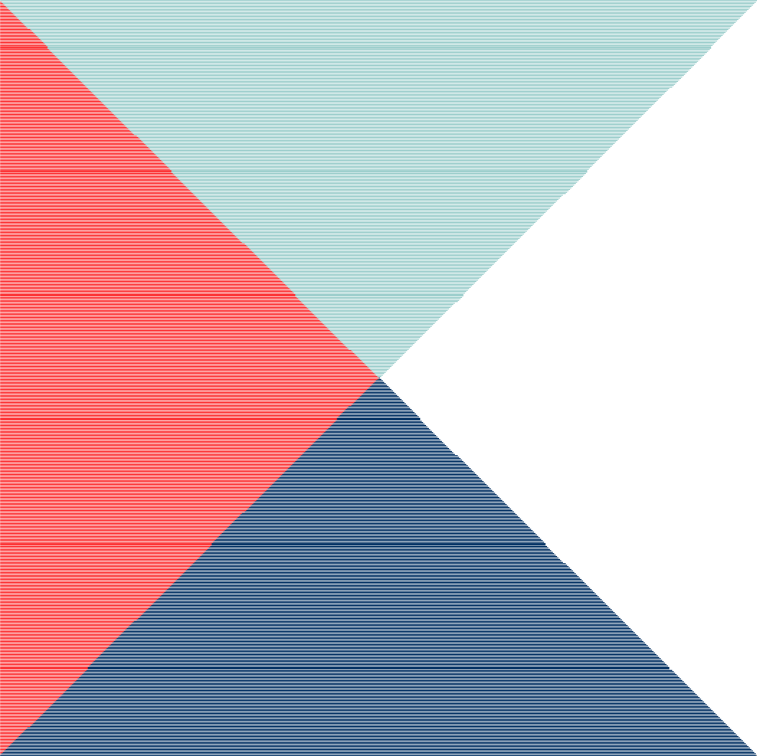
**UNCONTROLLED COPY**

**Defence Centre of Training Support Training Support Handbook**

**ASSESSMENTS IN TRAINING**

****

© CROWN COPYRIGHT 2007

**DOCUMENT QUALITY RECORD**

The following quality control statements are made in accordance with the requirements of the Defence Systems Approach to Training Quality System.

**Security Classification: UNCLASS**

**Authority:** This Training Support Handbook is published under the authority of the Commanding Officer Defence Centre of Training Support.

**Review Period:** This manual will be reviewed annually.

**Issue Number:** V1.0

**Amendment status:** See page ii.

**Disposal Instructions:**

1. **Amendments:** Individual pages that are replaced during the amendment process should be disposed of according to the document security classification in accordance with local instructions.
2. **Re-Issue:** On occasion the entire handbook will be re-issued and under those circumstances the old version of the handbook must be destroyed, without undue delay, upon receipt of the later version.

i

UNCONTROLLED COPY

**AMENDMENT RECORD**

**Change**

**Number**

**Description**

**Name Signature**

**Date**

**Incorporated**

ii

UNCONTROLLED COPY

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| Document Quality Record | | i |
| Amendment Record | | ii |
| Contents |  | iii |
| Foreword | | iv |
| Overview of DSAT Process | | v |
| Section 1 – Assessment in Training | | 1-1 |
| Section 2 – Planning Assessments | | 2-1 |
| Section 3 | – Developing a Practical Assessment | 2-9 |
| Section 4 | – Developing a Theory Assessment | 2-21 |
| Section 5 | – Evaluating Assessment | 2-38 |

This document is a Crown Copyright and is being made available for information purposes only. Any copying or commercial use of the document must be by way of a license for which an application can be made to:

IPR-PL1,

Directorate of Intellectual Property Rights,

Poplar 2a #2218,

MoD Abbey Wood

Bristol

BS34 8JH

iii

UNCONTROLLED COPY

**FOREWORD**

1. For the purpose of this document, the term “training”1 encompasses all Learning2, Education3 and Personal Development4 that has the objective of developing the knowledge, skills and/or attitudes of an individual towards preparing that individual for their role.
2. **Policy for the Management of Individual Training and Education in Defence.** This policy sets the framework for the management of individual training across Defence and details the key areas of Training Management (TM), Training Requirements Authority5 (TRA) and Training Delivery Authority6 (TDA) roles and Customer Executive Board (CEB) function. It is the high level policy that defines who is responsible for what in Defence Individual Training.
3. **Defence Systems Approach to Training Quality Standard (DSAT QS).** DSAT QS sets out the strategic principles to be applied to all Individual Training provided by, or on behalf of, Defence. The DSAT QS has been endorsed by the Training and Education Policy Group (TEPG) as the quality standard for the management of Individual Training across Defence. Any activity that has the objective of developing the knowledge, skills and/or attitudes of an individual for their current or future role must comply with DSAT QS.
4. **Defence Training Support Manuals (DTSMs).** The DSAT QS is underpinned and supported by DTSMs that direct its implementation. There are 6 DTSMs:

DTSM 1 - The Analysis, Design and Development of Training.

DTSM 2 - The Glossary of Defence Training Terminology.

DTSM 3 - Training Needs Analysis.

DTSM 4 - The Evaluation of Training.

DTSM 5 - Technology Based Training Solutions.

DTSM 6 - The Audit and Inspection of Individual Training.

1. An activity that aims to impart the specific knowledge, skills and/or inculcate appropriate attitudes required by an individual in order to perform adequately a task or job.
2. Learning is the acquisition of knowledge, skills and/or attitude.
3. Education encompasses the development of intellectual capacity, the acquisition of general supporting knowledge and inculcation of attitudes, which underpin performance, and engender understanding, commitment and ethos.
4. Personal Development is the enhancement of personal and/or professional attributes arising from a combination of training, education and experience.
5. The TRA represents the end-user of the trained output. It is the ultimate authority for the derivation and maintenance of the Operational Performance Statement (OPS) or the Learning Objectives/elements of the appropriate Competence Framework (CF), and is responsible for the evaluation of the effect of the training and education in achieving that OPS/CF (delivered both in the training school/organisation and the workplace).
6. The TDA is the organisation responsible for the provision of individual training or education, to agreed standards and in accordance with extant and funded Defence and single Service policies, on behalf of the customer(s). It is the conduit through which a Training Organisation/School is commanded/headed, resourced and administered.

