DEVELOPMENT OF COLLABORATIVE CRITERION-REFERENCED TESTING PROCEDURES AT THE ARMY RESERVE READINESS TRAINING CENTER (ARRTC)

by

Robert A. Woolsey

A Research Paper

Submitted in Partial Fulfillment of the
Requirements for the
Master of Science Degree in
Training and Development

Approved for Completion of 4 Semester Credits
TRHRD-735 Field Problem in Training and Development

_____

David A. Johnson, Ph.D.
Research Advisor

The Graduate School
University of Wisconsin-Stout
December, 2001

The Graduate School
University of Wisconsin-Stout
Menomonie, Wisconsin 54751

ABSTRACT

Woolsey Robert A.
(Writer) (Last Name) (First Name) (Initial)

Development of Collaborative Criterion-Referenced Testing Procedures at the Army Reserve Readiness Training Center (ARRTC)
(Title)

Training and Development	Dr. David A. Johnson	December, 2001	103
(Graduate Major)	(Research Advisor)	(Month/Year)	(No. of Pages)

American Psychological Association (APA) Publication Manual
(Name and Style Manual Used in this Study)

Traditional learning institutions have approached instruction and testing from the perspective of the individual student. For those subjects that are strictly knowledge-based, this method works best; that is, to test individually to evaluate the individual student's ability to recall the information covered in the course. Some training institutions are reluctant to train and test their students as groups or collaboratively. This comes from feelings that they will cheat, that one achiever will carry the whole group, and that the slower students will ride on the coattails of those who are stronger in the group. Professional training institutions and most companies are reluctant to train their employees as groups for many of the same reasons.

As these students enter the workforce, they are usually part of a team or collaborative effort, who are working as a team to complete mutual jobs. They often do not have references on the job, instead relying on instructions from a supervisor or manager. As situations or projects are completed, if there are questions that need answering, the team often works together, sharing information and problem solving toward a collective result. Yet these same individuals are evaluated based on their personal performance generally, and possibly on how well they support group cooperation and team cohesiveness as a secondary consideration.

The Army is no different than private industry in this regard. Soldiers are taught collaboratively and work as teams toward a common goal, but are then tested and evaluated as individuals. Only when soldiers are members of crew served weapon systems are they tested and evaluated on the system as a collaborative unit. The Army strives toward teamwork, inspecting units in different areas, toward a common goal or standard. Soldiers are encouraged to work together, but when they attend schools, most of the tests are geared toward assessing the individual's accomplishments, not the team accomplishments.

There needs to be a way to design tests that both evaluate the student's individual abilities and the collaborative accomplishments. A total testing program needs to evaluate both the individual and the collaborative accomplishments of instructional objectives. The researcher is not suggesting doing away with individual testing – as an instructional institution, there still needs to be a way to ensure each student can accomplish all the

tasks of a course or subject. However, in addition to the individual tests there needs to be a way to develop effective and validated collaborative testing procedures to ensure the students are not only self reliant, but able to work collaboratively, toward common goals or objectives. With this in mind the following questions are asked:

Is there a need to test an individual if a group test is used to ensure students can accomplish the class or course objectives?

Can a test procedure be created that will conclusively evaluate individual and group work?

What are the steps in validating a group test?

Are group tests as effective as individual tests?

This study will examine the different methods and procedures used to evaluate the validity and reliability of individual Criterion-Referenced Tests (CRT) for possible modification to measure the validity and reliability of a collaborative CRT. There are several types of methods and procedures for this purpose that have been accepted and used by training institutions.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I want to express my thanks and appreciation to Dr. David A. Johnson, my research advisor, for his patience, advice, and guidance with this study.

To Mary Snyder and Pat Upton, who between us kept us on the straight and narrow during the months of endurance, to persevere to the end of this part of our journey. Remember to smile and nod.

For my wife and soul mate, Cindy. The person God chose for me, the person I want to grow old and share the fruits of life with. Without your support, this would have not been possible. Thank you for your sacrifice for our future. This really is it.

For my son, Rob. I wish only the best for you as you now begin the journey. May God go with you and bless you as he has me.

The sacrifices you both have made for me to complete this journey can never be repaid.

Chapter I

Introduction

Introduction of the Army Reserve Readiness Training Center

The Army Reserve Readiness Training Center (ARRTC) is a second-generation government training institution. Established in 1975 as a consolidation of two training centers for the Army Reserve Military Technicians (MTs), it is located in west central Wisconsin, primarily due to its central location in the Midwest area of the United States. It has gone through several changes and growth over the years to its present size of 82 civilian employees and 76 soldiers. The ARRTC is the only institution designed to train Army Reserve personnel. This includes the full time workforce and Troop Program Unit (TPU) members. It is often called "The Schoolhouse of the Army Reserve" and trains the Reserve force to "go to work."

The ARRTC's mission statement is, "Design and provide quality training to increase our customers' readiness." The vision statement is:

A – Achieve a relationship with our customers that develops complete trust and confidence.

R – Readiness enhancement through quality training.

R – Reaching our customers through the application of all available technologies.

T – "Training Institution of Choice."

C – Committed to exceeding our customers' expectations in quality, delivery, and value.

The ARRTC's core values are:

V – Value or students/customers as priority one.

A – Advance professionalism at all levels.

L – Live by our integrity and ethics.

U – Use common sense in all processes.

E – Effectively communicate.

S – Stress constant improvement through teamwork.

The ARRTC provides pre-mobilization training in the areas of Human Resource, Budget and Finance, Network Operations and Security, Physical and Document Security, Logistics, Engineering, Mobilization, Training, Operations, Retention, Strength Management, and initial employee orientation, for both civilian and military personnel. The ARRTC trains approximately 3000 students each year, which keeps pace with the approximate 30 percent attrition rate in the workforce annually from retirements and reassignments.

The ARRTC provides training to civilian employees and soldiers in a number of areas. However, before courses and support materials are created or modified, the United States Army Reserve Command (USARC), located in Atlanta, Georgia, must first formally task the ARRTC to create or approve courses undergoing revalidation. The creation and revalidation processes are similar. The USARC must approve the needs analysis before a course is either created or modified. To create a course, a Task Analysis Board (TAB) and Job Analysis Board (JAB) are held. Subject Matter Experts (SMEs) in the field and one level of command above where the course material is targeted are

invited to analyze the job and the tasks that are necessary to perform the job in the field. The results are then sent out using usually a Difficulty, Importance, and Frequency (DIF) survey, to previous students and supervisors on the new task list that were picked as a result of the TAB. After the surveys are received, they are processed using the Perseus evaluation software, which ranks the results from the field in a relative manner. The results are then reported to the course team responsible for the development of the course for further action. For a course revalidation, the steps are the same, except in this case the boards are the JAB and Task Review Board (TRB). In either case, analysis is preformed to ensure the courseware is what the field needs and not what someone else feels the field should receive.

The ARRTC uses their own Systems Approach to Training (SAT) process, outlined in ARRTC Regulation 351-1-1 in the Analysis, Design, Development, Implementation, and Evaluation phases to create, revise, and make improvements to courseware being considered for initial development and to revalidate existing courseware. This SAT process provides for the validation of the effectiveness of the courseware, tests, handouts, and other components of the course. The ARRTC SAT is derived from the Training and Doctrine Command (TRADOC) SAT, TRADOC Regulation 350-70, which is a very lengthy document that entails all of the steps and tasks needed to develop a course and the courseware needed. TRADOC is the Department of Defense (DoD) command responsible for training centers and schooling institutions. They publish regulatory policies and procedures that outline what, where, how, when, and why training is conducted. The ARRTC SAT takes parts of the

TRADOC SAT, streamlining the process. For example, a course that TRADOC develops is a course of instruction that is between 16 weeks and 18 months in duration. On average, to develop this length of course, it takes two to three years. ARRTC courses are a maximum of two weeks in length have to be developed within six months after the school Commandant approves the working Program of Instruction (POI). College credit, both lower and upper division baccalaureate degree credit is given for several of the ARRTC courses. In order to receive college credits, the ARRTC courseware must meet standards in regards to validity of courseware, testing instruments, and effectiveness of student ability to meet or exceed course requirements according to the American Council on Education (ACE).

To train the instructional staff, the ARRTC has historically relied on the internal staff and faculty section to monitor the Instructional Professional Development Training Program. This formal training program was stopped in 1994, relying instead on TRADOC courses to train the instructional staff, providing a wider range of instructional development for traditional classroom instruction and Internet or Distributive Learning (DL) techniques using Video Tele-Training (VTT) to provide synchronous training. The ARRTC also relies on each training center for On-the-Job Training (OJT) for newly assigned personnel to provide the necessary skills and development of knowledge. On-the-Job training occurs from other team members on educational principles, policies, and activities. The Instructional Systems Specialist (ISS) provides educational principle training and guidance on creating and modifying course materials. The Educational Technician (Ed Tech) provides classroom and courseware assistance with creating and

modifying courseware. New employees will attend the Total Army - Instructor Training Course (TA-ITC) and the Systems Approach to Training (SAT) Basic Course within six months after assignment to an instructional position. They will also observe other team members in the classrooms, comparing instructional styles and techniques, observing effectiveness of instruction and student ability to meet the course requirements. The Training Center Chief (TCC), with input from the Course Manager, will evaluate the new employee's development, making recommendations on technique, process, and opportunities to further develop themselves professionally through additional OJT and by job shadowing other instructors. The training process has limited structure, as there has not been a formal program in place since 1994.

Statement of Problem

On 15 February 2001, research was conducted to find what guidance, if any, there was on collaborative testing in the ARRTC and TRADOC SAT regulations. This research was conducted in response to the ARRTC SAT recommendation that students work and be tested in groups (collaborative) whenever possible. Neither SAT regulation provides a process of measuring the validity and reliability of collaborative testing. The emphasis of this paper is on the steps needed to create an effective, validated, and reliable collaborative test process for Criterion-Referenced Tests (CRT).

Purpose of the Study

The purpose of this study is to identify the steps necessary to create a process that will measure the validity and reliability of a collaborative Criterion-Referenced Test (CRT), which would be appropriate to the ARRTC.

Significance of the Study

The ARRTC has set no timetable for a collaborative testing process. However, since the ARRTC SAT identifies collaborative testing as a way to measure students ability to meet course objectives, a process needs to be defined and approved as soon as possible, allowing for the instructional staff to create the tests and measure the effectiveness of instruction. Testing in existing courses may be attempting to create these collaborative tests now with no standardization or validation process, allowing for the lowest non-validated measurement of student accomplishment to erroneously report an invalid student degree of mastery.

Limitations of the Study

The limitations of this study are:

1. Limited to DoD regulatory policies and procedures. ARRTC must use DoD regulatory policies and procedures.

2. No surveys were used, only quantitative measurements.

3. Limited to the ARRTC.

4. Used on courses or subjects that use only Criterion-Referenced Tests (CRT).

5. Student selection. The student's higher headquarters select and schedule the training the students will attend at ARRTC. These students come from all over the world, from a variety of jobs, staying an average of two weeks while attending training for their particular jobs. Their higher headquarters must approve any changes to their primary duty or before any additional training or duties were

performed. Coordination with the commands to get the student's involvement on short notice is very difficult to impossible in the time the student is at ARRTC. Therefore, this was the first reason no students were used for the reliability check of the collaborative CRT.

6. Civilian timekeeping issues. Classes are scheduled from 0730 to 1630 daily. For a civilian employee to work more than eight hours per day, 40 hours per week, or 80 hours per pay period (two weeks), their timekeeper must approve any deviation of time before the work is performed. Since the civilian students come from all over the world, and the civilian employee's timekeeper is a different person than their higher headquarters whom may have approved the additional duty. To get permission from both their timekeeper and their higher headquarters on short notice while the employee was at ARRTC would be very difficult to impossible. Therefore, this was the second reason no students were used for the reliability check of the collaborative CRT.

7. Small pool. There were a very small number, or small pool, of individuals available at the ARRTC to act as "master groups" and "nonmaster groups" to take the different tests that will be developed under the test conditions. Therefore, only 12 individuals were available from the staff, three groups of two or six total in the "master groups" and three groups of two or six total in the "nonmaster groups".

8. Test reliability evaluation procedures. To perform the reliability evaluations of the tests, the test conditions needed to be conducted under the

ARRTC classroom conditions. This requirement eliminated the possibility of mailing the test to individuals in the field to complete and resubmit to ARRTC. The participants would not be in a controlled environment, would not have access to ARRTC classroom applications, and would not be in the proper work groups. ARRTC and TRADOC SAT regulations require that the test be evaluated for reliability in the development phase of the SAT. Therefore, the reliability evaluations of the tests were done at ARRTC.

Definition of Terms

For clarity of understanding, the following terms are defined as follows:

1. Active/Guard Reserve (AGR) – Army Reserve Soldiers on Active Duty in accordance with U.S. Title 10.

2. American Council on Education (ACE) – An independent, nonprofit organization that evaluates educational programs according to established college-level criteria and recommends college credit for those that measure up to these standards.

3. Andragogy – Emphasizes that adults are self-directed and expect to take responsibility for decisions. Andragogy means that instruction for adults needs to focus on the process and less on the content of what is being taught. Examples such as case studies, role-playing, simulations, and self-evaluation are the most useful. Andragogy also applies to any form of adult learning that has been used extensively in the design of organizational training programs (especially "soft skills" domains such as management development).

4. Army Regulation (AR) – Governing policies and procedures for the Army, Army Reserve, and National Guard.

5. Army Reserve Readiness Training Center (ARRTC) – Government training institution primarily for Full Time Unit Support (FTUS) personnel and Army Reserve soldiers.

6. Collaborative Testing – Testing done jointly or in cooperation with others.

7. Criterion-Referenced Instructional (CRI) – Instruction that uses a comprehensive set of methods of the design and delivery of training programs, which are based upon ideas of mastery learning and performance oriented instruction.

8. Criterion-Referenced Test (CRT) – A test that establishes whether or not a unit or soldier performs the learning objective to the established standard. Performance is measured as a "go" or "no-go" against a prescribed criterion or set of criteria - the learning objective standard. It is scored based upon absolute standards, such as job competency, rather than upon relative standards, such as class standings.

9. Degree of Mastery – The performance of the training objectives within the prescribed conditions and to the stated standards.

10. Department of Defense (DoD) – Civilian employees and military members of the Armed Forces in an Active, Reserve, or retired status.

11. Difficulty, Importance, and Frequency (DIF) survey – Asks job holders to rate each of their job's tasks in terms of how difficult they are to perform, how important they are, and how frequently they are performed.

12. Distributive Learning (DL) – Students take courses from a variety of sources (and delivery modes) to customize a program of study. Often is used synonymously with online learning.

13. Educational Technician (Ed Tech) – Individual responsible to support the instructors with their classroom needs. This is in addition to supporting the entire training center and course teams in areas such as brochures, survival kits, submissions, and publications.

14. Full Time Unit Support (FTUS) – Military Technicians and Active Guard/Reserve soldiers.

15. Instructional Systems Specialist (ISS) – Individual responsible to provide guidelines and expertise in the application of educational principles within the training center courses.

16. Job Analysis Board (JAB) – A board of Subject Matter Experts, from a specific job or career to be trained, and focus student learning on what needs to be learned. Both terminal and enabling objectives are learning objectives.

