterion-referenced measurement. From the literature search by the author of this study, it was determined that the present means for setting cutoff scores for criterion-referenced tests is based on either instructor opinion or the contemporary score-setting methodology developed for use with norm-referenced measurement. This study presents an empirically based methodology for setting proper cutoff scores for criterion-referenced tests utilizing the Angoff score-setting method. The procedure minimizes the subjectivity associated with the use of the present score-setting processes. In contrast with the present score-setting methods, application of the Angoff score-setting procedure yields a valid and reliable cutoff score that is in consonance with the minimum job knowledge that the entry-level job performer must possess.

For education and training applications, utilization of the Angoff score-setting method would minimize erroneous labeling of non-masters as masters and thereby reduce the serious effects of students and trainees being allowed to move to the next level of skill before they have acquired essential prerequisites. Also, application of the Angoff method would diminish the erroneous labeling of masters as non-masters, which would reduce the wastes of money, manpower, and time associated with reteaching and overtraining. These resource and training implications associated with erroneous labeling of course graduates
could be significant. Moreover, the objectivity of school administrators' and training managers' decisions concerning administrative, fiscal, and training matters could be enhanced through the use of the Angoff score-setting method.

For the organizations in which course graduates are employed, utilization of the Angoff cutoff score-setting method in the classroom could reduce the possibility of incompetent individuals being assigned to critical jobs. Use of the Angoff method could also enhance minimum job competency on the part of entry-level job performers and increase their units' support capabilities.

**Implications for Criterion-Referenced Measurement**

This study provides the empirical means to establish valid scores for criterion-referenced tests by employing the judgments of practitioners. It presents a practical and reliable methodology for determining minimum job competency and questions the present subjective procedures employed in setting cutoff scores. The study proffers a realistic way to assess test content validity using the test item job relevancy estimates rendered by the practitioners. It also provides a systematic means to assess the predictive validity of a criterion-referenced test by determining the strength of the relationships between the soldiers' classroom test scores, supervisory rating scores, and skill qualification test scores. This empirical means of assessing the predictive validity of a criterion-referenced test could be substituted for the opinions of commanders and managers, which are sometimes based on personal preference rather than on job requirements.
Implications for Training

The implications of the findings of this study concerning training are evident. The findings indicate that the present method for setting criterion-referenced test cutoff scores in specialist-level Army training programs is flawed.

For the Equipment Records and Parts Specialist Course, this study derived a new end-of-course cutoff score of 92 percent, compared to the present cutoff score of 85 percent. Application of this new score, adjusted to compensate for possible chance error in the performance of the test item judges and the students whose test scores were used to set the Angoff-derived cutoff score, would diminish the possibility of mislabeling some non-masters as masters. Based on calendar year 1985 student enrollment in the Equipment Records and Parts Specialist Course, 447 students, or 20.45 percent of the total annual enrollment of 2,186 students, would have been labeled non-masters rather than masters by applying the new empirically derived cutoff score adjusted for one standard error.

Soldiers who graduate from the Equipment Records and Parts Specialist Course are assigned to positions of great responsibility, where less than minimal competency performance could impact seriously on their units' readiness and mission support capability. For this reason, it is recommended that the new Angoff-derived passing score be adopted and adjusted to compensate for one standard error. Any further adjustment to this empirically derived performance standard to compensate for additional chance error should be made only after very careful consideration.
of its effect on soldier minimal competency and resultant impact on unit operations.

**Recommendations for Future Research**

The following recommendations are made for future research which concern the limitations of this study and the effect of motivation on the relationship between soldier job performance and classroom test scores received:

1. Research and develop an empirical means to assess the level of understanding of the concept of minimal competency on the part of the test item judges. The effect of the level of understanding of the concept would be utilized to assess its relative influence on the test item judges' probability estimates concerning the individual test items.

2. Study the effects of job experience on the linear relationships between supervisory rating scores and classroom test scores by placing strict upper limits on the length of job experience on the part of job specialists (test item judges) and job performers.