iv

UNCONTROLLED COPY

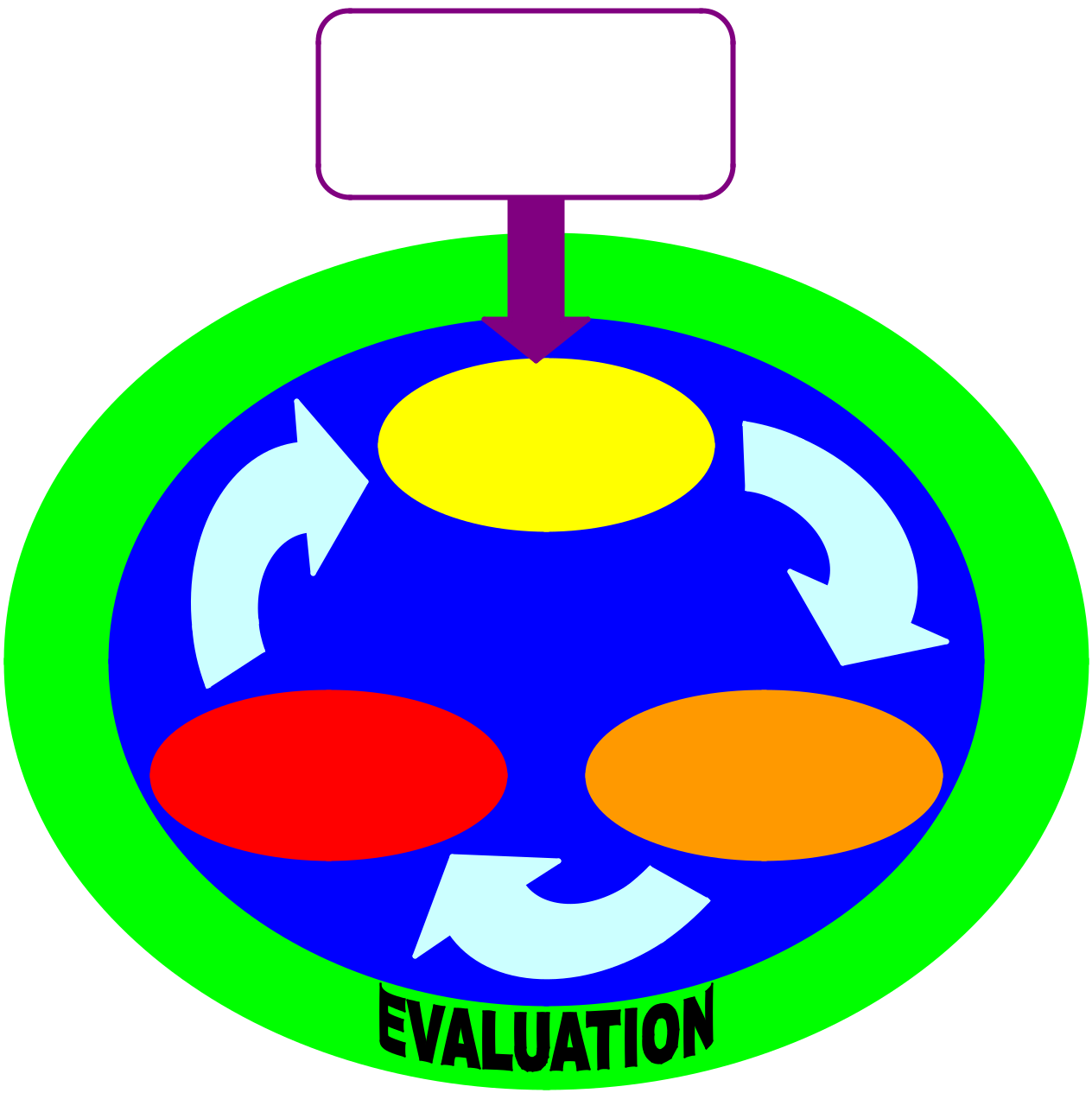
1. **Defence Centre of Training Support (DCTS) Training Support Handbooks.** In order to further amplify the information contained in DTSMs and in direct support of the training delivered to Training Support specialists, DCTS has developed a series of DCTS Training Support Handbooks. These publications are provided as reference guides for Training Support practitioners to give additional detailed guidance on specific areas relating to the training delivered by DCTS. DCTS Training Support Handbooks are available on the following topics:
   * Job Analysis and Formal Training Statements
   * Instructional Design incl Course Programming
   * Assessment
   * InVal and ExVal
   * Training Needs Analysis (TNA)
   * Data Gathering and Analysis Techniques
   * Audit

v

UNCONTROLLED COPY

**OVERVIEW OF THE DSAT PROCESS**

1. The DSAT process illustrated in Figure 1 demonstrates how the main process groups are needed for the comprehensive Analysis, Design and Development exercise. These groups are:
   1. Needs Analysis.
   2. Training Design and Development.
   3. Training Delivery.
   4. Evaluation (DTSM 4 deals with the Evaluation of Training).



Change in, or review of,

operations/business triggers a

perceived need for Training

**NEEDS**

**ANALYSIS**

**TRAINING**

**TRAINING** **DESIGN &**

**DELIVERY** **DEVELOPMENT**

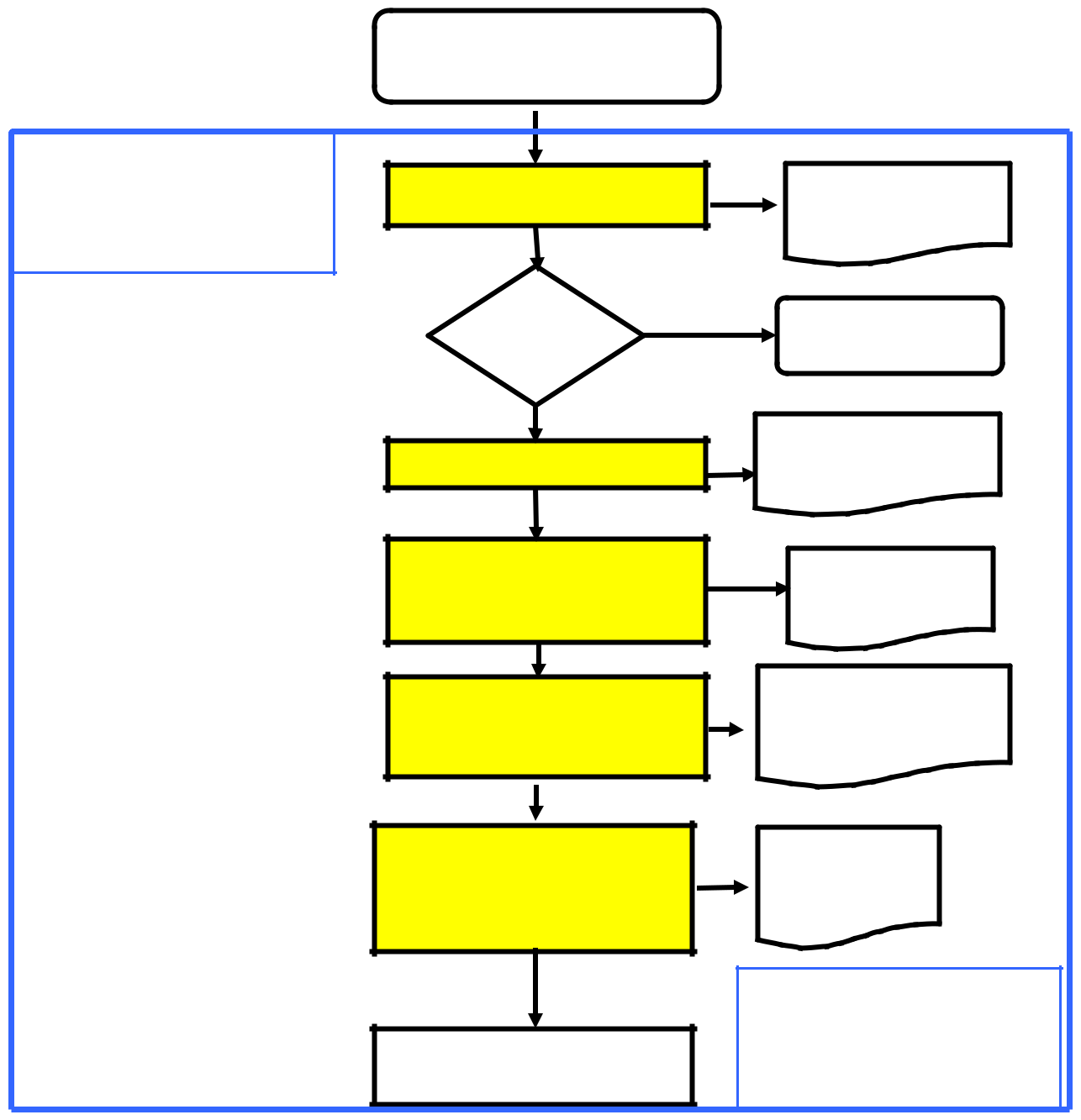
Figure 1 – DSAT Process

vi

UNCONTROLLED COPY

1. This handbook is designed primarily to accompany and to reinforce the training delivered within the Job Analysis course. Defence Training Support Manual
2. (DTSM1) – The Analysis, Design and Development of Training sets out the various processes and tools that may be used in order to meet the requirements of Defence Systems Approach to Training (DSAT QS) regarding activities required for the analysis, design and development of Defence individual training.
3. Figure 2 provides an illustrative process diagram of the analysis, design and development of

training and shows how the production of the key DSAT documentation is linked to the 3 stages of Training Design.



**EVALUATION**

Applied to all stages of the DSAT process as appropriate

A change in, or review of, operational/business practices triggers a perceived requirement for

**SCOPING EXERCISE**

|  |  |
| --- | --- |
| Is a | **NO** |
| training |  |
| interventio |  |

**Scoping Report**

Stop DSAT process.