17. Masters – Individuals who have command of the material being tested.

18. Master Groups – A group of individuals who have command of the material being tested.

19. Metacognition – Knowledge and awareness of your own cognitive processes, how they function, when it's likely to falter, etc. Examples of metacognition are:

"I don't recall"

"I understood this fairly well"

"I won't be able to solve this problem right away"

"I can't study with the TV on"

"Her name is on the tip of my tongue"

20. Military Technicians (MTs) – Full time Department of Defense civilian employees who are required to be members of the military in an active reserve status to hold their job.

21. Nonmasters – Individuals who know nothing about the material being tested.

22. Nonmaster Groups – A group of individuals who know nothing about the material being tested.

23. Norm-Referenced Test (NRT) – A test that grades a student based on the performance of other students taking the same test. Is scored based upon relative standards, such as class standings, rather than upon absolute standards, such as job competency.

24. On-the-Job Training (OJT) – Formal training for learning the skills and knowledge to perform a job that takes place in the actual work environment.

25. Perseus - Evaluation software that uses email to send and receive student DIF information, leaving the sender anonymous.

26. Phi Coefficient – Mathematical formula used to correlate the results of two activities that experience the same thing. Used to determine reliability and validity of a criterion-referenced test (CRT).

27. Program of Instruction (POI) – The program of instruction is a requirements document that provides a general description of course content, duration of instruction, types of instruction, and learning objectives.

28. Quantitative Skills for Trainers Reference Book (QSTRB) – Reference book that describe how to perform various quantitative skills and guide archetypal TRADOC trainers in their associated duties and tasks.

29. Systems Approach to Training (SAT) – The Army's training development process. It is a disciplined, logical approach to making collective, individual, and self-development training decisions for the total Army. It determines whether or not training is needed; what is trained; who gets the training; how, how well, and where the training is presented; and the training support/resources required to produce, distribute, implement, and evaluate those products. The SAT involves training related phases: analysis, design, development, implementation, and the evaluation process.

30. Subject Matter Experts (SMEs) – An individual who has a thorough knowledge of a job (duties and tasks). This knowledge qualifies the individual to

assist in the training development process (i.e., consultation, review, analysis, etc.) Normally, a SME will instruct in his area of expertise.

31. Synchronous Training – Training happening or existing at the same time but at possibly different locations.

32. Task Analysis – An analysis that results in the identification of task performance specifications for each task selected for training, i.e., initiating cues, task steps, conditions, standards, materials, references, and safety factors.

33. Task Analysis Board (TAB) – A team of subject matter experts that reviews the total task inventory and job performance data. Their purpose is to recommend which tasks should be included for training and to perform task analysis.

34. Task Review Board (TRB) – A team of subject matter experts that convenes to review an established course's task for Training List and recommends changes, as appropriate.

35. Task Selection Board (TSB) – See Task Analysis Board.

36. Test Reliability – Addresses whether a test gives dependable or consistent scores. Reliability refers to the consistency of a set of test results (precision). Next to validity, reliability is the most important measure of a test's quality. A test that is not valid, but highly reliable, may be measuring the wrong thing with great precision.

37. Test Validity – Addresses whether a test measures what is was intended to measure (accuracy). It is the most important single attribute of a good

test. Even if other practical and technical considerations are satisfactory, the test's quality is doubtful without proof of validity. Validity is a unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree to which that evidence supports the inferences made from scores.

38. Total Army - Instructor Training Course (TA-ITC) – Course of instruction that trains personnel to be instructors for the Army. Satisfactory completion awards a soldier qualification as an instructor.

39. Training and Doctrinal Command (TRADOC) – Command responsible for training centers and schooling.

40. Training Center Chief (TCC) – Individual primarily responsible for the overall management and support of the training center team.

41. Troop Program Unit (TPU) – An Army unit that when mobilized must function effectively in a very short time, usually three to five days.

42. Video Tele-Training (VTT) – Video training delivered via communication links such as satellite or cable links. There are two types of VTT: broadcast and desktop.

43. United States Army Reserve Command (USARC) – Major Army Command Headquarters located in Atlanta, Georgia

Chapter II

Review of Related Literature

Introduction

The Army Reserve Readiness Training Center (ARRTC) needs a collaborative testing process as a means of meeting one of the requirements of the ARRTC System Approach to Training (SAT) regulation, which requires a testing plan for all functional training (ARRTC SAT Regulation, 1999) requires a course testing plan for all functional training). Because there are so many courses that rely on collaborative testing, the ARRTC needs a collaborative testing plan. The majority of the subjects taught by the ARRTC are functional based training and the staff and faculty are to teach and test groups (collaboratively) where possible, emulating how the students will work on the job. Presently, there is no testing process or procedure for developing a test that will measure the validity and reliability of a collaborative test.

Traditional learning institutions have approached instruction and testing from the perspective of the individual student. For those subjects that are strictly knowledge based, this method works best; that is, to test individually to evaluate the individual student's ability to recall the information covered in the course. Some training institutions are reluctant to train and test their students as groups. This comes from feelings that they will cheat, that one achiever will carry the whole group, and that the slower students will ride on the coattails of those who are stronger in the group. Professional training institutions and most companies are reluctant to train their employees as groups for many of the same reasons.

Pray Muir and Tracy (1999) describe how businesses are approaching collaborative efforts in today's workplace. They discuss how today fewer people work in isolation.

To meet this new approach, many university programs, especially in professional schools, now engage students in teamwork. Some professors require students to collaborate during small-group discussions, in preparing formal group presentations, or by co-authoring papers.

People work and are supervised in groups ranging in size from three to seven (small) to groups from 10 to 15 (medium). Since people work together, usually as groups or sections, why should they train and be tested individually? Supervisors who evaluate their employee's job performance, base the evaluation on the person's ability to do the job and on how well the employee works with others in the company. While individual evaluation is a good method of measuring how well an individual is doing their job, the company or business measures success and failure on a holistic scale. When jobs come into a department or section to be completed, the supervisor will assign several people to work on the project, to ensure the job is done as quickly and as accurately as possible. To help in this process, the collective whole or collaborative effort, not the individual effort is used and completing the project is the collective work of many.

Wang (2000) provides that, although small group instruction is a strategy that is gaining acceptance in the United States and abroad, many adult education programs have not taken advantage of it.

The ARRTC provides training for the full-time unit support personnel, both military and civilian, in the functional areas of Human Resource, Training, Operations, Mobilization, Logistics, Engineering, Finance, Budget, Retention, Computer Security, and Network Operations. To maximize training time and efforts, students work in small groups, usually from three to five in number. Classroom instruction is geared toward both individual and collaborative projects, but there is no clear effective testing plan to measure collaborative ability to meet the class and course objectives. This paper will develop a collaborative testing program to effectively measure the validity and reliability of collaborative tests.

Needs Assessment and Performance Improvement Methodologies

ARRTC SAT Regulation 351-1-1 states that a systems approach to training is used to determine if there is a need for the creation, modification, or deletion of subjects or courses. Part of this SAT process requires that a needs assessment be conducted, and if necessary, a systematic plan be developed to explore how this need can be met. Since there are references to creating groups (collaborative) in the ARRTC SAT, but no process for developing the collaborative tests, the researcher feels that this meets the requirements for showing that a need exists. The need is for the development of a process that will identify the steps necessary to create a collaborative testing process that will measure the validity and reliability for CRTs.

Purposes to be served by Review of Research Literature

Creation of a collaborative testing process that measures the validity and reliability for CRTs will require research to look beyond what is currently on hand, with

an expanded view to look beyond what the existing information addressed. The models the researcher has discovered apply the validation process to individual test instruments. Consider the research conducted to date as the beginning of a journey, that more than one use could be found for the data that exists.

Tests, Testing, and Mastery

ARRTC SAT Regulation 351-1-1 states that all functional training will be tested. But why test? What is a test? How does testing improve both the student's abilities and the quality of the instruction or teaching? These are questions this section will address.

A test is defined as a device, technique, or measuring tool used to:

- Determine if a student or group can accomplish the objective to the established standard.
- Determine if training does what it is designed to do efficiently and effectively.
- Measure the skill, knowledge, intelligence, abilities, or other aptitudes of an individual or group.
- Collect data as a basis for assessing the degree that a system meets, exceeds, or fails to meet the technical or operational properties ascribed to the system.

As we can see by the definitions, all three of the questions are answered in the definition of a test. A test is given to evaluate and measure the student's ability to meet the established standard for a class or for a course. The test also measures the effectiveness of the instruction. Finally, tests also measure the degree of mastery that the

student exhibits in the area covered by the test. In 1963, John B. Carroll argued for the ideas of mastery learning. Mastery learning suggests that the focus of instruction should be the time required for different students to learn the same material.

The ARRTC orients its courses in two types. The first type is to instruct knowledge-based subjects only. In these courses, the objective is to provide information rather than develop specific capabilities to perform job tasks. The second type is skill-based. In these courses, the objective is to train the students on how to perform job/functional area tasks. All functional-based courses will have quantifiable standards verifying student attainment of learning objectives, unless the Commandant grants an exception to policy. All tests should be criterion-referenced and, where possible, performance-oriented. Course tests will be designed to validate and measure performance. After arrival at ARRTC, students are normally broken down into teams within a course, based on experience, knowledge, ability, and skills. This ensures that as much as possible, each team will have the same group level knowledge to perform the group tasks that will be assigned to the groups throughout the course.

Criterion-Referenced (CRT) and Norm-Referenced (NRT) Tests

Normally tests can be divided into two separate categories, based on the purpose and method of interpreting test results. When a relative ranking of students is the desired outcome with respect to the present group of students, NRT is used. When a description of the learning tasks a student can and cannot perform is desired with respect to specific knowledge and skills that can be demonstrated, CRT is used. What does this mean to an educator, instructor, or an employer? Consider the following:

NRT interpretation – Pat is the third highest in his class of 20 students.

Mary surpassed 90 percent of the students on the Math test.

CRT interpretation – Bob can identify all the parts of the telephone and demonstrate its proper use.

Steve correctly completed 15 of the 18 chemical equations.

The definitions of both CRT and NRT are found in the following publications, as noted.

Criterion-Referenced Test (CRT). ARRTC SAT Regulation 351-1-1 (1999). A test that establishes whether or not a unit or student performs the learning objective to the established standard. Performance is measured as a "Go" or "No Go" against a prescribed criterion or set of criteria, the learning objective standard. It is scored based upon absolute standards, such as job competency, rather than upon relative standards, such as class standings.

Criterion-Referenced Test (CRT). Measurement and Evaluation in Teaching (1990). A test designed to provide a measure of performance that is interpretable in terms of a clearly defined and delimited domain of learning tasks.

Stated another way, CRT is a test designed to measure each student's ability to achieve a set minimum degree of mastery on the test. Each student is measured against the test standard, not measured against other student's degree of mastery.

Norm-Referenced Test (NRT). ARRTC SAT Regulation 351-1-1 (1999). A test that grades a student based on the performance of other students taking the same test. Is scored based upon relative standards, such as class standings, rather than upon absolute standards, such as job competency.

Norm-Referenced Test (NRT). Measurement and Evaluation in Teaching (1990). A test designed to provide a measure of performance that is interpretable in terms of an individual's relative standing in some known group.

Quantitative Skills for Trainers Reference Book (QSTRB) (Sep 1997) provides that the Army generally, and the ARRTC specifically, does not accept the NRT approaches. These organizations are more concerned with each person's ability to master the instructional content rather than a relative standing in a group of students. The CRT is generally used to create a series of coordinated achievement tests designed to measure the behavioral objective(s) within a course or subject of study. Although mastery may not indicate complete knowledge of a subject or course, it may be defined as obtaining a score at least a minimum competency level that is an acceptable percent (standard) correct. The criterion-referenced approach to learning can encourage greater instructor emphasis on individualized instruction and enables the students to work at their own pace. In the group setting, the group dynamics should allow for a faster understanding of the material, without the competition between each member of the group. It is based on a philosophy that people differ not so much as a level of intelligence, but in the speed or pace with which a student will acquire facts and skills.

It should also be noted that the purpose of a CRT is not just to show the results of how well the student was able to complete the goal of the test. Rather, the CRT results provide additional information such as:

- Distinguish between properly and improperly stated instructional objectives.
- State instructional objectives as learning outcomes.
- Identify technical flaws in test items.
- Construct test items that are free of technical defects.

Finally, adults learn differently than children. Knowles theory of Andragogy (Knowles, M. 1984) emphasizes that adults are self-directed and expect to take responsibility for decisions. Andragogy means that instruction for adults needs to focus on the process and less on the content of what is being taught. Examples such as case studies, role-playing, simulations, and self-evaluation are the most useful. Andragogy also applies to any form of adult learning that has been used extensively in the design of organizational training programs (especially "soft skills" domains such as management development).

The Military approach to training and instruction emphasizes experiential learning as well as theories social learning. Decision-making and problem solving are two skill domains that are fundamental to most types of military tasks. Mager uses the Criterion-Referenced Instructional (CRI) approach as a comprehensive set of methods of the design and delivery of training programs, which are based upon ideas of mastery learning and performance oriented instruction. (Mager, R. 1975)

Collaborative Testing

Since February 2001, a search for any existing collaborative information in the ARRTC and TRADOC SAT regulations revealed numerous statements for the need of collaborative training techniques and procedures. At the user and management levels of Human Resource Management courses, for example, employees and managers seldom work alone. The instruction in the Personnel Management Course (PMC) is given to introduce or review the subject matter. Then, the class, which is already broken down into groups of four students each, work as a group to solve scenarios on topics that involve the use of computer applications, databases functions, and the research of Army policy. The students then report their results of the scenarios in a series of informational briefings to a simulated supervisor or commander. This emulates a Personnel Manager's role back at their units, which is to provide personnel and administration recommendations in the form of a briefing to their supervisor or commander. Likewise for the user levels, except that their briefings are more technical in nature. This research project is being conducted to provide a process of measuring the validity and reliability of collaborative testing. Neither the ARRTC nor TRADOC SAT regulations provide a procedure to create an effective, validated collaborative test process for CRT.

The ARRTC SAT regulation recommends that students work and be tested as groups (collaboratively) whenever possible. The researcher will investigate different sources to develop a collaborative testing process that will measure the validity and reliability of a collaborative test for CRT.

The ARRTC has set no timetable for the development of a collaborative testing process. However, since the ARRTC SAT identifies collaborative testing as a way to measure students ability to meet course objectives, a process needs to be defined and approved as soon as possible, allowing for the instructional staff to create the collaborative tests and measure the effectiveness of the tests and instruction.

Without a standardized and validated process, collaborative tests would report an invalid student degree of mastery and an inaccurate measurement of the effectiveness of the instruction. Current testing in existing courses was developed to test students in a collaborative environment, without a standardization or validation process. This allows for the lowest non-validated measurement of student accomplishment, resulting in an invalid student degree of mastery.

Pray Muir and Tracy (1999) provide these student comments on the collaborative tests.

1. Positive - Almost all partners reported that their contributions were equal. Two students reported, "We worked together throughout the exam. We asked questions of each other and created what we felt was an excellent example of what we knew." In one course, students knew in advance that they must mutually report on the amount of effort contributed by each partner. Over the years, 97 percent of the partners have reported that each person contributed equally.