3. Apply the findings and recommendations concerning the use of the Angoff methodology to other Army specialist-level courses which employ criterion-referenced measurement.

4. Study the effect of motivation on the linear relationship between supervisory rating scores and classroom test scores through the measurement of other achievement goals such as self-development, military appearance, and physical fitness.
5. Study the application of the Angoff methodology in mastery learning situations in the public schools. Specifically, present practices relating to the setting of mastery levels, the validity of mastery tests, and the reteaching of unlearned skills need to be examined in light of this study.
APPENDIX A

CONSENT FORM

Name of Researcher: Dari McDaniels

Study Title: A Predictive Validation Study of Criterion-Referenced Tests for the Certification of Soldiers in Specialist-Level Military Training Programs.

Study Participants' Role: You are asked to perform one of the following tasks:

- As instructor or job specialist (test item judge), you will assess the relevance, or relationship, of the content of each test item to the job of a beginning PLL clerk. You will also estimate the probability that a minimally competent PLL clerk could answer each test question correctly.

- As a member of the committee to assess test content validity, you will review the test and course (content topics) to determine if the universe of behavior sampled by the test is adequate.

There are no risks involved in your performance in this study. The data gathered will be kept confidential. No disclosure of participants' assessments will be made. Information concerning the study outcome will be provided to the participant through official channels, if requested.

I voluntarily agree to participate in this study.

<table>
<thead>
<tr>
<th>NAME/RANK</th>
<th>SOCIAL SECURITY NUMBER</th>
<th>DATE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B

CONSENT FORM

Name of Researcher: Dari McDaniels

Study Title: A Predictive Validation Study of Criterion-Referenced Tests for the Certification of Soldiers in Specialist-Level Military Training Programs.

Study Participants' Role: As a participant you will estimate the job effectiveness of select PLL clerks whom you supervise. Using the Army-Wide Performance Rating Scales, you will assess the level of effectiveness of four specific categories of performance for first-term soldiers. These categories of performance are technical knowledge/skill, effort, following regulations and orders, and maintaining assigned equipment.

There are no risks involved in making these assessments. The data gathered in this study will be kept confidential. No disclosure of participants' assessments will be made. Information concerning the study outcome will be provided to the participant through official channels, if requested.

I voluntarily agree to participate in this study.

Name/Rank _______________________________________________________

Social Security Number ________________________________________
Before you begin, I want to thank you for participating in this very important project. With your help a better way of setting the passing score for a classroom test can be developed.

First, I'd like to discuss briefly the importance of this project. Annually the United States Army trains thousands of soldiers in performance-oriented, specialist-level training programs such as the Equipment Records and Parts Specialist Course. Many of the soldiers graduating from these programs are assigned to very responsible jobs, where less than minimum satisfactory performance could affect seriously the unit's readiness and mission support capability. This is particularly true for "PLL clerks."

Today there is no reliable way to set the ideal difficulty level, or passing score, for these performance-based, specialist-level training programs such as the PLL Clerk Course. The passing score is usually set according to the expectations of the classroom instructor or the school. The test validity and score-setting methods to be developed through this study will replace the present arbitrary method, if the new method proves to be a better one.
Overtraining in these programs should be avoided, since it would constitute a waste of critical manpower by keeping the soldier away from the job for that extra period of time which he/she spends in the course. Undertraining, on the other hand, would be more serious. If this condition exists, the course graduates would not possess the minimum essential skills required for satisfactory performance on the job. A means to ensure that the graduates from these programs are trained to the proper standard is therefore essential.

**RATING THE TEST**

In this exercise you will be making two kinds of judgments concerning each item in the test.

First, you will judge the **relevance**, or relationship, of the content of each test item to the job. You should judge the relevance of each item by considering what knowledge a beginning PLL clerk should possess at the time he/she is assigned to the job.

The second judgment you will make concerns an estimate of the **probability** that a minimally competent person would answer each item correctly.