**YES**

**NEEDS ANALYSIS**

**TRAINING DESIGN &**

**DEVELOPMENT – STAGE 1**

(Determination of Training

Objectives)

**TRAINING DESIGN &**

**DEVELOPMENT – STAGE 2**

**TRAINING DESIGN &**

**DEVELOPMENT – STAGE 3**

(Production of Training and

Assessment Media)

**TRAINING DELIVERY**

**Operational Performance Statement/ Competence Framework**

**Formal training**

**Statement**

**Assessment Strategy**

**(incorporating**

**Assessment**

**Specification)**

**Instructional**

**Specification**

**EVALUATION**

Applied to all stages of the DSAT process as appropriate

Figure 2 – DSAT Illustrative Process Diagram

vii

UNCONTROLLED COPY

**SECTION 1 - ASSESSMENTS IN TRAINING**

**BACKGROUND**

1.1 Testing is an essential aspect of military training, yet many aspects of testing are often misunderstood and misapplied. The development of the Defence Systems Approach to Training has provided a clear focus for the use of tests in support of military training.

**AIM**

1.2 The aim of the handbook is to guide the use of tests in military training and it will be used as the main course material for the Assessment in Training module. Its contents may also be supplemented by DSAT publications as appropriate.

**SCOPE**

1.3 Although testing as a field of study has great breadth and depth, not all elements of this activity are relevant to military training. Because training is concerned with preparing servicemen/women to do their jobs, the focus of this handbook is on tests of ***achievement*** – measures of the extent to which learning has occurred.

1.4 Tests used for purposes of recruiting and selection, such as measures of intelligence or aptitude, are not discussed in this handbook. Use of such tests requires special training and experience, normally limited to personnel employed by the recruiting organisations.

**APPROACH**

1.5 This handbook is designed to serve as a job aid for those who design, administer, score and evaluate achievement tests. Accordingly, the sequence of the handbook follows the normal sequence of the work associated with test development (see Figure 1).

1.6 Section 2 – Planning Assessment, places test planning within the context of military training and provides guidance for the strategic direction of testing activities. This includes discussion of the fundamental issues of test purposes and responsibilities for testing. Note: the development of an assessment strategy and test specification are covered on the Instructional Design course.

1.7 Section 3 – Constructing Practical Tests, provides detailed guidance for the design and use of those tests, which measure soldiers’ ability to ***do*** things. Because these tests usually provide a direct measure of job ability, they should predominate in military training.

1

UNCONTROLLED COPY

1.8 Section 4 – Constructing Theory Tests, provides detailed guidance for the design and use of those tests which measure students’ ***knowledge*** of facts, concepts and principles, In training, these tests are best used to complement practical tests.

1.9 Section 5 – Evaluating Tests, examines the quality control methods essential to confident interpretation of test results. This includes checking that tests are consistent in measuring student achievement and that they are in fact measuring what they are supposed to measure.

1.10 Section 6 – Glossary of Testing Terms, provides explanations of the common terms using in testing

**The Test Development Process**

**Stage One – Plan Assessment** **SECTION 2**

**Stage Two – Construct Practical Tests** **SECTION 3**

* Determine test content and standard.
* Select appropriate format.
* Construct the test.
* Prepare instructions for students, administrators and scorers.
* Pilot tests

**Stage Three – Construct Theory Tests** **SECTION 4**

* Determine test content and standard.
* Select appropriate format.
* Construct the test
* Prepare instructions for students, administrators, and scorers
* Pilot the test.

**Stage Four – Evaluate Tests** **SECTION 5**

* Evaluate the assessment strategy/methodology.
* Evaluate practical tests for validity.
* Evaluate theory tests for validity.
* Interpret test results.

**Figure 1: The Test Development Process**

2

UNCONTROLLED COPY

**SECTION 2 - PLANNING ASSESSMENTS**

**GENERAL**

2.1 Results of tests are used to make critical decisions regarding soldiers’ progress through training and subsequent qualification for employment. Such decisions may have direct implications for economy or resources, preservation of life and success in battle.

2.2 To ensure that such decisions are sound the use of tests must be based upon rational and methodical planning. In this section, guidelines for planning are presented. This reinforces the fundamental principles covered in the Distance Learning Package – Assessment Principles.

**FUNDAMENTAL TERMS AND CONCEPTS**

2.3 **Testing Defined.** A test is an event during which a student is asked to demonstrate some aspect of knowledge or skill. Tests which measure the extent to which such knowledge or skill has been learned (for example, as a result of training) are termed achievement tests.

2.4 **Measurement and Assessment. *Measurement*** is the determination of the extent of some characteristic, often expressed in numerical terms. ***Assessment*** occurs when a measurement is compared with a standard, and a judgement is made which is based on that comparison. In achievement testing we seek to measure student learning, and to use the measures taken as a basis for assessment. For example, a soldier who dons a respirator in less than 9 seconds (measurement) is deemed to have passed that element of the NBC defence test (assessment).

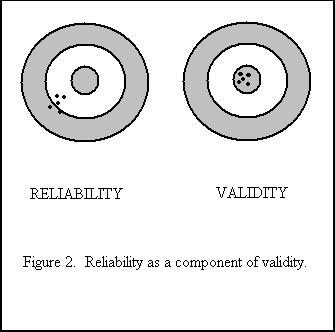
2.5 **Test Validity.** Because assessment is based on measurements collected through testing, continued accuracy of measurement ***Validity*** is essential. A ***Valid Test*** is one which measures what it is supposed to measure. Test validity has five components:

a*.* ***Relevance or Content Validity****.* A relevant test will measure only learning which is pertinent to the purpose of the test. Irrelevant student performances must not affect the scoring (e.g. Awarding marks for writing style, when knowledge of facts is the only requirement), and the test itself must not direct students to carry out irrelevant, or seemingly irrelevant, performances. This last point is known as ***face or content validity***– that a test appears to measure what it is supposed to measure.

1. ***Reliability.*** Measurement should be based on consistent and objective observation. A reliable test would, in theory, produce similar scores when, (1) the test is repeated by the same student, and (2) the same student performance is measured by different scorers. It is possible for a test to be reliable without being valid, but an acceptable test must be valid and reliable (see Figure 2).

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 1 |

UNCONTROLLED COPY



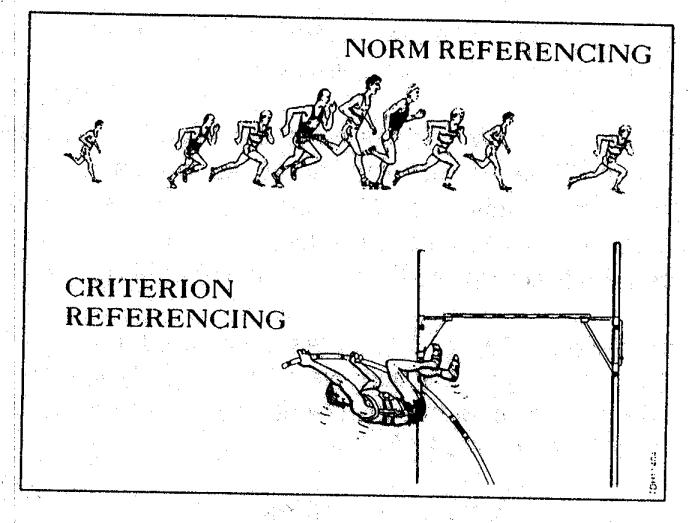
1. ***Practicable or useable.*** The degree to which a test is suitable for use i.e. Is it?:
   * + Administratively feasible
     + Cost effective
     + Easily interpreted by all users
     + Feasible in terms of time
     + Marked using a suitable and practicable scheme
2. ***Equity.* Equity** is the requirement for assessment to be gender and race fair and also the requirement for consideration of learning difficulties / special needs such as dyslexia or dyspraxia.
3. ***Integrity***. There is a requirement to maintain the **integrity** of all test items, assessments, marking guides and associated instructions. Methods of maintaining test security include:
   * restricted access to test papers and marking guides / DS solutions
   * random construction of equivalent test items from a question bank
   * validators to set and mark assessments (rather than instructors)
   * use of an exam board for standardisation and moderation and to investigate any unusual results or potential collaboration between students

2.6 **Choice of Assessment Standards.** The standard against which a measure can be applied will take one of two forms:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 2 |

UNCONTROLLED COPY

1. ***Norm-referenced***. The performance is compared to that of other students. Highest and lowest score, class average, and grades awarded “on a curve” are examples of norm-referenced expressions. While norm-referencing may have some advantages for purposes of selecting among students (assuming both that there is a need to do so, and that the test has been proven to be valid), it cannot tell us whether students have met the standard of required achievement.
2. ***Criterion-referenced***. The performance is compared to a fixed standard. By comparing measured performance against this standard. Each student may be assessed as having passed or failed. Criterion-referencing lends itself well to military training, because fixed standards are incorporated into Training and Enabling Objectives (TOs and EOs). Test standards can be established to reflect mastery of an objective, so that students who pass demonstrate achievement of that objective, a readiness to proceed to further training or employment.