2. Negative - Nevertheless, a few students described their efforts as unequal. One wrote, "The faculty would be shocked to discover how many

> students with high GPAs are always guilty of being a noncontributing work partner. . . . I have resented the times when my evaluation depended in any way on someone else's performance."

Pray Muir and Tracy (1999) also provided their observations on the collaborative tests. When using collaborative testing, they noticed several changes in students' behavior. Achievement increased slightly, test anxiety decreased greatly, and students engaged in reflective thinking similar to metacognition. Moreover, in most instances, partners devoted equal effort to test preparation and content. Even though we generally are pleased with the equal effort in most teams, the few unfavorable experiences reported leave these issues somewhat unresolved.

Flavell (1976) describes Metacognition as “Knowledge and awareness of your own cognitive processes, how they function, when it’s likely to falter, etc.” Examples of metacognition are:

> “I don’t recall”
>
> “I understood this fairly well”
>
> “I won’t be able to solve this problem right away”
>
> “I can’t study with the TV on”
>
> “Her name is on the tip of my tongue”

Russo and Warren (1999) provided these observations to the use of a collaborative test. After the exam was graded, the students' scores did not increase or decrease dramatically from the previous semester's mean average when collaborative testing was not employed. The use of collaborative testing prepared students to make the

transition from school to the work world much easier, by helping them develop problem-solving techniques and reducing test anxiety. Collaborative tests encouraged the students to present their view or answer, working through the intricacies of problem solving, embedding the concepts in their minds which resulted in their understanding and retaining the material on a higher level than if the speaker had merely internalized the same information. Instead of students trying to cram or being tempted to find ways to cheat, they solved the problems collaboratively, by debating among themselves and using resources.

Social Learning Theory

Bandura (1977) stated that learning would be exceedingly laborious, not to mention hazardous, if people had to rely solely on the effects of their own actions to inform them what to do. Fortunately, most human behavior is learned observationally through modeling: from observing others one forms an idea of how new behaviors are performed, and on later occasions this coded information serves as a guide for action. Bandura went on to focus his work on self-efficacy in a varity of contexts. The most common and pervasive example of social learning situations is television commercials. Depending upon the component process involved (such as attention or motivation), we may model the behavior shown in the commercial and buy the product being advertised.

Principles:

1. The highest level of observational learning is achieved by first organizing and rehearsing the modeled behavior symbolically and then enacting it

overtly. Coding modeled behavior into words, labels, or images results in better retention than simply observing.

2. Individuals are more likely to adopt a modeled behavior if it results in outcomes they value.

3. Individuals are more likely to adopt a modeled behavior if the model is similar to the observer and has admired status and the behavior has functional value.

Interpreting Criterion-Referenced Test Results

Unless otherwise stated, the following information was extracted and used from the Quantitative Skills for Trainers Reference Book (1997). Once the test is given, the training institution needs to be able to examine the results and use them to interpret the CRT results. As previously stated, the ARRTC will use CRT whenever possible for functional based training. The test will be used to measure a student's ability to meet the objectives of the classes and course, or to the desired outcome and not as an evaluation tool to measure individual standings relative to the rest of the students.

The interpretation of the CRT results can be interpreted many ways, but limits the interpretation to either the description of the tasks that a student can perform or to comparison of the student's performance to some performance standard. The "masters" of the subject area decide what is a minimum measurement of mastery or minimal competency of the subject. While this determination of degree of mastery or minimal competency does not necessarily mean complete knowledge of the subject it does mean that the student meets a certain percentage of acceptable percentage (standard) correct. It

is critical to remind those who develop and interpret CRT that the focus is on obtaining as clear of a description as possible of the student's performance on the test.

When using the performance description method of interpretation, it is understood that the CRT is commonly used to measure mastery or minimal competency. An assumption is that there is a standard cut-off score for this type of interpretation. However, there is no requirement to have a pre-established cut-off score. The criterion used in CRT is the domain of the task being measured. For example, the student performance on a set of tasks for a course in designing instruction could include and be described as follows:

- Distinguish between properly and improperly stated instructional objectives.
- State instructional objectives as learning objectives.
- Identify technical flaws in test items.
- Construct test items that are free of technical defects.

If there is a need to perform a more detailed analysis of test results, an item-by-item analysis is useful in identifying student learning errors. Table 1 is an Item-Response Table that can be used to analyze the students' performance.

Table 1

Item-Response Table

OBJECTIVE	**CHANGING A FLAT TIRE**											
CONTEXT AREAS	**Secure Car**			**Remove Flat Tire**			**Install Spare Tire**			**Unsecure Car**		
TEST ITEMS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Student 1	+	+	+	+	+	-	+	+	+	+	+	+
Student 2	+	+	+	+	+	-	-	+	+	+	+	+
Student 3	+	+	+	+	-	-	-	+	+	+	+	+
Student 4	+	+	-	-	+	-	-	+	+	-	+	+
Student 5	+	-	+	+	-	-	-	-	-	+	-	+
Student 6	+	-	-	-	-	-	-	+	-	-	-	+

The rows specify the performance for each student. The columns highlight the class response patterns for individual subjects (items) and clusters of subjects (items). The item-response-chart is also useful for tracking learning improvement since the chart provides the information for correcting the individual student weaknesses.

This same chart can be used to check the test in the development and validation phase of SAT and the instruction given later based on the student answers. If the test item is valid and a large number of students answer an item incorrectly, it is possible that the instruction, the materials, the reference materials, or possibly the test instructions are faulty.

When using test performance standards as a method of interpretation, students are divided into two groups. Those students who have mastered a given set of tasks and those who have not. To establish the standard of performance or cut-off score for this type of evaluation, those who are the subject matter experts or “masters” may choose to use speed, a precision level, or the percentage of items answered correctly. The most widely

used measurement is the percentage correct score to determine mastery and to report CRT results.

When determining test performance standards using CRTs to distinguish between those students who have mastered a given set of tasks and those who have not, there needs to be a test performance standard or cut-off score. The test standard may be based on accuracy, speed or a combination of both, as shown in Table 2.

Table 2

Test Performance Standards Examples

TYPE OF STANDARD	EXAMPLE
Speed	Solves ten computational problems in two minutes.
Precision	Measures an obtuse angle to the nearest whole degree.
Percentage of Items Correct	Defines 85% of the basic terms.

Setting performance standards on a CRT is both difficult and frustrating because of the many issues involved with the process and so few clear guidelines to follow. A simple and practical procedure is to arbitrarily set a standard based on a best guess and then adjust the percentage up or down, as shown in Table 3.

Table 3

Performance Standards Setting Procedure

STEP	PROCEDURE
1	Set mastery level on a multiple-choice test at 85% correct.
2	Increase the level if essential for the next stage of instruction.
3	Increase the level if essential for safety.
4	Increase the level if test or sub-test is short.
5	Decrease the level if repetition occurs at next stage.
6	Decrease level if tasks have low relevance.
7	Decrease level if tasks are extremely difficult.
8	Adjust the level up and down as teaching experience dictates.

Each test needs to be evaluated and adjusted, since test types vary. For short-answer items, a lower scale would be used (example 80% correct), while a higher score should be used for true-false items (95% correct). This allows for the differences in scores based on a comparison between filling out an answer vs. 50/50 chance where guessing plays a factor.

Test Item Evaluation

The next step in this process is the need to determine what method to use to perform test item evaluation on a test. Remember, at this stage the "students" are "masters" and "nonmasters" and neither group has received instruction in the area. There are several different methods to choose from.

The first method is called a Test Item Response Profile. Used primarily with NRT, it could also be used for CRT. Use of this profile determines whether a test item is appropriate or effective. The index is a frequency of student responses to test item options and can also be used to identify confusing, ambiguous, or flawed test items. In

the case of multiple choice, true-false, or matching tests, it will identify how well the distracters affect student choices when the student is forced to make a choice.

The second method is called the Phi Coefficient. It is used to correlate the results of two activities that experience the same thing. Used for CRT, the method can be used to determine test item discrimination, test validation, or test reliability. This method is the technique of choice when items or tests are scored as "Pass/Fail" or "Go/No Go". Additionally, the method can be used when individual test items are given point values. When used in this manner, it is necessary to set a "Pass/Fail" cut-off score for each item.

This technique can be used whether an item or test is a good discriminator between "masters" and "nonmasters". Bad items or tests are incapable of making this type of discrimination. When using the Phi Coefficient the "masters" and "master groups" must contain about the same number of individuals.

The Phi Coefficient is a measurement of the difference between those who were successful in both groups compared to those in the same group. The range of index values of Phi Coefficient is from –1.00 to +1.00. The test or test item is considered a better discriminator as its index moves toward +1.00. TRADOC recommends accepting an item or test that has a value of +.50 or greater.

Using the Phi Coefficient as the test item discriminator index can be determined in two different ways. These will be referred as the short and long methods. Comparing the two methods shows that the short method (see Table 4) is slightly less reliable than the long method (see Table 5).

Table 4

Phi Coefficient Test Item Discriminator Index Short Method

STEP	PROCEDURE
1	Determine the number of "masters" who gave the correct answer.
2	Determine the number of "nonmasters" who gave the correct answer.
3	Subtract the number of "nonmasters" who gave the correct answer (Step 2) from the "masters" who gave the correct answer (Step 1).
4	Divide the number in Step 3 by ½ of the total number of "masters" and "nonmasters" responding to the test item to get the phi coefficient. Round to the nearest hundred (2 decimal places).

Phi Coefficient Test Item Discriminator Index Short Method Example-

Assume that a group of 6 "masters" responded to a test item. The results of their responses, -

- 12 participants (6 "masters" and 6 "nonmasters")
- 6 "masters" answered the item correctly.
- 2 "nonmasters" answered the item correctly.

$$\frac{\text{("masters")} - \text{("nonmasters")}}{\text{(1/2 of total responses)}} =$$

$$\frac{6 - 2}{6} =$$

+.67 Phi Coefficient Test Item Discriminator Index Short Method

Table 5

Phi Coefficient Test Item Discriminator Index Long Method

STEP	PROCEDURE
1	Determine the number of "masters" who gave the correct answer
2	Determine the number of "masters" who gave the incorrect answer
3	Determine the number of "nonmasters" who gave the correct answer
4	Determine the number of "nonmasters" who gave the incorrect answer
5	Add number obtained in step 1 to the number obtained in step 2
6	Add the number obtained in step 3 to the number obtained in step 4
7	Add the number obtained in step 1 to the number obtained in step 3
8	Add the number obtained in step 2 to the number obtained in step 4
9	Multiply the number obtained in step 1 by the number obtained in step 4
10	Multiply the number obtained in step 2 by the number obtained in step 3
11	Subtract the number obtained in step 10 from the number obtained in step 9
12	Multiply the numbers obtained in steps 5, 6, 7, and 8
13	Take the square root of the number obtained in step 12
14	Divide the number obtained in step 11 by the number obtained in step 13 to determine the phi coefficient rounded to nearest 2 decimals.

Phi Coefficient Test Item Discriminator Index Long Method Example-

- 12 participants (6 "masters" and 6 "nonmasters")
- 6 "masters" answered the item correctly.
- 0 "masters" answered the item incorrectly.
- 2 "nonmasters" answered the item correctly.
- 4 "nonmasters" answered the item incorrectly.

$$\frac{(\text{Step 1})(\text{Step 4}) - (\text{Step 2})(\text{Step 3})}{\sqrt{(\text{Step 5})(\text{Step 6})(\text{Step 7})(\text{Step8})}} =$$

$$\frac{(6)(4) - (0)(2)}{\sqrt{(6)(6)(8)(4)}} =$$

$$\frac{24}{33.94} =$$

+.71 Phi Coefficient Test Item Discriminator Index Long Method

Interpretation of the Phi Coefficient Test Item Discriminator Index Short and Long Methods and Action to be Taken. Using the TRADOC acceptable index standard +.50 as our acceptability index, this item is acceptable since its index is greater than +.50 regardless of the method used (+.67 with the Short Method and +.71 with the Long Method). If the Phi Coefficient Test Reliability Discriminator Index were between –1.00 and +.50, then this would indicate that something may be wrong the test item, instructions, reference materials, or the course instruction. More detailed evaluation of the item would have to be conducted to determine whether it should be revised or discarded.

Test Validity

Test Validity addresses whether a test measures what is was intended to measure (accuracy). It is the most important single attribute of a good test. Even if other practical and technical considerations are satisfactory, the test's quality is doubtful without proof of validity. Validity is a unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree to which that evidence supports the inferences made from scores.

There are several types of techniques for determining test validity. They are as follows: Face-Related Evidence of Validity, Content-Related Evidence of Validity, Criterion-Related Evidence of Validity, Phi Coefficient, and Agreement Index. The researcher will cover each technique individually.

Face-Related Evidence of Validity also known as face validity has a superficial appearance of validity. Simply stated, it means that the test items or test looks appropriate. Experts in a specific field also call this "The Looks Right Principle". Good face validity helps to keep motivation high because people tend to try harder when a test seems reasonable. Also, good face validity may be important to public relations. This procedure is used to run a test or test item through another set of eyes to confirm that the test items or test does what you intend them to do. This is the least valid technique of validity because no empirical procedures are used to support it.

Content-Related Evidence of Validity uses logical evidence that the item content of a test is suitable for the purpose for which the test is to be used. Also called content or logical validity, it answers the question, "How adequately does the sample of tests items represent the domain of content to be measured?" Content Validity is the most common type of validation trainers use to determine if the test provides an accurate assessment of the instructional objectives. The key element in Content Validity is the adequacy of the sampling of the subject material. A test is always a sample of the many questions that could be asked. Content validation is a matter of determining whether the sample is representative of the larger domain of content it is suppose to represent. Content Validity is similar to face validity, but content is more systematic and more sophisticated. Like

face validity, no empirical procedures are used to establish Content Validity (i.e., it is non-statistical). Instead, the test items are examined in detail and are individually analyzed and compared with the levels of behavior specified in the instructional objectives. If the test items measure exactly what the associated objectives(s) calls for, the test is content valid; otherwise it isn't. This technique (see Table 6) can be used with CRT or NRT.

Table 6

Content-Related Evidence of Validity

STEP	PROCEDURE
1	Check to be sure the objectives have been properly derived from an analysis of what the students must know and/or do in order to meet the objectives(s).
2	Check to be sure that the test specifications and/or a test plan have been developed to specify the sample of items to be used for each objective/domain.
3	Check each test item against its associated objective to see if the item measures exactly what the objective says should be measured. Be sure that the test item covers all aspects of the objective. Compare the item with the test specifications and/or test plan.

Content Validity is best described as an absolute measurement. From the absolute point of view, the results of a CRT suggest that either the student does or does not possess the ability to adequately perform the task which the objective defines. If the test items(s) are precisely matched, the test is content valid. If all items are not precisely matched to their associated objectives, the test is not content valid. For example, if the objective involves applying a concept, which has three characteristics, the items must include all three characteristics. If there are many items on a test associated with one

objective, each item must be measured against what the objective indicates. Table 7 illustrates a one-item criterion.