**Judging Test Item Relevance:** In judging the relevance of each test item, consider the knowledge you believe an entry-level PLL clerk should possess. A 4-point response scale (see rating sheet provided) is used to make your rating. Look at the rating scale at the top of the rating sheet at this time.
If you believe the content of the test item is crucial to the performance of a beginning PLL clerk, you should darken the circle containing the number 4.

If you believe the test item content is important but not crucial, darken the circle containing the number 3.

If you believe the relevance of the test item content is questionable for a beginning PLL clerk, darken the circle containing the number 2.

Finally, if you believe the test item content is not relevant, you should darken the circle containing the number 1.

It is necessary for you to make a judgment and darken one of the circles for each question even though you are not completely sure of the appropriate classification. Also, do not be concerned over the number of questions you place in each of the relevance categories. Consider each question separately and use your best judgment.

Remember, the judgment you make for each question must be based on the knowledge that a beginning PLL clerk should possess.

**Estimating Probability:** The second rating you are asked to make is to estimate the probability that a minimally competent PLL clerk would possess the minimum amount of knowledge to answer each question correctly.

The scale for rating "probability" is shown at the top of the rating scale. Notice that a rating of 1, for example, represents the range of probabilities from .0 to .1. A rating of 2 represents the range of probabilities between .11 and .20, and so forth, through a rating of 1.0. You will use these ranges as discussed below.
Let us assume you feel that there is a .0-to-.10 probability that a minimally competent person would answer a particular question correctly. You would, in this case, darken the circle containing the number 1. Or, let's assume that your imaginary minimally competent PLL clerk would have between an .11-to-.20 probability of answering the question correctly. You would, in this case, darken the circle containing the number 2. Let's try one more example. Assume that your imaginary person has a .61-to-.70 probability of answering the question correctly. In this case, darken the circle containing the number 7. The scale continues in the same manner up to a rating of 10, which you should mark if you believe that a minimally competent person has a .91-to-1.00 probability of answering a particular question correctly.

Do not be concerned over the number of questions you assign the same probability estimate. Consider each question separately and use your best judgment.

Remember, a minimally competent person is one who possesses the minimum amount of job knowledge to perform effectively the duties of the PLL clerk at the beginning level.

Mini-Test: To help you decide on the probability that a minimally competent person would answer each question correctly, it may be helpful for you to know how hard or easy each question was for large groups of students at the Quartermaster School. On the 64-question mini-test you have been provided, you will see a figure in a box next to each response number. This figure is the number of students out of 725 students who recently completed the PLL Clerk Course who gave the correct response to the question. For example, 635 of the students
answered question number 17 correctly. It is emphasized that this is the number of all the students who gave the right answer, not the number that was judged to be minimally competent.

At this time I would like for you to record your relevance ratings (1 = not relevant; 2 = questionable; 3 = important; 4 = crucial) and probability ratings (the percentage chance that the job performer would answer the question correctly) for each question on the mini-test. Again, here are the probability intervals for your use:

- .0 to .10  ----  darken circle 1
- .11 to .20  ----  darken circle 2
- .21 to .30  ----  darken circle 3
- .31 to .40  ----  darken circle 4
- .41 to .50  ----  darken circle 5
- .51 to .60  ----  darken circle 6
- .61 to .70  ----  darken circle 7
- .71 to .80  ----  darken circle 8
- .81 to .90  ----  darken circle 9
- .91 to 1.0  ----  darken circle 10

You have approximately 15 minutes to record both the relevance and probability ratings for the questions on the mini-test, using the rating scale provided. Darken the circles you choose carefully, using a No. 2 pencil. If you wish to change a rating, be sure to erase your first mark completely.

Please begin your rating of the questions on the mini-test.

WAIT ABOUT 15 MINUTES
When you have recorded as many ratings for the 64 questions on the mini-test as time permits, look over the ratings and estimate your average probability scores across all of the items. It won't be necessary for you to actually calculate your average.

The purpose of this exercise and the evaluation of your ratings are to give you an idea of the difficulty of the questions on the PLL clerk examination.

The test you are going to review may be somewhat more difficult than the mini-test.