**Figure 3. Norm vs Criterion-Referencing**

2.7 **Approaches to Grading.** A ***grade*** is any label, which represents an assessment. Since assessment is based on norm- or criterion-referencing, grades should clearly reflect the approach taken. Various labels may be applied. For example, letter grading (A, B, C,..) normally would be taken to represent norm-referenced assessment, while Pass/Fail would imply criterion-referencing. There is a risk of confusion and misunderstanding where these conventions are ignored.

2.8 **Test Formats.** There are two main ways in which tests can be presented.

1. ***Practical Tests***. Practical tests require the student to perform job skills. They can require physical activity (e.g. equipment operation), written work (e.g. service writing) or oral communication (e.g. verbal orders). Performance is usually evaluated against a checklist or rating scale. Practical tests are recommended for use in

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 3 |

UNCONTROLLED COPY

military training because they tend to be job oriented, inherently valid, and relatively easy to maintain.

1. ***Theory Tests***. Since practical tests can demand much time and many resources, supplementary theory tests are sometimes required. Theory tests measure the knowledge, which supports job skills, usually by taking a sample of what must and should be known. These tests are usually in written form (e.g. multiple choice, short answer) although oral tests can also be used. To achieve validity, theory tests require much care in construction and/or scoring, and usually require considerable labour devoted to bookkeeping and statistical analysis.

2.9 **Formative and Summative Testing.** Achievement tests used in support of training can serve two purposes;

1. ***Formative Tests***. These tests are used to help both the student and the instructor to recognise lapses in learning, so that corrective action (such as additional practice or remedial instruction) can be initiated. They also serve to reinforce successful learning, so that it will be retained longer.
2. ***Summative Tests***. These tests are used to determine whether students have achieved Training Objectives, or those significant sub-objectives (Enabling Objectives) which are deemed prerequisite to further training.

**RESPONSIBILITIES FOR TESTING**

2.10 **The Ideal.** The formative and summative purposes of testing are mutually supportive, but are best achieved when distinct testing responsibilities are assigned (see Figure 4). Within a training establishment, these responsibilities should be allocated as follows:

1. Validation (sometimes called Standards) staff provide the School Comdt/CO with quality control – verification that standards are being met. Ideally, this staff would carry out all design, administration, and scoring of summative tests. This would include ***terminal tests*** of Training Objective achievement, ***enabling tests*** of Enabling Objective achievement, and any tests designed to confirm student achievement of course prerequisites - ***entry tests.***
2. Instructors concentrate on helping students, by means of formative testing. This includes use of various in-class tests, as well as any placement tests designed to help match instructional approaches to student needs.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 4 |

UNCONTROLLED COPY

**Testing Student Achievement**

**In Military Training**

**O?WHO?**

Instructors

Validators

**WHY?**

**HOW?**

To match

Student

Placement

Test (note 2)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | To Aid | | |  |  | To Verify that Students | | | |
|  |  |  |  | Learning | | |  |  | Meet Start Standards | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Oral | |  |  | End of | |  | Revision Tests | |  | Entry Test | |  |
| Questions to | |  |  | Lesson | |  | (Quizzes) | |  | (Note 3) | |  |
| Class | |  |  | Tests | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

To Verify Critical Student Progress

Progress/ Enabling Test (Practical/Theory)

To Verify Student Attainment of Objectives

Terminal Test

(Practical)

Notes: 1.These **Formative** tests aid learning by improving retention where learning has occurred, and by cueing remedial action where it has not.

2.**Placement** tests establish the appropriate level for initial instruction, and imply some flexibility in course design and instructional approach. Avoid pass/fail assessment here – students are not expected to demonstrate mastery at course start!

3**.Entry** tests check for attainment of prerequisite knowledge or skills. Pass/fail assessment is appropriate, and failure could result in RTU.

**Figure 4: Testing Student Achievement**

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 5 |

UNCONTROLLED COPY

2.11 **The Reality.** Few training establishments have sufficient validation staff to carry out all the summative testing activities. Instructors are often called upon to administer and score validators’ tests. The best compromise is to share the labour, but to clearly maintain the division of responsibility. To preserve the quality control function, validators must retain control of summative test construction and content; they should administer, monitor, and score as much summative testing as possible. Similarly, instructors must ensure that formative tests are used only for formative purposes; these scores should not be added, averaged, or otherwise factored into the summative assessment process.

**THE COURSE ASSESSMENT STRATEGY**

2.12 **Definition and Rationale.** Every course should have a document that clearly lays down the strategy for the use of tests in support of a training program. A carefully conceived assessment strategy will achieve the following:

1. It provides an overview of the sorts of tests to be used, the points during the training when they will occur, and how the results of tests are to be interpreted and acted upon.
2. It is both a, record of decisions taken about the best approach to assessment, and
3. guide for the later development of individual tests.
4. It helps ensure that tests match the requirements of the Training Objectives, and that adequate resource for testing are identified and obtained.
5. It will, quite rightly, influence the manner in which instruction is delivered (e.g. a gunnery instructor knows that the summative test will emphasise practical handling skills, and wisely ensures ample student practice during instructional periods).

2.13 **Elements of the Assessment Strategy.** As a minimum, the assessment strategy must include clear direction for each of the following:

1. A concept for the summative testing of each of the Training Objectives. This should be based upon a practical terminal test, which may be supported by selected enabling tests in either practical or theory format. At this point, a short description of the test is sufficient; e.g. “A practical test in which each student will command a tank during a troop advance” or “a theory test on the Highway Code”.
2. A concept for the formative testing of student progress. This might include a statement of purpose, an assignment of responsibility, and a caution about the use of formative test results (see para 11 above)
3. A policy for the assignment and interpretation of grades (see para 7 above).
4. A policy, for the action to be taken upon student failure of a (valid) test. Because appropriate action will depend upon many variables, it is recommended that this

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 6 |

UNCONTROLLED COPY

policy be flexible rather than restrictive. For example, a statement guaranteeing (or denying) a re-sit will seldom prove practicable; it would be better to establish a procedure (e.g. a Student Review Board) by which each case will be considered against criteria such as:

* 1. Resources required, to repeat the test, without compromise of test conditions and assessment standard.
  2. Requirements for, additional tuition and/or practice.
  3. Likelihood of student success during the re-sit.

1. A policy for determining course pass or failure. This can be a statement such as, “to pass this course, students must achieve all Training Objectives”, or “pass all summative tests”. The inclusion of such a simple statement provides focus to the testing; it can also prevent misunderstanding or even serious grievance later on.
2. A policy for the maintenance of test records. This should state a clear requirement for both of the following:
   1. A record for each student that includes a summary of all test results (both formative and summative), as well as a record of formative action taken, such as counselling notes or copies of written warnings. This record should be used to guide the preparation of the student course report.
   2. A consolidated tabular record of summative test results. This record, accumulated over several repetitions of a course, provides valuable information for internal validation of training in general, and evaluation of tests in particular (see Section 5 for details).

2.14 **Publication.** The assessment strategy is an essential part of the School Comdt/CO’s plan for the achievement of Training Objectives, and provides important policy guidance. Under DSAT the overall assessment strategy for a course will form an important section of the Course /Module Specification (CSpec), and then the Assessment Specifications (ASpec) will give more details on the individual assessments to be used. These documents are dealt with in more detail on the Defence Instructional Design (DID) course. Examples of the format and content of these documents can be found at Annex A and B.

**SECTION SUMMARY**

2.15 Military achievement tests enable trainers to access student performance in comparison to either a norm- or criterion-reference. In training (as opposed to selection), criterion-referencing is more appropriate, and lends itself to practical tests which act as rehearsals for the job (Figure 5). Similarly, this job orientation points to Pass or Fail assessment rather than assignment of letter grades.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 7 |

UNCONTROLLED COPY

**Test Standards, Formats and Purposes**

**Norm-** **Criterion-**

**referenced** **referenced**

Theory

Format

Practical

Format

Purpose: Formative

Placement tests

Instructors’ tests of Learning in facts lessons

Purpose: Selection

NOT appropriate for testing training achievement

Purpose:

Summative

Validators’, tests of

achievement in

theory-based entry

tests/EOS.