Table 7

Criterion-Referenced Test (CRT) Example

OBJECTIVE	CRT (ONE ITEM)
Given the appropriate tools, perform routine preventative maintenance on the 45 KW generator as specified in the operating and maintenance manual for same, within 30 minutes.	In front of you is a 45 KW generator and the appropriate tools. Perform routine preventative maintenance on the generator as specified in the operating and maintenance manual. You have 30 minutes to complete this task.

Criterion-Related Evidence of Validity also known as criterion-related or empirical validity illustrates how well the test measures the target material by indicating how relative the test is to some criterion (i.e., standard of performance). Criterion-related validity answers the question, “How accurately does test performance predict future performance or estimate present performance based on some criterion or other valued measure?”

A coefficient or correlation is used to express the degree of the relationship between a set of scores and some criterion-measure and them is called a validity coefficient. Skill is required to interpret validity coefficients because they are influenced by many factors. In general, the higher the value of the correlation coefficient the better the correlation is between the test and the criterion.

There are other factors that have to be considered: different test variables, criteria, groups, dispersion or variability; practical factors such as race, gender, etc.; and

additional information. These considerations show why a high validity coefficient is good, but it is not the entire story. Relatively speaking, the statement will be true; so the validity coefficient must be judged on a relative basis with the larger coefficient being more acceptable.

The Phi Coefficient can be used as a correlation coefficient because it correlates the results of two activities that experience the same thing. It will be used here to measure the relationship between the results of two criterion-measures (two activities) on one group of students (same thing). There are two types of studies used to obtain criterion-related evidence of validity.

1. Concurrent Study is used to test performance to estimate current performance on some criterion and is sometimes referred to as Concurrent Validity.

2. Predictive Study uses test performance to predict future performance on some other valued measure and is sometimes referred to as Predictive Validity.

Sometimes these two examples have been treated as separate types of validity, but they are both considered as examples of empirical validity since they differ only in time sequence.

Criterion-Related Evidence of Validity (Concurrent Validity) compares a teacher-made test with a similar assessment measure. A major reason for the establishment of Concurrent Validity is to substitute a test for a more time-consuming or more complex measuring instrument. In Concurrent Validity, both test scores and criterion values are obtained at almost the same time. In CRT, individuals' results on the CRT are compared

with their results on some other measure of the performance being tested. The other measure must be the best available assessment of performance on the objective(s) in question.

A quantitative determination of the association between the results on the test and the other performance measures will be provided as an estimate of the Concurrent Validity of the test. Other performance measures commonly used to establish Concurrent Validity with a test may conclude-

- Existing tests already in use.
- Instructor ratings of students' performance.
- Higher fidelity versions of the test being validated.

Remember, you must ensure that your sample is representative of the population for which the test is intended. Random sampling from the population can accomplish this. Additionally, your sample must be relatively large, preferably around 100 individuals.

Once you have chosen the other measure to use in establishing the Concurrent Validity of the test, you can use the Phi Coefficient or the Agreement Index to make that quantitative determination.

Examples of measurement instruments that can be used in Concurrent Validity are as follows:

1. A CRT on first aid techniques may be validated against instructor ratings of first aid achievement.

2. A CRT on first aid techniques may be validated against an existing first aid test that has worked well.

3. A multiple-choice test on vocabulary may be validated against a supply-response type version of a vocabulary test. (The supply-response type test has a higher fidelity than the multiple-choice test.)

The other measure of performance must be a suitable one. If you don't have another measure that you consider suitable, you cannot establish Concurrent Validity.

Concurrent Validity Example-

In the past, instructors' ratings of students' leadership skills have been used-reportedly with good results. To establish the Concurrent Validity of a new CRT, the instructor evaluated a sample group for leadership skills. Students were categorized and evenly distributed into two groups as acceptable or unacceptable based on the results of the instructors' ratings. Next, the two groups were tested using the new CRT and then scored as either passing or failing. The Phi Coefficient Concurrent Validity Short Method can be applied to determine the Concurrent Validity of the new CRT.

Total sample population = 84 students.

Acceptable leadership skills rating = 42 students.

Unacceptable leadership skills rating = 42 students.

Number of acceptable students who passed the CRT = 36

Number of unacceptable students who passed the CRT = 2

Phi Coefficient Concurrent Validity Short Method Example-

$$\frac{(\#\text{ of Acceptable Passing CRT}) - (\#\text{ of Unacceptable Passing CRT})}{(\frac{1}{2}\text{ Total taking the CRT})}$$

$$\frac{36 - 2}{\frac{1}{2}(84)} =$$

$$\frac{34}{42} =$$

+.81 Phi Coefficient Concurrent Validity Short Method

Interpretation of the Concurrent Validity Short Method Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above, the new CRT appears valid and useful for determining student leadership.

Criterion-Related Evidence of Validity, Predictive Study or Predictive Validity compares test performance with a future outcome. A good example would be using 9th grade SAT scores to predict success in the 10th grade high school chemistry. In Predictive Validity, there must be some lapse of time (6 months or more) between testing and obtaining the criterion values. Predictive Validity tell the extent to which results on the test predicts results on the job. Typical types of measures used in Predictive Validity may include:

- Supervisor's ratings of on-the-job performance.
- Other existing tests.
- Peer ratings of on-the-job performance.
- Objective indices of on-the-job performance, such as amounts of products turned out per day (acceptable or unacceptable), numbers of mistakes committed (acceptably few or unacceptably many).

The same cautions that apply to Concurrent Validity also apply to Predictive Validity.

1. The validation sample must be representative of the population for which the test is intended. Random sampling from the population can accomplish this.

2. The sample must be relatively large, preferably around 100 individuals.

3. The measures against which you will validate the test must be suitable – not just the only measures available.

Phi Coefficient is used to correlate the results of two activities that experience the same thing. Used for CRT, the method can be used to determine test item discrimination, test item analysis, test validation, or test reliability. This method is the technique of choice when items or tests are scored as "Pass/Fail" or "Go/No Go". Additionally, the method can be used when individual test items are given point values. When used in this manner, it is necessary to set a "Pass/Fail" cut-off score for each item.

This technique can be used whether an item or test is a good discriminator between "masters", individuals who have command of the material being tested, and "nonmasters", individuals who know nothing about the material. Bad items or tests are incapable of making this type of discrimination. When using the Phi Coefficient the "masters" and "non-masters" groups must contain about the same number of individuals.

The Phi Coefficient is a measurement of the difference between those who were successful in both groups compared to those in the same group. The range of index values of Phi Coefficient is from –1.00 to +1.00. The test or test item is considered a better

discriminator as its index moves toward +1.00. TRADOC recommends accepting an item or test that has a value of +.50 or greater.

Using the Phi Coefficient as the test discriminator index can be determined in two different ways. These will be referred as the short and long methods. Comparing the two methods shows that the short method (see Table 8) is slightly less reliable than the long method (see Table 9).

Table 8

Phi Coefficient Test Discriminator Index Short Method

STEP	PROCEDURE
1	Determine the number of "masters" who passed the test.
2	Determine the number of "nonmasters" who passed the test.
3	Subtract the number of "nonmasters" who passed the test (Step 2) from the "masters" who passed the test (Step 1).
4	Divide the number in Step 3 by ½ of the total number of "masters" and "nonmasters" responding to the test to get the Phi Coefficient. Round to the nearest hundred (2 decimal places).

Phi Coefficient Test Discriminator Index Short Method Example-

Assume that a group of 6 "masters" responded to a test. The results of their responses-

- 6 "masters" passed the test.
- 2 "nonmasters" passed the test.

$$\frac{\text{("masters")} - \text{("nonmasters")}}{\text{(½ of total responses)}} =$$

$$\frac{6 - 2}{6} =$$

+.67 Phi Coefficient Test Discriminator Index Short Method

Table 9

Phi Coefficient Test Discriminator Index Long Method

STEP	PROCEDURE
1	Determine the number of "masters" who passed the test.
2	Determine the number of "masters" who failed the test.
3	Determine the number of "nonmasters" who passed the test.
4	Determine the number of "nonmasters" who failed the test.
5	Add number obtained in step 1 to the number obtained in step 2.
6	Add the number obtained in step 3 to the number obtained in step 4.
7	Add the number obtained in step 1 to the number obtained in step 3.
8	Add the number obtained in step 2 to the number obtained in step 4.
9	Multiply the number obtained in step 1 by the number obtained in step 4.
10	Multiply the number obtained in step 2 by the number obtained in step 3.
11	Subtract the number obtained in step 10 from the number obtained in step 9.
12	Multiply the numbers obtained in steps 5, 6, 7, and 8.
13	Take the square root of the number obtained in step 12.
14	Divide the number obtained in step 11 by the number obtained in step 13 to determine the Phi Coefficient rounded to nearest 2 decimals.

Phi Coefficient Test Discriminator Index Long Method Example-

- 6 "masters" who passed the test.
- 0 "masters" who failed the test.
- 2 "nonmasters" who passed the test.
- 4 "nonmasters" who failed the test.

$$\frac{(\text{Step 1})(\text{Step 4}) - (\text{Step 2})(\text{Step 3})}{\sqrt{(\text{Step 5})(\text{Step 6})(\text{Step 7})(\text{Step8})}} =$$

$$\frac{(6)(4) - (0)(2)}{\sqrt{(6)(6)(8)(4)}} =$$

$$\frac{24}{33.94} =$$

+.71 Phi Coefficient Test Discriminator Index Long Method

Interpretation of the Phi Coefficient Test Discriminator Index Short and Long Method Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above, both the Short and Long method examples are acceptable for this particular test since both are higher that the pre-established minimum of +.50 or higher, +.67 and +.71 respectively. If the Phi Coefficient Test Discriminator Index Long Method Data were between –1.00 and +.50, this would have indicated that something maybe wrong with the test, instructions, reference materials, or the course instruction. In this case, more detailed evaluation of the test would have to be conducted to determine whether it should be revised or discarded.

The Agreement Index (see Table 10) technique can be used when items or tests are scored “Pass/Fail” or “Go/No Go”. This technique is similar to the Phi Coefficient and is appropriate when determining the discrimination of tests or the validity of a CRT. Although its values are generally proportional to the Phi Coefficient, it is considered less reliable.

The Agreement Index is used to determine whether an item or a test item or an overall test is a good discriminator between ““masters”” and “nonmasters.” When using

the Agreement Index, the “masters” and “nonmasters” groups must contain about the same number of individuals.

This index is a nonstatistical value that can be either a positive or negative number. A zero or a negative number means that there could be a problem, which requires a closer evaluation to determine if a revision or if the item needs to be discarded. This method is a quick way to determine that quality of tests or test items when time does not permit using the Phi Coefficient or other more complicated methods. The Agreement Index can also be used to determine the Concurrent or Predictive Validity of a CRT. When validating tests, score each student as a “Go/No Go” (“Pass/Fail”) and use the same steps in determining test item discrimination to determine a realistic validity index.

The only difference between using the procedure to determine validity and using the procedure to determine item discrimination is that the variable changes. In item discrimination, it is the number of students who answered the item correctly/incorrectly. In test validity, it is the number of students who passed/failed the test as compared to the similar assessment. The guidelines are the same for both applications (i.e., a positive value).

Table 10

Test Item Discriminator Agreement Index

STEP	PROCEDURE
1	Determine the number of "masters" who passed the test.
2	Determine the number of "masters" who failed the test.
3	Determine the number of "nonmasters" who passed the test.
4	Determine the number of "nonmasters" who failed the test.
5	Multiply the number of "masters" who passed the test (step 1) with the number of "nonmasters" who failed the test (step 4).
6	Multiply the number of "masters" who failed the test (step 2) with the number of "nonmasters" who passed the test (step 3).
7	Subtract the number obtained in step 6 from the number obtained in step 5 to get the Agreement Index.

Test Item Discriminator Agreement Index Example-

- 5 "masters" passed the test.
- 1 "master" failed the test.
- 3 "nonmasters" passed the test.
- 3 "nonmasters" failed the test.

(passed "masters") x (failed "nonmasters") – (failed "masters") x (passed "nonmasters")

(5 x 3) – (1 x 3) = 15 – 3 = +12 Test Item Discriminator Agreement Index

Interpretation of the Test Item Discriminator Agreement Index Data and Action to be Taken. With +12 as the Agreement Index, this test is acceptable since it is a positive number. No action is required. If the number had been zero or a negative number, this would have indicated a problem, which may need further examining and a possible revising or discarding.

Test Reliability

Test reliability addresses whether a test gives dependable or consistent scores. Reliability refers to the consistency of a set of test results (precision). Next to validity, reliability is the most important measure of a test's quality. Good test reliability is necessary, but not sufficient for good quality. A test that is not valid, but is highly reliable, may be measuring the wrong thing with great precision. A test with high reliability is one that will produce very much the same relative scores for a group of the same type of individuals under different conditions or situations. Using Test-Retest, Parallel Forms, Test-Retest with Parallel Forms, Factors Affecting Reliability, Phi Coefficient, and Percent of Consistency techniques, a CRT can be systematically assessed.

Test-Retest Reliability is a method of establishing consistency by correlating scores for the same test to the same group of students in a given time interval. This will provide evidence of stable test scores over a period of time.

The length of time interval should fit the type of interpretation to be made from the results. For example, if we are interested in using the test scores only to group students for more effective learning, short-term stability may be sufficient. If we were attempting to predict vocational success or make another long-range predictions, stability over a period of years would be desired.

If the test is reliable, students who pass the first time should pass the second time, while students who fail the first time should fail the second time.

Use the following precautions when using the Test-Retest method for determining reliability:

1. Use a sample group of at least 30 students for the test reliability. These students should be chosen randomly from the student population who would ordinarily take this test.

2. Longer time periods between testing will result in lower reliability coefficients because there will be greater changes in the students. For classroom tests, it is recommended that the test to be administered be the same test to the same group of students twice. Allow only one day between the administration of the two tests.

3. Do not tell the students that they will be tested again. You do not want them to practice or gain any skills/knowledge between test administrations or try to recall the test in detail. Test-Retest reliability assumes that no practice or gain in learning has occurred between the administration of the tests.

4. Test the students under the same conditions both times, including the same time of day. Test-Retest reliability assumes equivalent conditions. Students should be equally hungry, tired, rested, etc., during each administration.

Normally, reliability is expressed as a correlation coefficient reported on a scale ranging from 0.00 to +1.00. If the results show the students' scores were approximately the same on both administrations of the test, then a positive relationship would exist (+1.00 indicates a perfect positive relationship; 0.00 indicates no relationship).

The Test-Retest Phi Coefficient long method can be used as a simple estimate of the Test-Retest reliability, especially for a CRT where scores are reported as “Go/No Go” (“Pass/Fail”).