MAKE CERTAIN THE SOLDIERS (JUDGES) UNDERSTAND THE CONCEPT OF MINIMAL COMPETENCY BEFORE THEY PROCEED TO REVIEW THE ACTUAL TEST.

Please pass the mini-test and rating sheet forward at this time.

Now, open the folder containing the test you were invited to review today. In the folder you should find:

1. A copy of a two-part PLL clerk test.
2. A relevance and probability rating scale.
3. A sheet containing a description of the test and a summary of the instructions for rating the test.

MAKE CERTAIN THE SOLDIERS (JUDGES) UNDERSTAND THAT THEY ARE NOT TAKING THE TEST. THEY ARE SIMPLY JUDGING TEST ITEM RELEVANCE AND MAKING PROBABILITY ESTIMATES. ALSO, INSTRUCT THE JUDGES TO EQUATE MINIMAL COMPETENCE WITH THE UPPERMOST POINT OF THE RANGES -- OR INTERVALS --
SHOWN ON THE RESPONSE SCALE. STRESS THAT THIS IS IMPORTANT BECAUSE THE
PROBABILITY ESTIMATES WHICH THEY ARE MAKING CONCERN PERFORMANCE THAT IS
BASED ON A SPECIFIC STANDARD OF PERFORMANCE FOR EACH TEST ITEM.

You are asked to work independently and not discuss your ratings
with other persons until you have completed your review of the entire
test. Do not write on your copy of the test since it will be used
again.

You may take a break whenever necessary. Do not take the tests
from the room. Turn your work over on your desk when you take a break.

Fill in all of the blanks at the top of the rating scale before
you begin.

Page 14 of the test is a foldout. This page, the "Document
Register for Supply Actions" (Figure 16), will be used at various times
in rating the "Manual PLL" portion of the test. Fold page 14 out before
you begin.

Also, page 20 is a foldout. Page 20 (Figure 22) contains the
"Document Register for Supply Actions" for the "Automated PLL" portion
of the test. It will be necessary to fold out page 20 before you begin
rating the automated PLL portion of the test.

Turn in all materials when you are finished.

DON'T FORGET TO HAVE SOLDIERS SIGN THE CONSENT FORM!
Mr. Dari McDaniels  
201 Norwood Drive  
Colonial Heights, VA 23834

Dear Mr. McDaniels:

You certainly have my permission to use the optical scanning sheet that I developed for the Virginia validation study of the National Teacher Examinations. In fact, you are welcome to use four boxes of misprinted forms that I received from the printer before getting the corrected forms. I've enclosed one of these misprinted forms for you to consider. You could also contact the printer to obtain additional correct copies, since they should still have the original master.

If I can be of further help, please do not hesitate to call on me.

Sincerely,

Lawrence H. Cross  
Associate Professor

/cw
Enclosure
June 18, 1987

Mr. Dari McDaniels
201 Norwood Drive
Colonial Heights, VA 23834

Dear Mr. McDaniels:

This is in response to your letter of May 26 asking me for my rationale for recommending against using analysis of covariance (ANCOVA) in your dissertation.

As I understand the design of your study, you have a sample of approximately 100 soldiers for whom you have scores on a) a clerical aptitude test taken before enrolling in the course, b) a comprehensive course examination, and c) supervisory ratings of on-the-job performance. Your original intent was to compare the mean supervisory ratings of those who scored above the present cut-off score of 85 but below 92 on the comprehensive examination with those who scored above 92. The clerical aptitude scores were to serve as the covariate.

In order to understand why I recommended against using ANCOVA, it would be helpful to explain what ANCOVA can properly be used for. ANCOVA was designed for use in experimental studies (where the groups have been randomly formed) to increase statistical power by removing variance in the dependent variable that can be predicted on the basis of the covariate.

Not only do you not have a treatment, as required by an experimental study, but also the two groups are anything but randomly equivalent. Indeed, you defined your two groups by the scores on the comprehensive examination, and thus, they differ systematically by definition of group membership. To use ANCOVA with groups known to differ systematically and substantially is a misuse of ANCOVA [See Lord, F. M., (1967) Psych. Bulletin 68, 304-305].