Purpose: Formative

Instructors’ tests of

learning in skills

lessons

Purpose:

Summative

Validators’ tests of

achievement in skill-

based entry tests? EOs,

and ALL TOs.

**Figure 5: Test Standards, Formats and Purposes**

2.16 Instructors and Validators have distinct responsibilities and purposes for testing, and have many testing tools available. Tests can occur at the start, during and at the end of training. Placement, Entry and Enabling Tests may not always be necessary, but as a minimum, sound training assessment will rely upon a combination of formative classroom tests and summative terminal tests.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 8 |

UNCONTROLLED COPY

**SECTION 3 – DEVELOP A PRACTICAL ASSESSMENT**

**THE ROLE OF PRACTICAL TESTS**

3.1 When a testing situation matches the requirements of a job, students must demonstrate ability to do the job in order to pass the test. This is the aim of practical testing – to test the application of learned skills in realistic situations.

3.2 In the military, Training objectives or Training Performance Statements provide a detailed description of the training requirements for the job. Training Objectives are inherently practical – they describe what the student must do. A practical test that matches a Training Objective does not merely predict future job ability, it demonstrates job ability already achieved.

3.3 Most military achievement testing should be practical in format. Practical tests should be used for:

1. Summative testing of every TO
2. Summative testing of skill oriented Enabling Objectives (EOs), when indicated in the assessment plan/strategy.
3. Formative testing of skills lessons.

3.4 Limitations of time, staff, equipment, and other resources often constrict the degree of realism that can be achieved in practical testing. To ensure test validity, test developers must strive to provide as much realism as possible, where necessary achieving economies by the use of simulation and sampling techniques. The substitution of theory tests for practical tests is not recommended, unless unavoidable.

3.5 Similar to the design of theory tests, there are 5 steps in the construction of a practical test.

**STEP 1 – DETERMINE CONTENT AND STANDARD**

3.6 **Examine the Objective**. The first step in the construction of a practical test is to examine the relevant Training Objective, Enabling Objective, or Instructional Specifications. The developer must pick out the relevant details needed to achieve the best possible match between the objective and the test.

3.7 **Determine test performances** . The performance statement describes the end performance or final outcome, but the tester must consider the intermediate step by which that outcome is achieved. In particular, the tests must decide the extent to which successful performance relies upon each of the following:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2 - 9 |

UNCONTROLLED COPY

1. **Process**. Is the performance based on the execution of a set procedure; eg, delivery of orders in a standard format, or preparing a weapon for firing? If this is the case, the designer should list all those procedural steps that ***must*** be carried out.
2. **Product**. Is there a finished item by which the performance can be measured; e.g. a prepared meal, or a written estimate. If so, the designer should list all those measures that are ***essential*** to make the final product acceptable.
3. **Combination**. Acceptable performance will often require elements of both

process and product; e.g. following a trouble-shooting sequence to locate a radio fault, then repairing the fault to produce a functional radio. When such a combination is indicated, the designer should prepare separate lists of the required steps and measures.

3.8 **Determine test conditions**. The objective describes the conditions under which the performances must occur. Each of the aspects ‘given’ or ‘denied’ in the objective (tools, cues for action, supervision, etc) must be recreated or simulated in the test situation. Where alternative conditions are stated, the more restrictive or demanding one should apply; e.g. if the conditions state “by day or by night”, the test should usually occur at night.

3.9 **Determine the test standard**. By definition, standards stated in objectives should be expressed in observable and measurable terms. This will generally aid the development of test standards, but the test designer must be prepared to translate single, somewhat broad statements, into more specific details. For example, the requirement to type a service letter “ in accordance with JSP 101” might be developed into more specific statements dealing with margins, indentation, abbreviations, and so on. Similarly, the designer should be alert for statements, which refer to attitudes that students must display (e.g. safety consciousness, aggressiveness, confidence). To incorporate such attitudes into a test standard, the designer should describe indicators – student behaviours that are assumed to reflect the presence or absence of the desired attitudes.

**STEP 2 – SELECT APPROPRIATE FORMAT**

3.10 **Decide upon sampling strategy.** In some cases, resource limitations may preclude complete testing of a practical skill. When this occurs, the designer must decide the best manner in which to sample the required behaviour, while preserving confidence in the validity of the test. To minimise reduction in validity, the sample taken must be as comprehensive as possible – at least some of the elements of each of the performances identified at para 7 above should be required by the test. Where many similar activities share an underlying skill, placing the student in a few random but typical situations might satisfactorily test that skill. In some cases, those elements, which must be omitted in a practical test, can form the basis for a supporting theory test. For example:

1. During a practical driving test, it would not be feasible to present enough situations to test all applications of the rules of the road. A theory test might be used

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-10 |

UNCONTROLLED COPY

to sample a wide range of knowledge, and a few random practical situations could check the student’s ability to apply that knowledge.

1. Emergency servicing of a complicated piece of equipment might entail checks involving hazards or complexities not reproducible in training. During an otherwise practical test, the tester might pose oral theory questions of the “what if ” variety, calling upon the student to explain the processes which cannot be acted out.
2. A repair procedure may require the location and correction of a large number of possible faults. A practical test might randomly present a few of these, and a theory test could be used to sample knowledge of the variation involved in the other situations.

3.11 **Select test format**. For each of the performances selected for the testing, the designer should consider the standard required, and selects the most appropriate test item format. Practical test items fall into three general categories:

1. **Checklist items**. If a performance can be rated either as satisfactory or unsatisfactory, or as having occurred or not occurred, a checklist item should be used.
   1. They offer high reliability, because test scores are less likely to apply subjective judgements, and all scores will be looking for the same performances and standards.
   2. They are generally easy for the scorers to complete, and lend themselves readily to criterion- referenced assessment for the overall test.
   3. The score sheet itself can serve as a precise guide for instruction, and student rehearsal for the test.
2. **Rating scale items.** In some cases, a simple yes or no measure may not allow an adequate assessment. There may be a requirement for additional gradations, e.g., unsatisfactory/satisfactory/very good. A rating scale item is often better expressed as several checklist items, each of which corresponds to one of the gradations on the scale. In formative testing, where the student can most benefit from guidance for improvement, rating scale items may have merit. They should otherwise be avoided, unless there is a genuine need for norm-referencing, because of the following disadvantages:
   1. Rating scale items threaten the reliability of the test, and the more gradations on the scale, the greater the threat. Different scorers will likely differ in how they interpret the difference, for example, between “satisfactory” and “good”. Some scorers favour certain parts of the scale, tending to rate high, low, or down the middle for all students. Others may allow their general impression of a student to influence their rating of that student’s behaviours.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-11 |

UNCONTROLLED COPY

Finally, scorers may incorrectly link ratings for different items together, based on preconceived but false notions.

* 1. It is difficult to determine the pass standard for a rating scale item, and for the test overall. If “satisfactory” is good enough, why use additional gradations?

1. **Anecdotal record items.** Some performances, such as attitudes, are difficult to measure with the precision implied by checklist and rating scale items, and may only be evident in exceptional circumstances. If it can be assumed that the described behaviour is a valid ***indicator*** of an attribute that is otherwise difficult to measure, occasions that reflect the presence or absence of such learning may be recorded as descriptive anecdotes, providing a record that may be used for formative counselling. In the interest of validity, such records must include precise examples of the observed behaviour.

3.12 At the end of this step, the designer should have a clear plan for the overall structure of the test, including the appropriate mix of checklist, rating scale, and anecdotal record items.