Test-Retest Phi Coefficient Long Method Example-

You want to determine the reliability of a new CRT using Test-Retest method. 30 students are chosen, sampled randomly from a population who would ordinarily be given the test. The test is administered (“Pass/Fail”) and scored (the students) accordingly. The next day, the same test is administered to the same group at the same time under the same conditions, and you score the second test. No learning has occurred between the first and second test. The results of both administrations of the same test were:

- A = # of students who passed the test both times = 14
- B = # of students who failed the 1st test, but passed the test the 2nd time = 5
- C = # of students who passed the 1st test, but failed the test the 2nd time = 1
- D = # of students who failed the test both times = 10

Using the long method for computing the Test-Retest Phi Coefficient Long Method you will get the following:

$$\text{Test-Retest Phi Coefficient Long Method} = \frac{(AD - BC)}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} =$$

$$\frac{(14)(10) - (5)(1)}{\sqrt{(19)(11)(15)(15)}} =$$

$$\frac{135}{\sqrt{47{,}025}} =$$

$$\frac{135}{216.85} =$$

+.62 Test-Retest Phi Coefficient Long Method

Interpretation of the Test-Retest Phi Coefficient Long Method Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above as the acceptable reliability index, the new CRT sufficiently reliable. If Test-Retest Phi Coefficient Long Method Data were between –1.00 and +.50, this would have indicated that something maybe wrong with the test, instructions, reference materials, or the course instruction. In this case, more detailed evaluation of the test would have to be conducted to determine whether it should be revised or discarded.

The Parallel Forms method of establishing reliability uses correlating scores from two different forms of the test is also called equivalent or alternate forms. This method will provide an estimate of the consistency of the test scores over different forms of the test, or different samples of items.

Use the following precautions when using the Parallel Forms method for determining reliability:

1. Use a sample group of at least 30 students for the test reliability. These students should be chosen randomly from the student population who would ordinarily take this test.

2. Administer both forms of the test at the same time to the same group.

3. Do not tell the students that they will be tested again. You do not want them to practice or gain any skills/knowledge between test administrations or try to recall the test in detail. Parallel Forms reliability assumes that no practice or gain in learning has occurred between the administration of the two tests.

4. Test the students under the same conditions both times, including the same time of day. Parallel Forms reliability assumes equivalent conditions. Students should be equally hungry, tired, rested, etc., during each administration.

Normally, reliability is expressed as a correlation coefficient reported on a scale ranging from 0.00 to +1.00. If the results show the students' scores were approximately the same on both administrations of the test, then a positive relationship would exist (+1.00 indicates a perfect positive relationship; 0.00 indicates no relationship).

The Parallel Forms Phi Coefficient long method can be used as a simple estimate of the Parallel Forms reliability, especially for a CRT where scores are reported as "Go/No Go" ("Pass/Fail"). This is like the Test-Retest method.

Parallel Forms Phi Coefficient Long Method Example-

You want to determine the reliability of a new CRT using the Parallel Forms method. 30 students are chosen, sampled randomly from a population who would ordinarily be given the test. The test is administered, Form A test ("Pass/Fail") and scored

the students accordingly. Then, Form B is administered to the same group at the same time under the same conditions, and you score the second test. No learning has occurred between Form A and B tests. The results of both administrations of the different test were:

- A = # of students who passed both forms of the test = 14
- B = # of students who failed Form A test, but passed Form B test = 5
- C = # of students who passed Form A test, but failed Form B test = 1
- D = # of students who failed both forms of the test = 10

Parallel Forms Phi Coefficient Long Method = $\frac{(AD - BC)}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$

$\frac{(14)(10) - (5)(1)}{\sqrt{(19)(11)(15)(15)}} =$

$\frac{135}{\sqrt{47{,}025}} =$

$\frac{135}{216.85} =$

+.62 Parallel Forms Phi Coefficient Long Method

Interpretation of the Parallel Forms Phi Coefficient Long Method Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above as the acceptable reliability index, the new CRT sufficiently reliable. If Parallel Forms Phi Coefficient Long Method Data were between –1.00 and +.50, this would have indicated that something maybe wrong with the test, instructions, reference materials, or the course instruction. In this case, more detailed evaluation of the test would have to be conducted to determine whether it should be revised or discarded.

The Test-Retest with Parallel Forms method of establishing reliability uses correlating scores from two different forms of the same test with time intervening. This method is a combination of two methods where reliability is estimated based on the consistency of the test scores over both a time interval and on different forms of the test.

This is the most demanding estimate of reliability, since it takes into account all possible sources of variation. For most purposes, it is probably the most useful type of reliability, since it enables the evaluator to use the reliability estimate of the test results over various conditions.

With this method, two Parallel Forms of a test are administered to the same group over a period of time. A highly reliable test indicates that a test score represents not only present test performance, but also what the test performance is likely to be at another time or on a different sample of parallel items.

Use the following precautions when using the Test-Retest with Parallel Forms method for determining reliability:

1. Use a sample group of at least 30 students for the test reliability. These students should be chosen randomly from the student population who would ordinarily take this test.

2. Administer both forms of the test to the same group close together in time. Allow only one day between the administration of the first form and second form of the test.

3. Do not tell the students that they will be tested again. You do not want them to practice or gain any skills/knowledge between test administrations or try

to recall the test in detail. Test-Retest with Parallel Forms reliability assumes that no practice or gain in learning has occurred between the administration of the two tests.

4. Test the students under the same conditions both times, including the same time of day. Test-Retest with Parallel Forms reliability assumes equivalent conditions. Students should be equally hungry, tired, rested, etc., during each administration.

Normally, reliability is expressed as a correlation coefficient reported on a scale ranging from 0.00 to +1.00. If the results show the students' scores were approximately the same on both administrations of the test, then a positive relationship would exist (+1.00 indicates a perfect positive relationship; 0.00 indicates no relationship). The Test-Retest with Parallel Forms Phi Coefficient long method can be used as a simple estimate of the Test-Retest with Parallel Forms reliability, especially for a CRT where scores are reported as "Go/No Go" ("Pass/Fail"). This is like the Test-Retest or Parallel Forms method.

The Test-Retest with Parallel Forms Phi Coefficient Long Method Example:

You want to determine the reliability of a new CRT using the Parallel Forms method. Thirty students are chosen, sampled randomly from a population who would ordinarily be given the test. The test is administered, Form A test ("Pass/Fail") and scored the students accordingly. Then, Form B is administered to the same group at the same time under the same conditions, and you score the second test. No learning has

occurred between Form A and B tests. The results of both administrations of the different test were:

- A = # of students who passed both forms of the test = 14
- B = # of students who failed Form A test, but passed Form B test = 5
- C = # of students who passed Form A test, but failed Form B test = 1
- D = # of students who failed both forms of the test = 10

$$\text{Phi Coefficient} = \frac{(AD - BC)}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} =$$

$$\frac{(14)(10) - (5)(1)}{\sqrt{(19)(11)(15)(15)}} =$$

$$\frac{135}{\sqrt{47{,}025}} =$$

$$\frac{135}{216.85} =$$

+.62 Test-Retest with Parallel Forms Phi Coefficient Long Method

Interpretation of the Test-Retest with Parallel Forms Phi Coefficient Long Method Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above as the acceptable reliability index, the new CRT sufficiently reliable. If the Test-Retest with Parallel Forms Phi Coefficient Long Method Data were between –1.00 and +.50, this would have indicated that something maybe wrong with the tests, instructions, reference materials, or the course instruction. In this case, more detailed evaluation of the tests would have to be conducted to determine whether they should be revised or discarded.

The Factors Affecting Reliability are to be considered during the initial designing of a test as well as when evaluating the reliability and validity of the test. Table 11 lists a few of the major factors to be considered.

Table 11

Reliability Factors

FACTOR	AFFECT
1. Test Length	In general, the longer a test, the more reliable it will be--provided other factors remain the same. Longer tests sample the instructional objectives better.
2. Spread of Scores/Heterogeneity	Groups of students which vary more in ability will give you higher reliability. A group of students with heterogeneous ability will produce a larger spread of test scores than will a group with homogeneous ability.
3. Item Difficulty	Tests composed of moderately difficult items will have greater reliability that those tests composed of mainly easy or difficult items.
4. Item Discrimination	Tests composed of more discriminating items will have greater reliability that tests composed of less discriminating items.
5. Time Limits	Adding a time factor may improve reliability for lower-level cognitive test items. Since all students do not function at the same pace, a time factor adds another criterion to the test that causes discrimination, thus improving reliability.
6. Shorter Time between Testing	The length of time between the administration of the two tests in a temporal reliability coefficient is obvious importance. When the time between the administration of the two tests is short, reliability is high.
7. Type of Reliability Estimate	Reliability coefficients will differ according to the type of reliability estimate being used.

The Phi Coefficient can be used to determine the reliability of CRT results over time or on parallel (equivalent) forms of a test. When determining test reliability, each student is scored as a "Pass/Fail" ("Go/No Go") on each administration of the test, and use the steps in the long method (see Table 12) only to determine a realistic reliability index. The only difference between using the long method to determine reliability and using the long method to determine item discrimination is that the variable changes. In item discrimination, it is the number of students who answer the item correctly/incorrectly. In test reliability, it is the number of students who passed/failed different administrations of a test. The guidelines are the same for the acceptability of the Phi Coefficient (i.e., the closer to +1.00, the more reliable; with +.50 being the acceptable index).

Table 12

Phi Coefficient Test Reliability Discriminator Index Long Method

STEP	PROCEDURE
1	Determine the number of "masters" who passed the test.
2	Determine the number of "masters" who failed the test.
3	Determine the number of "nonmasters" who passed the test.
4	Determine the number of "nonmasters" who failed the test.
5	Add number obtained in step 1 to the number obtained in step 2.
6	Add the number obtained in step 3 to the number obtained in step 4.
7	Add the number obtained in step 1 to the number obtained in step 3.
8	Add the number obtained in step 2 to the number obtained in step 4.
9	Multiply the number obtained in step 1 by the number obtained in step 4.
10	Multiply the number obtained in step 2 by the number obtained in step 3.
11	Subtract the number obtained in step 10 from the number obtained in step 9.
12	Multiply the numbers obtained in steps 5, 6, 7, and 8.
13	Take the square root of the number obtained in step 12.
14	Divide the number obtained in step 11 by the number obtained in step 13 to determine the Phi Coefficient rounded to nearest 2 decimals.

The Phi Coefficient Test Reliability Discriminator Index Long Method Example-

- 6 "masters" who passed the test.
- 0 "masters" who failed the test.
- 2 "nonmasters" who passed the test.
- 4 "nonmasters" who failed the test.

$$\frac{(\text{Step 1})(\text{Step 4}) - (\text{Step 2})(\text{Step 3})}{\sqrt{(\text{Step 5})(\text{Step 6})(\text{Step 7})(\text{Step8})}} =$$

$$\frac{(6)(4) - (0)(2)}{\sqrt{(6)(6)(8)(4)}} =$$

$$\frac{24}{\sqrt{1152}} =$$

$$\frac{24}{33.94} =$$

+.71 Phi Coefficient Test Reliability Discriminator Index Long Method

Interpretation of the Phi Coefficient Test Reliability Discriminator Index Long Method Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above as the acceptable reliability index, the new CRT sufficiently reliable. If the Phi Coefficient Test Reliability Discriminator Index Long Method Data were between –1.00 and +.50, this would have indicated that something maybe wrong with the test, instructions, reference materials, or the course instruction. In this case, more detailed evaluation of the test would have to be conducted to determine whether it should be revised or discarded.

The Percent of Consistency is the last method of determining an estimate of reliability for a CRT by determining the percent of students who consistency passed or failed the test. CRTs are not normally designed to emphasize differences among individuals; therefore, the spread of scores should not vary extensively.

Since a CRT is used to determine knowledge or skill mastery, our primary concern is how consistently our test classifies “masters” and “nonmasters” (those who “Pass/Fail” or get a “Go/No Go” consistency). To perform this technique, administer

parallel or equivalent forms of a CRT. Then compute the Percent of Consistency (see Table 13) on two equivalent forms of the test. This estimate or reliability is expressed in percent, with 0% being a perfect negative relationship and 100% being a perfect positive relationship.

Table 13

Percent of Consistency

STEP	PROCEDURE
1	Administer parallel or equivalent forms of a CRT.
2	Determine the number of students who passed both forms of the test.
3	Determine the number of students who failed both forms of the test.
4	Determine the total number of students who took both forms of the test.
5	Add the students who passed both forms of the test (step 2) to the students who failed both forms of the test (step 3).
6	Divide the number in step 5 by the total number of students who took both forms of the test (step 4), and multiply the results by 100 to get percent of consistency.

Percent of Consistency Example-

Suppose two forms of a 25-item test were administered to 40 students. Mastery was set at 80% correct (20 items), so all students who scored 20 or higher passed the tests. After the equivalent forms of the test were administered, the scores were computed. Of the 40 students, 30 students passed both forms of the test and 6 failed both forms of the test.

$$\text{Percent of Consistency} = \frac{(\text{step 2}) + (\text{step 3})}{(\text{step 4})} \times 100$$

$$\text{Percent of Consistency} = \frac{30 + 6}{40} \times 100 = 90\%$$

Interpretation of the Percent of Consistency and Action to be Taken. As with other measures of reliability, the greater the percentage of consistency, the more satisfied we would be. There is no minimum acceptable level of consistency. Again, this is a judgement call that depends on the number of items in the test and the consequences of the decision.

If a "Failed" decision for a student means further study and later re-testing, low consistency might be acceptable. However if "Pass/Fail" decision determines whether to give a student a certificate or diploma, then a high level of consistency would be demanded.

Since there are no clear guidelines for setting minimum levels of percent of consistency, the evaluator will need to depend on experience in various situations to determine reasonable expectations. As for this case, a 90% consistency seems appropriate and the equivalent forms of the test seem to be reliable.

Summary

Collaborative testing by itself is not a sole indicator of student's minimum competencies. It is an additional tool to be used to build on a course test plan, combining individual knowledges, skills, and abilities on an individual and a group basis. Testing done on the individual students knowledges, skills, and abilities give the instructor an indication of how well the student can complete a task in a situation where they are alone, without the assistance of others. The group or collaborative test allows the instructor to build on the basic level competencies and expand the situations where the group uses the collaborative process to research and problem solve. By applying what they have learned

individually and using the group dynamic, students will process the new information on their own, using the previous steps in the process, causing them to be interactive and with application of the new material. This will drive the knowledge deeper in their minds, assisting them in learning past the short term. This change in behavior, either from learning the process or skill for the first time or from relearning the process or skill correctly will allow them to take back to their jobs the correct way to do these tasks. More importantly, this adjustment of their behavior will give them the confidence to start working to find their own answers to the questions and situations on the job, both from an individual and group level.

Selection of the appropriate modes to conduct validity and reliability checks for CRTs will require careful consideration. The current models are based on individual test instruments. Adaptation of these models to measure collaborative tests will be necessary.

Chapter III

Research Methodology and Approach

Introduction

This chapter describes the research design, instrumentation, data collection, data processing, and study limitations. The purpose of this study was to evaluate existing individual Criterion-Referenced Test (CRT) validity and reliability processes and procedures that the Training and Doctrine Command (TRADOC) has approved for use for adaptation to evaluate the validity and reliability of collaborative CRT at the Army Reserve Readiness Training Center (ARRTC). ARRTC SAT Regulation 351-1-1 states that a systems approach to training will be used to determine if there is a need for the creation, modification, or deletion of subjects or courses. Part of this SAT process requires that a needs assessment be conducted, and if necessary, a systematic plan be developed to explore how this need will be met. There are references to creating groups in the learning environment (collaborative efforts) in the ARRTC SAT, but no processes or procedures for developing the collaborative CRT. Therefore, the need for collaborative CRT validity and reliability processes exists.