More importantly, however, I question the rationale for comparing the mean supervisory ratings of those scoring above and below 92 on the comprehensive examination, whether or not ANCOVA or ANOVA is used. If the test is a validity predictor of the supervisory ratings, the mean for those scoring above 92 must be higher than those scoring below 92. This result will be true regardless of where the cut-off score is set as long as the test is valid. All that you can hope to accomplish via ANCOVA is
June 18, 1987

Mr. Dari McDaniels
Page 2

I hope these thoughts are of value to you. Feel free to use this letter in any way you choose. Best wishes in completing the study.

Sincerely,

Lawrence H. Cross
Associate Professor
Educational Research


Cross, L. H. Validation Study of the National Teacher Examinations for Certification of Entry-Level Teachers in the State of Virginia, Phase I. A research report by the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1982.
Validation Study of the National Teacher Examination for Certification of Entry-Level Teachers for the State of Virginia, Phase II. A research report by the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1984.


Government Documents

Department of Defense (DOD) Publication 1304.12 AA. ASVAB Test Manual for Forms 8, 9, 10, 11, 12, 13, and 14. (1 July 1984).


VITA

NAME: Dari McDaniels

BIRTH: August 4, 1928

RESIDENCE: Colonial Heights, Virginia


Awarded Bachelor of Science in Secondary Education, Virginia Commonwealth University, 1971

Awarded Master of Education in Administration, 1978

Awarded Certificate of Advanced Graduate Study in Educational Administration, 1983

Candidate for the degree Doctor of Education in Educational Administration, 1988.
A PREDICTIVE VALIDATION STUDY OF CRITERION-REFERENCED TESTS FOR THE CERTIFICATION OF SOLDIERS IN SPECIALIST-LEVEL MILITARY TRAINING PROGRAMS

Dari McDaniels, Ed.D

The College of William and Mary in Virginia, 1988

Chairman: Dr. Robert J. Hanny

Problem: This study assessed the predictive validity of criterion-referenced tests in a military setting with cutoff scores set by the Angoff and conventional score-setting methods.

Procedure: Thirty-six instructors and thirty-six job specialists assessed each test item for job relevance and the probability that a minimally competent person would answer each question correctly, resulting in a new test cutoff score. Intragroup variability and interrater reliability of judgments were calculated. Test predictive validity assessment compared classroom test scores, supervisory rating scores, and skill qualification test scores of 100 job performers based on the two score-setting methods. Sample sizes varied from 17 to 100. A behaviorally anchored rating scale was used to estimate soldier performance effectiveness. The hypotheses were tested using analysis of variance, a correlation procedure by Ebel, $t$-test, and Pearson Product-Moment correlation. Null was accepted or rejected at .05 level of significance.

Results: Findings follow: (1) intragroup variability and interrater reliability of judgments were statistically significant; (2) strengths of correlation coefficients for classroom test scores and supervisory rating scores under Angoff methods exceeded $r$ values for scores under conventional method; (3) strength of correlation coefficient for classroom test scores and skill qualification test scores under conventional method exceeded $r$ value for scores under Angoff method; (4) correlation coefficients for classroom test scores and supervisory rating scores were statistically significant for Angoff "accepts" (examinees who scored 92 through 100 percent) but not for Angoff "rejects" (examinees who scored 85 through 91 percent) in three out of four job performance areas, and the means of the supervisory rating scores of Angoff "accepts" and "rejects" were significantly different; and (5) the correlation coefficient for classroom test scores and skill qualification test scores was statistically significant for Angoff "accepts" but not for Angoff "rejects," and the means of the skill qualification test scores of the two groups of job performers were significantly different.

Conclusions: The Angoff cutoff score-setting method provides an effective means for setting criterion-referenced test cutoff scores. The Angoff and present score-setting methods yield significantly different test standards. The score derived by the empirical method is a better...
measure of minimum job requirements of an entry-level performer, thereby enhancing the predictive validity of the classroom test. Recommendations for future research are included.