**STEP 3 – CONSTRUCT THE TEST**

3.13 **Construct checklist items.** Designers should adhere to the following guidelines:

1. Each item should refer to a single, observable performance. E.g., “clean & oil bolt case” should be reworded as two items “clean bolt case” and “oil bolt case”.
2. Describe in positive terms what the student must do to achieve a satisfactory rating. This clarifies what the scorer must look for. Examples are as follows:
   1. Not “Doesn’t break the rules of safety”, but “Avoids safety infringements”.
   2. Not “No more than 3 typographical errors” but “Maximum of three typographical errors”.
3. Structure all items consistently. If, for example, the “satisfactory” of “yes” rating is always the first option, scorers are unlikely to tick the wrong box by error.
4. Keep each item descriptor short. The scorer should be able to see the required performance and standard at a glance. Avoid unnecessary words such as “correct” or “appropriate”. Necessary clarifications or supplementary instructions can be provided elsewhere, but during the test, reading the checklist should not detract from monitoring student performance.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-12 |

UNCONTROLLED COPY

1. For summative tests, it is usually best to include only items, which *must* be completed. The pass standard for the test is then unambiguous – all items must be rated “satisfactory”, because only essential items are listed.
2. For formative tests, non-essential but desirable performances may be included, but these should be clearly identified as such. The pass standard remains the same – successful completion of all the “must do” items.

3.14 **Construct rating scale items.** Rating scale items generally appear in one of the following forms:

Numerical Rating Scale

Directions: Circle the number, which best describes the student’s performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Poor | Fair | Good Very Good | |
| 1. | Issue a warning order: | 1 | 2 | 3 | 4 |

Graphic Rating Scale

Directions: Place an “X” *anywhere* on the line, at the place, which, best describes the student’s performance.

Poor Fair Good Very Good

1. Issue a warning order

In either case, validity of the rating can be enhanced by descriptive guidance for the scorer, either as part of the item, or in a separate guide. For example:

Scoring Guide

1. Issue a warning order.
   1. Very good – order issued includes all of the following:
      1. Statement of the probable task.
      2. Notice for movement.
      3. Administration instructions.
      4. Rendezvous for orders.
   2. Good – any three of the above.
   3. Fair – any two of the above.
   4. Poor – one or none of the above.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-13 |

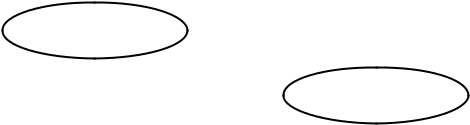
UNCONTROLLED COPY

For whichever format is selected, rating scale items should conform to the following general rules:

1. All points on the scale should be clearly defined. Scorers must be able to discriminate between rating points in a standard way.
2. All points on the scale should reflect a single dimension of performance. For example, rating a memorandum on a scale of “illegible/clear/accurate” would not be valid, because legibility and accuracy are independent performances.
3. The scale should encompass the full range of possible behaviour. For example, “sometimes/usually/always” may leave the scorer in doubt, if the behaviour “never” occurs.
4. Between three and seven reference points should be provided, depending on the degree of discrimination required. Use as few points as will serve the all of the assessment. If an odd number or points is used, scorers may tend to favour the central rating; if it is preferable to force discrimination, an even number of rating points should be used. When particularly fine judgements are both desirable and possible, the graphic rating scale format should be used.
5. Items should be structured in a consistent manner. As far as possible, use the same type of item (with the same number of rating points) throughout the test. Block similar items together to form a grid; then use a single set of instructions for the entire block. Associate higher numbers with higher ratings, and lay out graphic scales with the same orientation; eg, highest ratings always at the right end of the scale.

3.15 **Construct anecdotal record items.** These items usually provide space for short narratives describing the presence or absence of certain behaviours. An aide-memoir for the sorts of attributes to be reported can be helpful. For example:

Leadership Attributes



Circle related attributes(s): Delegation Mental Agility

Forcefulness Composure

Incident Description: Tasked group member to carry out Recce.

Remained calm when during casualty simulation***.***

Date/Time: 11/2/03 Observed by: Ft Lt Jones

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-14 |

UNCONTROLLED COPY

**STEP 4 – PREPARE TEST INSTRUCTIONS**

3.16 In carefully preparing the test sheet, the designer has laid the foundation for a valid test. To preserve validity when the test passes out of the designer’s control and through repeated use by many individuals, requires clear instructions for those who will take the test, and for those who will administer and score it. An example of a checksheet plus the administration instructions can be found at Annex A to this section.

3.17 **Prepare instructions for the student.** Students who are told what to expect in a test situation, and what must be done to succeed, are better able to demonstrate their true abilities. Instructions to be read by (or to) the student should address the following points:

1. The purpose of the test. Usually this can be based on the performance statement for the related objective; e.g., “During this test, you will be required to [performance statement]”. The student should also be told, in common terms, whether the test is formative or summative (e.g., the implications of test failure).
2. A brief description of how the test will be conducted. For example, “You will be presented with a PC 351 radio with a Complete Equipment Schedule. You must assemble the radio and establish communications with… “
3. Time allowed for the test, including details of any pauses or intermediate deadlines.
4. Resources available or denied. This must be consistent with the conditions of the objective, and might refer to:
   1. Reference materials (manuals, notes, job aids).
   2. Tools and equipment.
   3. Assistance (physical or advisory, including whether questions may be asked of the administrator/scorer).
5. The standard of achievement required, to pass the test. This should also specify any critical errors, such as safety infringements, which may result in immediate failure of the test.
6. The action to be taken upon completion of the test.

3.18 **Prepare instructions for the administrator.** The manner in which the test situation is presented to students may have unforeseen and detrimental effects on the validity of the test. To avoid such effects, administrators should be given clear guidance for the following:

1. The purpose, brief description, and time allowed (as in sub-paras 17a-c above) for the test.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-15 |

UNCONTROLLED COPY

1. How to set up for the test, including:
   1. Number of administrators/scorers required.
   2. Test conditions (e.g., tools and equipment, reference material – see para 17d above).
   3. Supervision, safety and hygiene considerations.
   4. Rubbish disposal.
2. How and when to present instructions for the student.
3. How and when cues for student action (e.g., verbal commands, written scenarios) are to be presented.
4. How much assistance, if any may be given by the administrator. Clarification, of instructions should be allowed, but not in the form of hints to students.
5. Control of student and third party movement, as necessary to protect the validity of the test and to avoid distractions.
6. Action to be taken upon completion of each student’s test, and at the end of the test session. This might include resetting the test situation for the next student, clearing up, collecting and securing test materials, and so on.

3.19 **Prepare instructions for the scorer.** Often, test administration and scoring are carried out by the same person. Because these are separate functions, however, it is good practice to write a separate set of instructions for each. Directions for scoring should include:

1. Guidance for the observation of the student; e.g. useful vantage points from which to see certain behaviours.
2. How and when cues for student action (e.g. verbal commands, written scenarios) are to be presented.
3. How much assistance, if any, may be given by the scorer.
4. Circumstances under which the test may be interrupted or halted; e.g. safety infringements, critical errors.
5. How the test sheet is to be completed. This might include guidance for assessing individual items, as well as confirmation of the pass standard.
6. Action to be taken upon completing of the test; e.g. what the student should be told, how results are to be recorded, where the test sheet should be directed.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-16 |

UNCONTROLLED COPY

**STEP 5 – PILOT THE TEST**

3.20 Before the test is put into service, the designer should confirm that the test appears to be valid, and that the test instructions are clear. The following approach is recommended:

1. **Check range and relevance.** Have a knowledgeable colleague scrutinise the complete test, with reference to the related objective, making note of:
   1. Typographical errors.
   2. Incorrect or confusing terminology.
   3. Ambiguous instructions.
   4. Any apparent mismatch between objective and test.
2. **Check reliability across scorers.** Try out the test with one person acting as the student and several others acting, without collusion, as scorers. Compare the scores to determine how closely they match. Where items are scored differently, determine whether this can be attributed to the test or test instructions and, if so, how it can be corrected.
3. **Check reliability across students.** Try out the test with several persons (the more typical of the student group, the better) acting as students, and briefed to act out both satisfactory and unsatisfactory performances. Determine the extent to which scoring correctly discriminates between the behaviours. At the same time, confirm administration details.

**SECTION SUMMARY**

3.21 Practical tests, by which we assess the application of learned skills in realistic situations, are particularly useful for summative testing.