This chapter will explain the validation and reliability processes used to analyze the Enlisted and Officer Reclassification collaborative CRTs.

Research Design

The research design used was a quasi-experimental consideration between the approved validity and reliability processes for individual CRTs and their applicability to collaborative CRTs.

After reviewing the TRADOC approved individual CRTs validation and reliability processes, the researcher hypothesizes that further examination of the Content-Related Evidence of Validity and Phi Coefficient Test Reliability Discriminator Index Long Method was needed to determine the validity and reliability of collaborative CRTs.

The validity and reliability processes considered for this study were taken from the Quantitative Skills for Trainers Manual (QSTRB), a TRADOC publication. The ARRTC Systems Approach to Training (SAT) Regulation 351-1-1 refers to the TRADOC Regulation 350-70 and the TRADOC Systems Approach to Training (SAT) Desk Reference as source documents for the policy and procedures for the ARRTC SAT. The TRADOC SAT Desk Reference referred to the QSTRB as a source for additional materials and information not specified in the TRADOC SAT Regulation or TRADOC SAT Desk Reference. Since the ARRTC SAT did not specify what the process and procedures were for the validation and reliability of collaborative CRTs, research of the TRADOC references was necessary to acquire the necessary information.

The primary reason for the use of these processes was based on their limitations of sample size to perform the validation and reliability evaluations of collaborative CRTs. There were five techniques or procedures used to determine test validity and six techniques or procedures used to determine test reliability defined in the QSTRB. The Content-Related Evidence of Validity process does not require “masters” and “nonmasters” to participate in determining the collaborative CRT validity. “Masters” are defined as, individuals who have command of the material being tested. “Nonmasters” are defined as, individuals who know nothing about the material being tested. The Phi

Coefficient Test Reliability Discriminator Index Long Method requires the amount of "masters" and "nonmasters" to be the same size, but no minimum number of each to determine the reliability of collaborative CRT results. Modification of the Content-Related Evidence of Validity process will be called the Collaborative Content Validity Procedure and the Phi Coefficient Test Reliability Discriminator Index Long Method will be modified and called the Collaborative Phi Coefficient Test Reliability Discriminator Index Long Method Procedure or simply the Collaborative Phi Coefficient.

When evaluating a collaborative CRT for validity, the Collaborative Content Validity Procedure, was used to provide logical evidence that the test was suitable for the purpose for which the test was to be used. Originally developed for individual CRT validation, the same procedure was used to validate a collaborative CRT for a small group of individuals (no fewer than two and no more than five individuals per group). The only difference between an individual and a collaborative CRT is who answers the questions. The individual answers the CRT questions for an individual CRT and the group consensus was used to answer the same question for the collaborative CRT. The Collaborative Content Validity Procedure answered the question, "How adequately does the test represent the domain of content to be measured?" The key element in the Collaborative Content Validity Procedure was the adequacy of the sampling of the subject material. A test is a sample of the many questions that could be asked. Collaborative Content Validity is a matter of determining whether the sample was representative of the larger domain of content it was suppose to represent. Collaborative Content Validity Procedure did not use empirical procedures (non-statistical). Instead, the

collaborative CRT was examined in detail, analyzed, and compared with the levels of behavior specified in the instructional objectives. If the collaborative CRT measures exactly what the associated objectives(s) calls for, the collaborative CRT is Content Valid; otherwise it is not. The original Content Validity Procedure is the most common type of validation trainers use to determine if the collaborative CRT provides an accurate assessment of the instructional objectives.

When evaluating a collaborative CRT for reliability, The Collaborative Phi Coefficient Test Reliability Discriminator Index Long Method Procedure, also known as the Collaborative Phi Coefficient, was used to provide logical evidence that the test was suitable for the purpose for which the test was to be used. Originally developed for evaluating individual CRT reliability, the same procedure was used to evaluate a collaborative CRT for a small group of individuals (no fewer than two and no more than five individuals per group). When used to determining collaborative CRT reliability, each group collaborative CRT was scored as a "Pass/Fail" ("Go/No Go") on each administration of the test and the steps in the long method was used to determine a realistic reliability index. In test reliability, it was the number of groups whom passed/failed different administrations of a test. The difference between evaluating the reliability of an individual CRT and a collaborative CRT was the total number of individuals needed to validate the test. To evaluate the reliability of an individual CRT, a minimum of six individuals, three "master" individuals and three "nonmaster" individuals would be needed. To evaluate the reliability of a collaborative CRT, a minimum total of 12 individuals, six individuals for the three "master groups" of two

individuals each and six individuals for the three "nonmaster groups" of two individuals each would be needed.

Instrumentation

After establishing a need to create or revise a test, the Subject Matter Expert (SME) wrote the test, using the course Training Learning Activity Worksheet (TLAW), the Program of Instruction (POI), and the Test Plan. The test was then submitted for evaluation to the Instructional Systems Specialist (ISS). The ISS analyzed the test, compared the test objectives to the TLAW, the POI, and the Test Plan for accuracy and applicability. This was done using the Collaborative Content Validity Procedure (see Table 14) as the procedural guide.

Table 14

Collaborative Content Validity Procedure

STEP	PROCEDURE
1	Check to be sure the objectives have been properly derived from an analysis of what the group must know and/or do in order to meet the objectives(s).
2	Check to be sure that the test specifications and/or a test plan have been developed to specify the sample of items to be used for each objective/domain.
3	Check each test item against its associated objective to see if the item measures exactly what the objective says should be measured. Be sure that the test item covers all aspects of the objective. Compare the item with the test specifications and/or test plan.

Using the Collaborative Content Validity Procedure, the ISS compared the test Action, Condition, and Standard, which orients the students on their role, what materials were needed, and what the expected outcome was of the test.

The ISS compared the stated test objective (see Appendix A) to the TLAW (see Appendix B) and POI (see Appendix C), ensuring that the objective was the same. Then, the ISS compared the test objective/domain to the Test Plan (see Appendix D), ensuring that they match. Finally, the ISS compared the test stated Action, Condition, and Standard to the TLAW and POI, ensuring that the Action, Condition, and Standard were the same (see Figure 1). If any of these did not match, the test would be returned to the SME for revision.

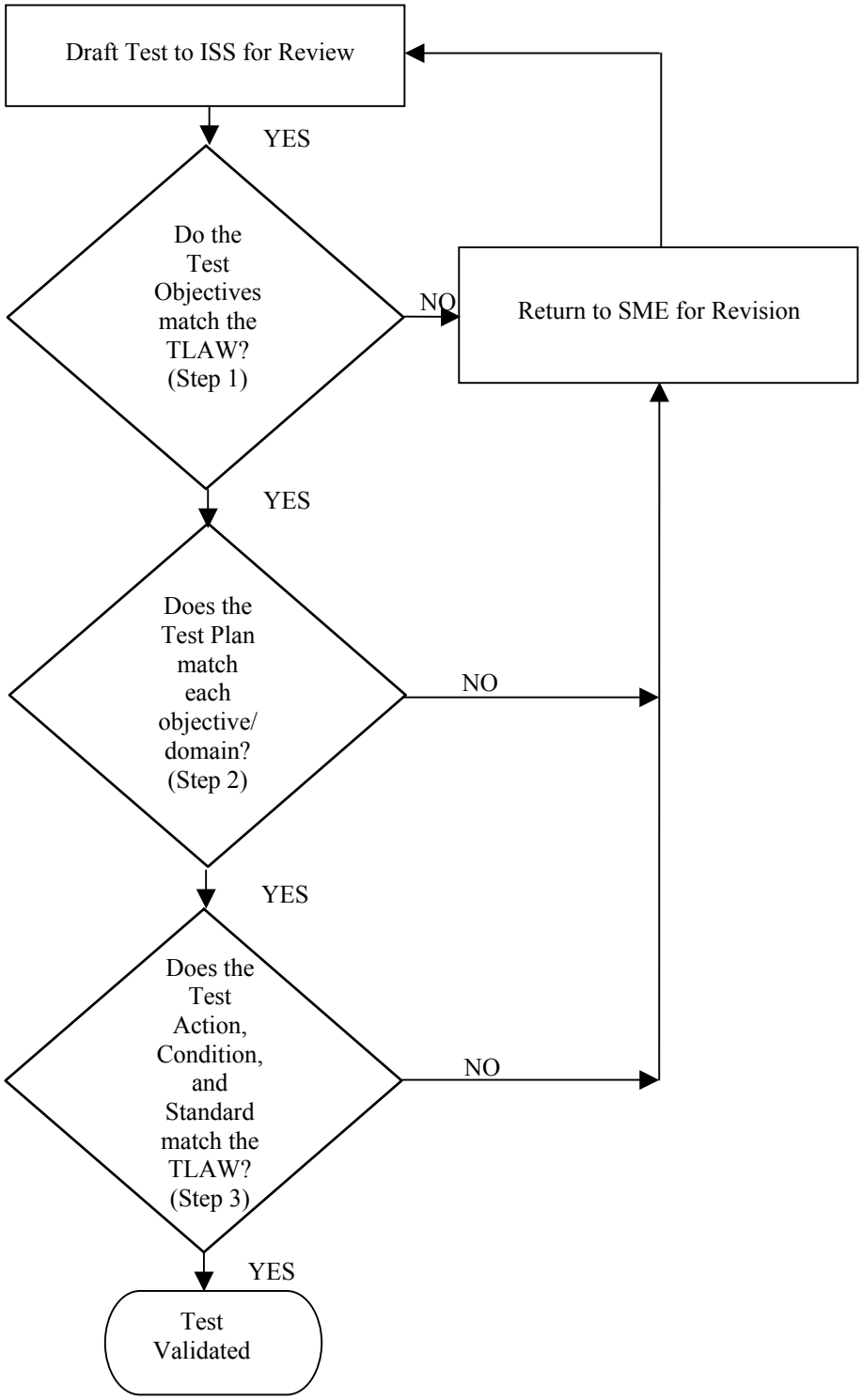

Figure 1. Procedural steps when validating a collaborative CRT.

The Collaborative Content Validity Procedure is best described as an absolute measurement. From the absolute point of view, the results of a CRT suggest that either the test does or does not possess the ability to adequately perform the task the objective defines. If the test items(s) are precisely matched, the test is Content Valid. If all items are not precisely matched to their associated objectives, the test is not Content Valid. For example, if the objective involves applying a concept, which has three characteristics, the items must include all three characteristics. If after using the Collaborative Content Validity Procedure the test met the TLAWS, POI, and Test Plan, the test was considered content valid and was moved to the next step in the review process.

After the test had been found to be valid, the next step was to evaluate the reliability of the test. Using the Collaborative Phi Coefficient, the ISS would conduct the test, using the same number of "master groups" and "nonmasters groups." Each "master group" would consist of individuals who were considered experienced in the job or task being tested. Each "nonmasters group" would be comprised of individuals who would be doing the job or task, but had not received training on the job or task. The groups could only receive help from within their group and not from others taking or administering the test. The test would be given under the same conditions as in the classroom except that training would not be conducted prior to the test.

Upon completing the test, the groups would turn in their test answers for grading. Once the test was graded, the ISS would analyze the test results using the Collaborative Phi Coefficient (see Table 15) as a procedural guide. In test reliability, it is the number of groups whom passed/failed different administrations of a test. The guidelines are the

same for the acceptability of the Collaborative Phi Coefficient (i.e., the closer to +1.00, the more reliable, with +.50 being the acceptable index).

The ISS would coordinate within the ARRTC, getting a total of 12 individuals, six individuals for the three "master groups" of two individuals each and six individuals for the three "nonmaster groups" of two individuals each to participate in this test. Additional coordination would be made to have a classroom with the same automation equipment, applications, seating arrangements, and courseware that the classroom students would have in the test setting. No instruction was given to the test individuals and the test individuals would de directed to follow the directions just as the students in the classroom would be directed. They could only work with those in their group and not between groups to come to consensus to finish the test. After collecting the test individuals' responses, the ISS would correct the tests.

Table 15

Collaborative Phi Coefficient

STEP	PROCEDURE
1	Determine the number of "master groups" who passed the test.
2	Determine the number of "master groups" who failed the test.
3	Determine the number of "nonmaster groups" who passed the test.
4	Determine the number of "nonmaster groups" who failed the test.
5	Add number obtained in step 1 to the number obtained in step 2.
6	Add the number obtained in step 3 to the number obtained in step 4.
7	Add the number obtained in step 1 to the number obtained in step 3.
8	Add the number obtained in step 2 to the number obtained in step 4.
9	Multiply the number obtained in step 1 by the number obtained in step 4.
10	Multiply the number obtained in step 2 by the number obtained in step 3.
11	Subtract the number obtained in step 10 from the number obtained in step 9.
12	Multiply the numbers obtained in steps 5, 6, 7, and 8.
13	Take the square root of the number obtained in step 12.
14	Divide the number obtained in step 11 by the number obtained in step 13 to determine the Phi Coefficient rounded to nearest 2 decimals.

Interpretation of the Collaborative Phi Coefficient Data and Action to be Taken. The ISS would grade the tests and use the test scores to calculate the Collaborative Phi Coefficient Index. If the Collaborative Phi Coefficient results were +.50 or higher, then using the TRADOC acceptable index standard of +.50 or above as the acceptable reliability index, the new CRT would be sufficiently reliable. If the Collaborative Phi Coefficient results were between –1.00 and +.50, this would have indicated that something maybe wrong with the test, instructions, reference materials, or the course instruction. This would result in a more detailed evaluation of the test to determine whether it should be revised or discarded.

The Collaborative Phi Coefficient was originally designed for evaluating individual test reliability. When used to determine collaborative test reliability, each group collaborative test was scored. Each collaborative test was scored as a "Pass/Fail" ("Go/No Go") on each administration of the test and the steps in the long method are used to determine a realistic reliability index.

Data Collection

In order to evaluate the effectiveness of collaborative CRT, two collaborative comprehensive CRTs were used. Both collaborative CRTs came from the Personnel Management Course (PMC). Each collaborative CRT was processed by the ISS, using the Content Validity Procedure for validation and the Phi Coefficient Test Reliability Discriminator Index Long Method Procedure for reliability.

Data Processing

First, each collaborative CRT was examined by the ISS, whom performed the validity check on each test, using the Collaborative Content Validity Procedure (see Table 14).

Then, each collaborative CRT was examined by the ISS, whom performed the reliability check on each test, using Collaborative Phi Coefficient (see Table 15).

Summary

This chapter described the research design, instrumentation, data collection, data processing, and study limitations. The purpose of this study was to evaluate existing individual Criterion-Referenced Test (CRT) validity and reliability processes and procedures that the Training and Doctrine Command (TRADOC) has approved for use

for adaptation to evaluate the validity and reliability of collaborative CRT at the Army Reserve Readiness Training Center (ARRTC). Chapter four describes the results of these processes and techniques.