3.22 To be valid, a practical test should:

1. Comprehensively measure the essential processes and/or products detailed in the training or enabling objective.
2. Rely as much as possible on checklist items, supplemented as necessary by rating scale and anecdotal record items.
3. Include detailed instructions for students, administrators and scorers.
4. Be piloted.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-17 |

UNCONTROLLED COPY

**ANNEX A TO**

**SECTION 3**

**EXAMPLE PRACTICAL TEST**

**Administrative Instructions including Student Briefing**

1. In this test, the student must demonstrate achievement of EO 1.4, by performing the Chemical Immediate Action Drill. Allow five minutes to brief, test and score each student. One scorer (who can also administer the test) is required to separately observe each student.
2. This test should be conducted outdoors. The student must wear full IPE (detector paper affixed) with helmet, respirator in haversack and NBC gloves stowed. No assistance is allowed. The administrator requires a spray gun with CATM. Scorers each require a watch or stopwatch.
3. Brief each student as follows:

“During this test, you will be required to perform the Chemical Immediate Action Drill, when triggered by warning of an attack. You must pass this test to pass EO 1.4. To pass the test, you must complete the drill in the correct sequence, donning the respirator within 9 seconds, and without inhaling or opening your eyes until the respirator has been cleared of bad air. You will not receive any assistance during the test. When you have completed the drill, stand to attention. Are there any questions about these instructions?

Your unit is at MOPP 3, and you have been warned that chemical weapons have been used on neighbouring units. Standby…”

1. Initiate the chemical alarm by any of the standard methods, at the same time spraying a small amount of CATM, ensuring that some of it falls on the student’s detector paper.
2. When students finish the test, they should be sent immediately to the Chemical Immediate Decontamination Drill test area. Instruct students **not** to remove respirators or gloves en route.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-18 |

UNCONTROLLED COPY

**Marking/ Scoring Instructions**

1. You should have a watch with which to time seconds passed, and a copy of the attached checklist for each student to be scored. Start timing from the instant the chemical alarm is given, and stop when the respirator has been donned and checked for fit. Do not assist or interfere with the student during the test, unless the student shows signs of fainting or panic.
2. While completing the checklist, watch for the following:
   1. The student does not inhale from the time the alarm is given until blowing out hard and shouting “GAS, GAS, GAS” (steps 1 top 10).
   2. When adopting a position to lessen contamination (step 4) the student leans forward, head down, back to the wind and tries to protect the hands with the body.
   3. When the student removes the helmet (step 5), it is held between the knees or hung over the arm by the chinstrap.
   4. The student dons the respirator (step 8) chin first, pulling the harness over the head, *and* feels to ensure proper fit.
   5. The student decontaminates the hands (step 14) using the BLOT-BANG-RUB technique.
   6. NBC gloves (step 15) are tucked under and over the sleeves of the NBC smock.
   7. Correct fitting of IPE (step 16) includes closing all fasteners, leaving no skin exposed, and ensuring a good seal between hood and respirator.
   8. When the helmet is replaced (step 17) the chinstrap must be placed behind the canister.

Upon completion of the test, immediately inform the student whether the test was passed or failed. If *all* checklist items are rated Yes, tell the student that the performance was “excellent”. Pass the completed checklist to the test administrator before testing the next student

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-19 |

UNCONTROLLED COPY

**Checklist – Perform the Chemical Immediate Action Drill**

Student Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Course\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

No:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date\_\_\_\_\_\_\_\_\_

Instructions: Beside each item in the following list, place a tick mark in the appropriate box. Record the time taken for steps 1-10 against serial 13. ***Steps 1- 10 must be performed in the order shown.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Step |  |  | Description | |  | Yes | No |
|  |  |  |  |  |  |  |  |
| 1 | Hold breath |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 2 | Close eyes |  |  |  |  |  |  |
|  |  | | |  |  |  |  |
| 3 | Remove spectacles (if worn) | | |  |  |  |  |
|  |  | | | | |  |  |
| 4 | Adopt position to lesson liquid contamination *(leans forward, head down, back* | | | | |  |  |
|  | *to the wind and tries to protect the hands with the body).* | | | |  |  |  |
|  |  | |  |  |  |  |  |
| 5 | Remove helmet | |  |  |  |  |  |
|  | *(held between the knees or hung over the arm by the chinstrap)* | | | |  |  |  |
|  |  | |  |  |  |  |  |
| 6 | Clear hood from head | |  |  |  |  |  |
|  |  | | |  |  |  |  |
| 7 | Take respirator from haversack | | |  |  |  |  |
|  |  | | | | |  |  |
| 8 | Put on respirator (*chin first, pulling the harness over the head, and feels to ensure* | | | | |  |  |
|  | *proper fit)* |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 9 | Pull up hood |  |  |  |  |  |  |
|  |  | | | |  |  |  |
| 10 | Blow out hard and shout “GAS, GAS, GAS” | | | |  |  |  |
|  |  | | |  |  |  |  |
| 11 | Steps 1-10 completed in correct order | | |  |  |  |  |
|  |  | | | | |  |  |
| 12 | Steps 1-10 completed without inhaling or opening eyes | | | | |  |  |
|  |  | | |  |  |  |  |
| 13 | Steps 1-10 competed within 9 seconds | | | Time = | secs |  |  |
|  |  | | | |  |  |  |
| 14 | Decontaminate hands the (*BLOT-BANG-RUB)* | | | |  |  |  |
|  |  | | | | |  |  |
| 15 | Put on gloves (*tucked under and over the sleeves of the NBC smock)* | | | | |  |  |
|  |  | | | | |  |  |
| 16 | Check correct fitting of IPE (*closing all fasteners, leaving no skin exposed, and* | | | | |  |  |
|  | *ensuring a good seal between hood and respirator)* | | | |  |  |  |
|  |  | | | | |  |  |
| 17 | Replace helmet (*chinstrap must be placed behind the canister)* | | | | |  |  |
|  |  | | |  |  |  |  |
| 18 | Close respirator haversack flap | | |  |  |  |  |
|  |  | | |  |  |  |  |
| 19 | Check for signs of liquid contamination | | |  |  |  |  |
|  |  | |  |  |  |  |  |
| 20 | Stow spectacles (if worn) | |  |  |  |  |  |
|  |  |  |  |  | |  |  |
| Grade (circle one): | | Fail | Pass | Pass (excellent) | | | |

Fail – any shaded item is rated “No” (excluding Step 3 if not applicable)

Pass – all shaded items are rated “Yes” (excluding Step 3 if not applicable)

Pass (excellent – all items are rated “Yes” (excluding Steps 3 & 20 if not applicable)

Signature of Scorer:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-20 |

UNCONTROLLED COPY

**SECTION 4 - DEVELOP A THEORY ASSESSMENT**

**THE ROLE OF THEORY TESTS**

4.1 The emphasis on training is on developing the ability to do things, but much of this ability depends on the mastery of underlying facts, concepts, and principles, theory tests are used to measure this abstract dimension of learning. They provide a check on student progress toward a more concrete goal- the practical application of knowledge.

4.2 In military training, theory tests fill an intermediate, supporting role; they should never serve as the culmination of training. Theory tests should be used for:

1. Supplementing practical tests, where indicated by the selected sampling strategy. Especially where the assessment of underpinning knowledge is essential e.g. for safety issues, or for accreditation by awarding bodies.
2. Summative testing of theory orientated Enabling Objectives, when indicated by the Assessment Specifications or Assessment Strategy.
3. Formative testing of facts lessons.

4.3 The administration of theory tests is usually less demanding of time and resources than that of practical tests. This apparent economy is offset by a need for painstaking test preparation, scoring, and follow-up analysis, all of which are essential to establish and maintain test validity. For these reasons, it is recommended that theory testing be used only when practical testing cannot achieve the aims of formative and summative assessment.

4.4 There are 5 steps in the construction of a theory test:

Step 1 – Determine the Test Content and Standard

Step 2 – Select the Appropriate Format

Step 3 - Construct the Test

Step 4 – Prepare Test Instructions

Step 5 – Pilot the Test

**STEP 1 – DETERMINE THE TEST CONTENT AND STANDARD**

4.5 **Examine the objective**. The first step in the construction of a theory test is to examine the relevant Training Performance Statement (Training Objective), Enabling Objective, or Instructional Specification. The developer must pick out the relevant details needed to achieve the best possible match between the objective and the test. Further guidance will be given by the analysis of the Assessment Strategy part of the Course Specification Documentation or the Assessment Specification. This will give details of the test type, purpose, administration, TOs and EOs covered, details of marking, pass/fail criteria and consequences and actions to be taken should a trainee fail the test.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-21 |

UNCONTROLLED COPY

4.5 **Determine the test performances or learning outcomes**. The designer should list the facts, concepts and principles that ***must*** be learned by the student to achieve the stated objective. By definition, these “must knows” can usually be found among the teaching points provided in the instructional specifications. Where instructional specifications are not available, essential theory elements can be identified among the conditions and standards of Enabling and Training Objectives. For example, if an objective calls for a patrol to be conducted “in accordance with the first principles of armoured reconnaissance stated in RAC Tactics Pamphlet No 2”; knowledge of these principles might be included in a theory test.