Chapter IV

Findings of Analysis and Results

Introduction

The purpose of this study was to meet the requirements of ARRTC SAT Regulation 351-1-1, which states that a systems approach to training is used to determine if there is a need for the creation, modification, or deletion of subjects or courses. Part of this SAT process required that a needs assessment be conducted, and if necessary, a systematic plan be developed to explore how the need can be met. There were references to creating groups (collaborative efforts) in the ARRTC SAT, but no process for developing collaborative Criterion-Referenced Tests (CRT). Therefore, the need for collaborative CRT development existed. This study developed a process that identified the steps necessary to create a collaborative testing process to measure the validity and reliability for CRT. This research examined existing validity and reliability processes and procedures for individual CRT employed by the Training and Doctrine Command (TRADOC) and determined which individual CRT validity and reliability process or procedure can be used to validate and measure the reliability of a collaborative CRT.

This chapter reports the findings of the collaborative CRT validity and reliability processes of the Enlisted and Officer Reclassification collaborative CRTs and the interpretations of the results.

Findings of the PMC Reclassification Collaborative CRTs Validity Process

First, the Instructional Systems Specialist (ISS), whom performed the validity check on each test, examined each Personnel Management Course (PMC) collaborative CRT. The Collaborative Content Validity Procedure was used and produced the following information.

Enlisted Reclassification Collaborative CRT. Following the Collaborative Content Validity Procedure (see Table 16), the results of the Enlisted Reclassification Collaborative CRT were as follows.

Table 16

Collaborative Content Validity Procedure

STEP	PROCEDURE
1	Check to be sure the objectives have been properly derived from an analysis of what the group must know and/or do in order to meet the objectives(s).
2	Check to be sure that the test specifications and/or a test plan have been developed to specify the sample of items to be used for each objective/domain.
3	Check each test item against its associated objective to see if the item measures exactly what the objective says should be measured. Be sure that the test item covers all aspects of the objective. Compare the item with the test specifications and/or test plan.

The first step was to compare the objective on the PMC Enlisted Reclassification Test objective (see Appendix A) to the PMC Training Learning Activity Worksheet (TLAW) (see Appendix B) objective. The PMC Enlisted Reclassification Test objective was to Validate Enlisted Request for Reclassification. The PMC TLAW objective was to Validate Enlisted Classification/Reclassification Requirements. Finally, the PMC

Program of Instruction (POI) (see Appendix C) objective was to Validate Enlisted Classification/Reclassification Requirements. Based on the comparison of these instruments, all are in agreement on the objective.

The second step was to compare the PMC Enlisted Reclassification Test sample for each objective/domain to the PMC Course Test Plan (see Appendix D). The PMC Enlisted Reclassification Test required the students to validate the request, making any necessary corrections on the form itself. The PMC Course Test Plan required the students, using analysis, to use a hands-on performance test to demonstrate that the students could validate an Enlisted Reclassification Request. Based on this comparison, these instruments are in agreement.

The third step was to compare each PMC Enlisted Reclassification Test item and its associated objective to the PMC Test specifications and/or PMC Course Test Plan, ensuring that the PMC Enlisted Reclassification Test item measures exactly what the objective says should be measured. The PMC Enlisted Reclassification Test Action was, Validate Enlisted request for reclassification. The PMC Enlisted Reclassification Test Condition was, in a classroom setting, using automation applications, AR 611-1, AR 140-158, DA PAM 611-21, and AR 611-6 review the DA Form 4187 and supporting documents and validate the request. Make any necessary corrections on the form itself. The PMC Enlisted Reclassification Test Standard was, validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

The PMC POI Action was, Validate Enlisted Classification/Reclassification Requirements. The PMC POI Condition was, given course automation equipment,

automation applications, Internet Explorer, Inter and Intranet domains, and course reference material. The PMC POI Standard was, validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

The PMC Course Test Plan required the students, using analysis, to use a hands-on performance test to demonstrate that the students could validate an Enlisted Reclassification Request. Based on this comparison, these instruments are in agreement.

Based on the comparison of the PMC Enlisted Reclassification Test, the PMC TLAW, the PMC POI, and the PMC Test Plan, the PMC Enlisted Reclassification Test was validated.

PMC Officer Reclassification Collaborative CRT. Following the Collaborative Content Validity Procedure (see Table 16), the results of the Officer Reclassification Collaborative CRT were as follows.

The first step was to compare the objective on the PMC Officer Reclassification Test objective (see Appendix A) to the PMC Training Learning Activity Worksheet (TLAW) (see Appendix B) objective. The PMC Officer Reclassification Test objective was to Validate Officer Request for Reclassification. The PMC TLAW objective was to Validate Officer Classification/Reclassification Requirements. Finally, the PMC Program of Instruction (POI) (see Appendix C) objective was to Validate Officer Classification/Reclassification Requirements. Based on the comparison of these instruments, all are in agreement on the objective.

The second step was to compare the PMC Officer Reclassification Test sample for each objective/domain to the PMC Course Test Plan (see Appendix D). The PMC

Officer Reclassification Test required the students to validate the request, making any necessary corrections on the form itself. The PMC Course Test Plan required the students, using analysis, to use a hands-on performance test to demonstrate that the students could validate an Officer Reclassification Request. Based on this comparison, these instruments are in agreement.

The third step was to compare each PMC Officer Reclassification Test item and its associated objective to the PMC Test specifications and/or PMC Course Test Plan, ensuring that the PMC Officer Reclassification Test item measures exactly what the objective says should be measured. The PMC Officer Reclassification Test Action was, Validate Officer request for reclassification. The PMC Officer Reclassification Test Condition was, in a classroom setting, using automation applications, AR 611-1, DA PAM 600-3, DA PAM 611-21 and AR 611-6 review the DA Form 4187 and supporting documents. Make any necessary corrections on the form itself. MAJ Anderson has orders assigning him to an 11A position at an Infantry Brigade. The PMC Officer Reclassification Test Standard was, validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

The PMC POI Action was, Validate Officer Classification/Reclassification Requirements. The PMC POI Condition was, given course automation equipment, automation applications, Internet Explorer, Inter and Intranet domains, and course reference material. The PMC POI Standard was, validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

The PMC Course Test Plan required the students, using analysis, to use a hands-on performance test to demonstrate that the students could validate an Officer Reclassification Request. Based on this comparison, these instruments are in agreement.

Based on the comparison of the PMC Officer Reclassification Test, the PMC TLAW, the PMC POI, and the PMC Test Plan, the PMC Officer Reclassification Test was validated.

Findings of the PMC Reclassification Collaborative CRT Reliability Process

The PMC Enlisted and Officer Reclassification collaborative CRTs were completed and submitted to the ISS to measure the reliability with the following results.

The PMC Enlisted Reclassification Collaborative CRT results were as follows.

- 2 “master groups” passed the test.
- 1 “master groups” failed the test.
- 0 “nonmaster groups” passed the test.
- 3 “nonmaster groups” failed the test.

The PMC Officer Reclassification Collaborative CRT results were as follows.

- 3 “master groups” passed the test.
- 0 “master groups” failed the test.
- 0 “nonmaster groups” passed the test.
- 3 “nonmaster groups” failed the test.

Using the Collaborative Phi Coefficient Procedure (see Table 17) the following results were found.

Table 17

Collaborative Phi Coefficient Procedure

STEP	PROCEDURE
1	Determine the number of "master groups" who passed the test.
2	Determine the number of "master groups" who failed the test.
3	Determine the number of "nonmaster groups" who passed the test.
4	Determine the number of "nonmaster groups" who failed the test.
5	Add number obtained in step 1 to the number obtained in step 2.
6	Add the number obtained in step 3 to the number obtained in step 4.
7	Add the number obtained in step 1 to the number obtained in step 3.
8	Add the number obtained in step 2 to the number obtained in step 4.
9	Multiply the number obtained in step 1 by the number obtained in step 4.
10	Multiply the number obtained in step 2 by the number obtained in step 3.
11	Subtract the number obtained in step 10 from the number obtained in step 9.
12	Multiply the numbers obtained in steps 5, 6, 7, and 8.
13	Take the square root of the number obtained in step 12.
14	Divide the number obtained in step 11 by the number obtained in step 13 to determine the Phi Coefficient rounded to nearest 2 decimals.

The Enlisted Collaborative Phi Coefficient Procedure Results-

$$\frac{(\text{Step 1})(\text{Step 4}) - (\text{Step 2})(\text{Step 3})}{\sqrt{(\text{Step 5})(\text{Step 6})(\text{Step 7})(\text{Step8})}} =$$

$$\frac{(2)(3) - (1)(0)}{\sqrt{(3)(3)(2)(4)}} =$$

$$\frac{6}{\sqrt{72}} =$$

$$\frac{2}{8.50} =$$

+.71 Enlisted Collaborative Phi Coefficient

The Officer Collaborative Phi Coefficient Procedure Results-

$$\frac{(\text{Step 1})(\text{Step 4}) - (\text{Step 2})(\text{Step 3})}{\sqrt{(\text{Step 5})(\text{Step 6})(\text{Step 7})(\text{Step8})}} =$$

$$\frac{(3)(3) - (0)(0)}{\sqrt{(3)(3)(3)(3)}} =$$

$$\frac{9}{\sqrt{81}} =$$

$$\frac{9}{9.00} =$$

+1.00 Officer Collaborative Phi Coefficient

Interpretation of the Collaborative Phi Coefficient Data and Action to be Taken. Using the TRADOC acceptable index standard of +.50 or above as the acceptable reliability index, the new collaborative CRTs are sufficiently reliable, with the Enlisted Reclassification Test at +.71 and the Officer Reclassification Test at +1.00. If the Collaborative Phi Coefficient Data were between –1.00 and +.50, this would have indicated that something maybe wrong with the test, test instructions, or reference materials. In this case, more detailed evaluation of the test would have to be conducted to determine whether it should be revised or discarded.

The Collaborative Phi Coefficient Procedure was originally designed for evaluating individual test reliability. When used to determine collaborative test reliability, each group collaborative test is scored. Each collaborative test is scored as a “Pass/Fail” (“Go/No Go”) on each administration of the test and the steps in the long method are used to determine a realistic reliability index.

Summary

This chapter gave the results of using the Collaborative Content Validity Procedure to validate collaborative CRTs and the Collaborative Phi Coefficient Procedure to evaluate the reliability of collaborative CRTs. The processes for evaluating the validity of the individual and the collaborative CRTs are the same. The process for evaluating the reliability of the individual and the collaborative CRTs are slightly different. To evaluate the reliability of an individual CRT, a minimum of six individuals, three "master" individuals and three "nonmaster" individuals would be needed. To evaluate the reliability of a collaborative CRT, a minimum total of 12 individuals, six individuals for the three "master groups" of two individuals each and six individuals for the three "nonmaster groups" of two individuals each would be needed. Using at least three master groups and three nonmaster groups is encouraged. Using fewer that three master groups and three nonmaster groups is not recommended. For example, if two master groups and two nonmaster groups were used to evaluate the reliability of a collaborative CRT, the worse case would be, if one master group failed the test, no nonmaster group could pass the test in order to meet the TRADOC minimum Phi Coefficient reliability index of +.50 or higher. This allows for very little error of the test with such a small number of groups.

These methods meet their own criteria for the validation and reliability processes for both individual and collaborative CRTs.

Chapter V

Summary, Conclusions, and Recommendations

Introduction

This chapter will summarize the previous chapters and report the findings of chapter four. Conclusions and recommendations will also be addressed.

Summary

Chapter one provided the Statement of the Problem, outlining the need for the Army Reserve Readiness Training Center (ARRTC) to have a process to determine the validity and measure the reliability of a collaborative Criterion-Referenced Test (CRT). Research indicated that there was no process to measure the validity or reliability of a collaborative CRT. The need was based on the ARRTC and Training and Doctrine Command (TRADOC) requirements to use a Systems Approach to Training (SAT) for the analysis, design, development, implementation, and evaluations phases of a course. Chapter one outlined the roles and responsibilities of those involved in the development of training products. Further, chapter one identified where the students come from, what the ARRTC mission and vision statements were, defined terms, and gave a reference list of publications that were used to create this study and support its findings. Finally, chapter one listed the study limitations.

Chapter two outlined how people are traditionally taught as individuals and then spend the majority of their lives working in small groups. Chapter two explains how some companies are now approaching training from the perspective of training people in small groups, just as they will work on the job. Chapter two defines what tests are, what

testing is designed to accomplish, and what is needed to master a task or subject. Also defined are Criterion-Referenced Tests (CRT), Norm-Referenced (NRT) Tests, Collaborative Testing, Social Learning Theory and how it applies to collaborative CRTs, how to interpret Criterion-Referenced Test Results, when Test Item Evaluation is conducted, and what is involved with Test Validity and Test Reliability. Finally, chapter two examined the five existing techniques or procedures used to validate an individual CRT and the six existing techniques or procedures used to evaluate the reliability of individual CRT that TRADOC has approved for use for adaptation to evaluate the validity and reliability of collaborative CRT at the ARRTC.

Chapter three described the research design, instrumentation, data collection, data processing, and study limitations of the Content-Related Evidence of Validity procedure to determine the validity and the Phi Coefficient Test Reliability Discriminator Index Long Method to evaluate the reliability of the Personnel Management Course (PMC) Enlisted and Officer Reclassification collaborative CRTs.

Chapter four reported the findings of the collaborative CRT validity and reliability processes of the Enlisted and Officer Reclassification collaborative CRTs and the interpretations of the results.

Conclusions

Both the Collaborative Content Validity to determine validity and the Collaborative Phi Coefficient Procedure to evaluate reliability have been proven to work for individual CRTs. The development of a collaborative CRT is possible, using the Collaborative Content Validity Procedure to determine the validity and the Collaborative

Phi Coefficient Procedure to evaluate the reliability as outlined in chapter three and proven effective in chapter four. These two procedures provide a standardize method of measuring the validity and reliability of both individual and collaborative CRTs. These procedures can be used for government and private sector training institutions whose mission is to train and test personnel collaboratively.

Recommendations

1. The ARRTC endorse the Collaborative CRT. Course teams who are teaching small groups can create a collaborative CRT that is valid and reliable.

2. Implementation of the Collaborative Content Validity Procedure to determine the validity and the Collaborative Phi Coefficient Procedure to evaluate the reliability of collaborative CRT as the minimum standards that collaborative CRT are measured against for application at the ARTTC.

3. Implement the Collaborative Content Validity Procedure to determine the validity and the Collaborative Phi Coefficient Procedure to evaluate the reliability of collaborative CRT processes for inclusion in the ARRTC SAT Regulation 351-1-1.

4. These applications be made available for use in a non-military educational environment.

Reference List

ARRTC Systems Approach to Training Regulation 351-1-1. (1999). United States Army Reserve Readiness Training Center, Fort McCoy, WI.

Bandura, A. Social Learning Theory. <http://tip.psychology.org/bandura.html>

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. Educational Psychologist, 28 (2), 117-148.

Dick, W. & Carey, L. The Systemic Design of Instruction (3rd Ed.). United States of America: HarperCollins Publishers.

Flavell, J. (1976). Metacognitive aspects of problem solving. <http://tip.psychology.org/meta.html>

Knowles, M. (1984). The Adult Learner: A Neglected Species (3rd Ed.). Houston, TX: Gulf Publishing.

Mager, R. (1975). Preparing Instructional Objectives (2nd Ed.). Belmont, CA: Lake Publishing Co.

Gronlund, N. & Linn, R. Measuring and Evaluation in Teaching (6th Ed.). New York, New York: Macmillan Publishing Company

Wang, C. (2000). Establishing a productive climate for adult learning in the small instructional group. Performance Improvement, 39, pp. 13-17.

Pray Muir, S. & Tracy, D. (1999). Collaborative essay testing. College Teaching. Winter 1999, 47, Issue 1, p33, 3p.

The Quantitative Skills for Trainers Reference Book (QSTRB). (1997). United States Army Logistics Management College, Fort Lee, VA.

Training and Doctrinal Command (TRADOC) Regulation 350-70. (1999). United States Army Training and Doctrinal Command, Fort Monroe, VA.

Training and Doctrinal Command (TRADOC) Systems Approach to Training Regulation Desk Reference. (1997). United States Army Training and Doctrinal Command, Fort Monroe, VA.

Appendix A

Tests

EXTRACT FROM PMC FINAL POI, ARRTC, FORT MCCOY. WI

INSTRUCTIONS. The Staff Administrative Assistant (SAA) for the 888th EN BN is absent. Our subordinate Company Commanders are seeking guidance on several classification/reclassification actions. The Battalion Commander has asked you, the S-1, to review and recommend solutions.

ACTION. Validate Enlisted request for reclassification.

CONDITION. In a classroom setting, using automation applications, AR 611-1, AR 140-158, DA PAM 611-21, and AR 611-6 review the DA Form 4187 and supporting documents and validate the request. Make any necessary corrections on the form itself.

STANDARD. Validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

Circle the appropriate copy designator

Copy 1 Copy 2 Copy 3 Copy 4

PERSONNEL ACTION

For use of this form, see AR 600-8-6 and DA PAM 600-8-21; the proponent agency is ODCSPER

DATA REQUIRED BY THE PRIVACY ACT OF 1974

AUTHORITY: Title 5, Section 3012; Title 10, USC, E.O. 9397.

PRINCIPAL PURPOSE: Used by soldier in accordance with DA PAM 600-8-21 when requesting a personnel action on his/her own behalf *(Section III)*.

ROUTINE USES: To initiate the processing of a personnel action being requested by the soldier.

DISCLOSURE: Voluntary. Failure to provide social security number may result in a delay or error in processing of the request for personnel action.

1. THRU *(Include ZIP Code)*	2. TO *(Include ZIP Code)*	3. FROM *(Include ZIP Code)*
Commander HQ, 888th Engineer Battalion 751 East 13th Avenue Fort McCoy, WI 54656-5137	Commander 88th Regional Support Command Bldg 56, Fort Snelling St. Paul, MN 55111	Commander HHC, 888th Engineer Battalion 751 East 13th Avenue Fort McCoy, WI 54656-5137

SECTION I - PERSONAL IDENTIFICATION

4. NAME *(Last, First, MI)*	5. GRADE OR RANK/PMOS/AOC	6. SOCIAL SECURITY NUMBER
Pie, Quentin T.	SGT/63Y20H8	979-64-5325

SECTION II - DUTY STATUS CHANGE *(AR 600-8-6)*

7. The above soldier's duty status is changed from _____ to

_____ effective _____ hours, _____ _____

SECTION III - REQUEST FOR PERSONNEL ACTION

8. I request the following action: *(Check as appropriate)*

	Service School *(Enl only)*		Special Forces Training/Assignment		Identification Card
	ROTC or Reserve Component Duty		On-the-Job Training *(Enl only)*		Identification Tags
	Volunteering For Oversea Service		Retesting in Army Personnel Tests		Separate Rations
	Ranger Training		Reassignment Married Army Couples		Leave - Excess/Advance/Outside CONUS
	Reassignment Extreme Family Problems	X	Reclassification		Change of Name/SSN/DOB
	Exchange Reassignment *(Enl only)*		Officer Candidate School		Other *(Specify)*
	Airborne Training		Asgmt of Pers with Exceptional Family Members		

9. SIGNATURE OF SOLDIER *(When required)*	10. DATE *(YYYYMMDD)*

SECTION IV - REMARKS *(Applies to Sections II, III, and V) (Continue on separate sheet)*

1. Request award of PMOS: 63Y2EH8JT and SMOS: 12B20P5JT.

_____ _____

2. SGT Pie meets the physical demands rating and PULHES IAW AR 611-201. SGT Pie has normal color vision.

3. The following documents are attached:

a. DA Form 1059 - Combat Engineer Course
b. DA Form 1059 - Master Fitness Trainer Course
c. Certificate of Training - Mountaineer Course
d. Current MOS Order
e. DA Form 330 - Language Proficiency Questionnaire
f. DA Forms 2-1 and 2A

SECTION V - CERTIFICATION/APPROVAL/DISAPPROVAL

11. I certify that the duty status change *(Section II)* or that the request for personnel action *(Section III)* contained herein -

☐ HAS BEEN VERIFIED ☒ RECOMMEND APPROVAL ☐ RECOMMEND DISAPPROVAL ☐ IS APPROVED ☐ IS DISAPPROVED

12. COMMANDER/AUTHORIZED REPRESENTATIVE	13. SIGNATURE	14. DATE *(YYYYMMDD)*
BIGGEY B. ELK, CPT, EN, USAR		

DA FORM 4187, JAN 2000 PREVIOUS EDITIONS ARE OBSOLETE USAPA V1.00

EXTRACT FROM PMC FINAL POI, ARRTC, FORT MCCOY. WI

INSTRUCTIONS. The Staff Administrative Assistant (SAA) for the 888th EN BN is absent. Our subordinate Company Commanders are seeking guidance on several classification/reclassification actions. The Battalion Commander has asked you, the S-1, to review and recommend solutions.

ACTION. Validate Officer request for reclassification.

CONDITION. In a classroom setting, using automation applications, AR 611-1, DA PAM 600-3, DA PAM 611-21 and AR 611-6 review the DA Form 4187 and supporting documents. Make any necessary corrections on the form itself. MAJ Anderson has orders assigning him to an 11A position at an Infantry Brigade.

STANDARD. Validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

Circle the appropriate copy designator

Copy 1 Copy 2 Copy 3 Copy 4

PERSONNEL ACTION

For use of this form, see AR 600-8-6 and DA PAM 600-8-21; the proponent agency is ODCSPER

DATA REQUIRED BY THE PRIVACY ACT OF 1974

AUTHORITY: Title 5, Section 3012; Title 10, USC, E.O. 9397.

PRINCIPAL PURPOSE: Used by soldier in accordance with DA PAM 600-8-21 when requesting a personnel action on his/her own behalf *(Section III)*.

ROUTINE USES: To initiate the processing of a personnel action being requested by the soldier.

DISCLOSURE: Voluntary. Failure to provide social security number may result in a delay or error in processing of the request for personnel action.

1. THRU *(Include ZIP Code)*	2. TO *(Include ZIP Code)*	3. FROM *(Include ZIP Code)*
Commander HQ, 888th Engineer Battalion 751 East 13th Avenue Fort McCoy, WI 54656-5137	Commander 88th Regional Support CommanI:\	Commander HHC, 888th Engineer Battalion 751 East 13th Avenue Fort McCoy, WI 54656-5137

SECTION I - PERSONAL IDENTIFICATION

4. NAME *(Last, First, MI)*	5. GRADE OR RANK/PMOS/AOC	6. SOCIAL SECURITY NUMBER
Anderson, Fernando L.	MAJ/39B00	979-98-2229

SECTION II - DUTY STATUS CHANGE *(AR 600-8-6)*

7. The above soldier's duty status is changed from _____ to

_____ effective _____ hours, _____ _____

SECTION III - REQUEST FOR PERSONNEL ACTION

8. I request the following action: *(Check as appropriate)*

	Service School *(Enl only)*		Special Forces Training/Assignment		Identification Card
	ROTC or Reserve Component Duty		On-the-Job Training *(Enl only)*		Identification Tags
	Volunteering For Oversea Service		Retesting in Army Personnel Tests		Separate Rations
	Ranger Training		Reassignment Married Army Couples		Leave - Excess/Advance/Outside CONUS
	Reassignment Extreme Family Problems	X	Reclassification		Change of Name/SSN/DOB
	Exchange Reassignment *(Enl only)*		Officer Candidate School		Other *(Specify)*
	Airborne Training		Asgmt of Pers with Exceptional Family Members		

9. SIGNATURE OF SOLDIER *(When required)*	10. DATE *(YYYYMMDD)*

SECTION IV - REMARKS *(Applies to Sections II, III, and V) (Continue on separate sheet)*

1. Request the following reclassification action:

   AOC: 11B FA: 48 SI: 5P LIC: GS or 11B485PGS

Using appropriate references, determine if the officer meets the criteria for award of the requested classification code. If correction(s) are necessary, record them above.

2. Included in the supplemental packet are the following forms of documentation:
   a. DA Form 1059 - Infantry Officer Basic Course
   b. DA Form 1059 - Infantry Officer Advance Course
   c. DA Form 1059 - Ranger Course Completion
   d. DA Form 1059 - Parachutist (Airborne) Course Completion
   e. DA Form 1059 - Psychological Operations Officer Course
   f. DA Form 330 - Language Proficiency Questionnaire - Proficient in German Bavarian
   g. CGSC Form 128 - 50 percent completion certificate - Command General Staff Officer Course
   h. M.A. in German from University of Delaware (abroad program)
   i. HQDA approved

SECTION V - CERTIFICATION/APPROVAL/DISAPPROVAL

11. I certify that the duty status change *(Section II)* or that the request for personnel action *(Section III)* contained herein -

☐ HAS BEEN VERIFIED ☒ RECOMMEND APPROVAL ☐ RECOMMEND DISAPPROVAL ☐ IS APPROVED ☐ IS DISAPPROVED

12. COMMANDER/AUTHORIZED REPRESENTATIVE	13. SIGNATURE	14. DATE *(YYYYMMDD)*
BIGGEY B. ELK, CPT, EN, USAR		

DA FORM 4187, JAN 2000 PREVIOUS EDITIONS ARE OBSOLETE USAPA V1.00

Appendix B

Training Learning Activity Worksheets (TLAWS)

EXTRACT FROM PMC FINAL POI, ARRTC, FORT MCCOY. WI

TASK/LEARNING ANALYSIS
WORKSHEET (T/LAW)

COURSE: Personnel Management Course (PMC) **POI#: 120 DATE: 22 Sep 1998** **TASK#: 130.50.02.01**

DUTY AREA: Manage EPMS **INSTRUCTIONAL SYSTEMS SPECIALIST/TRAINING CENTER CHIEF**: Woolsey/Due

TASK TITLE: Task # B1, Validate Enlisted Classification/Reclassification Requirements.

CRITICALITY: 2 JOB STANDARD: **Validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.**

Can the student identify Enlisted classification codes? Can the student determine Enlisted classification codes? Can the student identify processes for Enlisted reclassification? Can the student identify reasons for Enlisted reclassification?	RELATED KNOWLEDGE: Computer devices Computer components Internet Browser Navigation Skills	REFERENCES: Computer Manuals Operating Instructions Classroom notes Classroom scenarios AR 135-205 AR 611-201 AR 601-210 AR 140-158 AR 140-10 AR 135-178 AR 623-205 AR 140-111 AR 600-8-105 AR 635-200	MATERIALS/EQUIPMENT: Personnel Computer Monitor Keyboard Mouse User ID, Password Computer software (MS-Word, Form Flow, etc.)

Train as is: X **Modify:** **Not Trained:**	**JUSTIFICATION FOR MODIFICATION OR NOT TRAINED**:

EXTRACT FROM PMC FINAL POI, ARRTC, FORT MCCOY. WI

TASK/LEARNING ANALYSIS
WORKSHEET (T/LAW)

COURSE: Personnel Management Course (PMC) **POI#: 120 DATE: 22 Sep 1998** **TASK#: 130.50.02.02**

DUTY AREA: Manage OPMS **INSTRUCTIONAL SYSTEMS SPECIALIST/TRAINING CENTER CHIEF**: Woolsey/Due

TASK TITLE: Task # C5, Validate Officer Classification/reclassification Requirements.

CRITICALITY: 2 **JOB STANDARD**: **Validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.**

STEPS:	RELATED KNOWLEDGE:	REFERENCES:	MATERIALS/EQUIPMENT:
Can the student identify Officer classification codes? Can the student determine Officer classification codes? Can the student identify processes for Officer reclassification? Can the student identify reasons for Officer reclassification?	Computer devices Computer components Internet Browser Navigation Skills	Computer Manuals Operating Instructions Classroom notes Classroom scenarios AR 611-1 AR 611-6 DA PAM 611-21 DA PAM 600-3 DA PAM 600-11	Personnel Computer Monitor Keyboard Mouse User ID, Password Computer software (MS-Word, Form Flow, etc.)

Train as is: X **Modify:** **Not Trained:**	**JUSTIFICATION FOR MODIFICATION OR NOT TRAINED**:

Appendix C

Program of Instruction (POI)

EXTRACT FROM PMC FINAL POI, ARRTC, FORT MCCOY. WI

ARMY RESERVE READINESS TRAINING CENTER
POI SYSTEM
PERSONNEL MANAGEMENT COURSE
PMC – POI 921-130
3 March, 2001

SCOPE: This course provides students with premobilization, USAR-unique skills and knowledge required to perform USAR personnel management. The target audience for this course are Personnel Officers, Senior NCOs (SSG and above), or Staff Administration Assistants (SSAs) with personnel management responsibilities at battalion level and above. Personnel with little or no previous personnel experience should complete the Unit Administration Basic Course (UABC). Company/Detachment level personnel are not eligible for this course without a waiver.

STANDARDS: Instructors must meet the terminal learning objectives. Students must meet or exceed the academic standards established for this course.

LENGTH: 80 Academic Hours.

CLASS SIZE: OPTIMUM: 20
MINIMUM: 08
MAXIMUM: 20

NUMBER	SUBJECT
130	SPECIAL SUBJECTS
130.31	PERFORMANCE EXAMINATIONS

//**EXTRACT**//
///////

130.50.02.01 ACTION: Validate Enlisted Classification/Reclassification Requirements.

CONDITION: Given course automation equipment, automation applications, Internet Explorer, Inter and Intranet domains, and course reference material.

STANDARD: Validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

130.50.02.02 ACTION: Validate Officer Classification/Reclassification Requirements.

CONDITION: Given course automation equipment, automation applications, Internet Explorer, Inter and Intranet domains, and course reference material.

STANDARD: Validate a request for reclassification within 30 minutes and a minimum proficiency of 70 percent.

Appendix D

Course Test Plan

EXTRACT FROM PMC FINAL POI, ARRTC, FORT MCCOY. WI

PERSONNEL MANAGEMENT COURSE (PMC)
COURSE TESTING PLAN
POI 921-130
19 OCTOBER 1998

Testing Plan: The academic plan for this course is based on a 70 percent minimum mastery level for each individual's ability to all graded tests. Students are expected to work together in a collaborative effort using team consensus to complete the tests and share their expertise with each other in a cooperative learning environment.

Instrument	POI Number	Topics	Critical Percent	Highest L of L*	Additional Comments
AS-301a	130.50.02.01	Validate Enlisted Reclassification Request	70%	Analysis	Hands-on Performance
AS-301b	130.05.02.02	Validate Officer Reclassification Request	70%	Analysis	Hands-on Performance