4.6 **Determine test conditions.** Instructional specification or objectives should also indicate which references, if any, should be provided to the student during a test. Generally, if a reference is used on the job, the same reference should be allowed during the test.

4.7 **Determine test standard.** In theory tests, two issues must be addressed: the degree of difficulty appropriate for each of the listed test performances, and the overall pass standard for the test:

1. The degree of difficulty required for each test performance can be related to the **Levels of learning** shown in figure 6. The designer plans for range and relevance by selecting the appropriate levels of learning and by assigning a mix of test items that reflects the requirement of the objective. This is often laid out in a **Test Plan** or **Test Grid** an example of which is shown at figure 7.
2. The overall pass standard for the test must be consistent with its purpose:
   1. The pass mark for a formative test serves primarily as a motivator for students, and, if used at all, should be established at a level that indicates adequate progress.
   2. The pass mark for a summative tests serves as evidence of mastery of a subject. Because the test items have been planned to match the requirements of the objective, a high standard is usually indicated.

E.g. A training objective for basic training might require that the soldier “discriminate between enemy and friendly Armoured Fighting Vehicles”, within certain conditions, to a standard of “without error”. (Considering the potential consequences of error, would a lesser standard – even for testing – be acceptable?)

This does ***not*** necessarily imply a high failure rate; if students are properly selected, motivated, instructed and practised, most or all can and should expect to pass the test. (N.B. The Army Training and Recruiting Agency (ATRA) suggests all tests in phase 3 training should have at least an 85% pass rate and anything less than 80% means either the tests, the training or the students need to be reviewed).

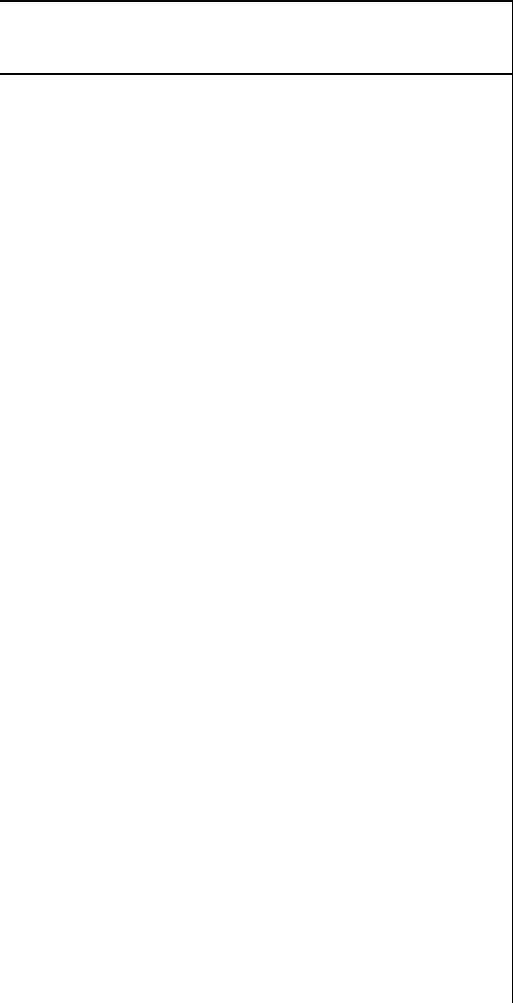
|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-22 |

UNCONTROLLED COPY

**FIGURE 6: Levels of Learning**

|  |  |  |
| --- | --- | --- |
| Category of | Description | Explanation |
| Learning |  |  |
| Knowledge | Recall of terms, facts, methods, | Knowledge is defined as the remembering of |
|  | concepts, principles. | previously learned material. This may involve |
|  | i.e. | the recall of a wide range of material, from |
|  | Knows common terms | specific facts to complete theories, but all that is |
|  | Knows specific facts | required is the bringing to mind of the |
|  | Knows methods and procedures | appropriate information. Knowledge represents |
|  | Knows basic concepts | the lowest level of learning outcomes in the |
|  | Knows principles | cognitive domain. |
| Comprehension | Explain or summarise material, or | Comprehension is defined as the ability to grasp |
|  | predict consequences or effects | the meaning of material. This may be shown by |
|  | i.e. | translating material from one form to another |
|  | Understands facts and principles | (words to numbers), by interpreting material |
|  | Interprets verbal material | (explaining or summarizing), and by estimating |
|  | Interprets charts and graphs | future trends (predicting consequences or |
|  | Translates verbal material to | effects). These learning outcomes go one step |
|  | mathematical formulas | beyond the simple remembering of material, |
|  | Estimates future consequences implied | and represent the lowest level of understanding. |
|  | in data |  |
|  | Justifies methods and procedures |  |
| Application | Use rules, methods, concepts, | Application refers to the ability to use learned |
|  | principles, laws and theories in new and | material in new and concrete situations. This |
|  | concrete situations. | may include the application of such things as |
|  | i.e. | rules, methods, concepts, principles, laws, and |
|  | Applies concepts and principles to new | theories. Learning outcomes in this area require |
|  | situations | a higher level of understanding than those |
|  | Applies laws and theories to practical | under comprehension. |
|  | situations |  |
|  | Solves mathematical problems |  |
|  | Constructs charts and graphs |  |
|  | Demonstrates correct usage of a |  |
|  | method or procedure |  |

Sample Military Objectives or Teaching Points



* Name officers of a unit or HQ
* Name parts of a weapon
* Recite the National Anthem
* List the Principles of War
* List the parts of a Training Objective
* Define assessment
* State the Principles of Defence
* Explain the different parts of a TO
* Describe the difference between Norm and Criterion Referenced Testing
* State a problem in your own words
* Describe an example of ‘Fire and Movement’
* Describe the operation of a weapon
* Give examples of the use of “depth” in defence.
* Use a manual to repair an equipment or vehicle,
* Calculate leave entitlement, etc
* Decide whether a soldier’s conduct is Insubordination.
* Design a defensive position with “Depth”
* Tactically site a weapon
* Layout a defensive position
* Identify a vehicle from a formation



Issue V1.0 Aug 07 2 - 23

UNCONTROLLED COPY

|  |  |  |
| --- | --- | --- |
| Category of | Description | Explanation |
| Learning |  |  |
| Analysis | Examine the structure of material to | Analysis refers to the ability to break down |
|  | determine the principles and | material into its component parts so that its |
|  | relationships between parts. | organizational structure may be understood. This |
|  | i.e. Recognizes unstated assumptions | may include the identification of the parts, analysis |
|  | Recognizes logical fallacies in | of the relationships between parts, and |
|  | reasoning | recognition of the organizational principles |
|  | Distinguishes between facts and | involved. Learning outcomes here represent a |
|  | inferences | higher intellectual level than comprehension and |
|  | Evaluates the relevancy of data | application because they require an |
|  | Analyzes the organizational structure | understanding of both the content and the |
|  | of a work (art, music, writing) | structural form of the material. |
|  |  |  |
| Synthesis | Form new patterns or structures in a | Synthesis refers to the ability to put parts together |
|  | creative fashion | to form a new whole; This may involve the |
|  | i.e. Writes a well organized theme | production of a unique communication (theme or |
|  | Gives a well organized speech | speech), a plan of operations (research proposal), |
|  | Writes a creative short story (or poem, | or a set of abstract relations (scheme for |
|  | or music) | classifying information). Learning outcomes in this |
|  | Proposes a plan for an experiment | area stress creative behaviors, with major |
|  | Integrates learning from different areas | emphasis on the formulation of new patterns or |
|  | into a plan for solving a problem | structures. |
|  | Formulates a new scheme for |  |
|  | classifying objects (or events, or ideas) |  |
| Evaluation | Judge the value of material with | Evaluation is concerned with the ability to judge |
|  | relation to definite criteria. | the value of material (statement, estimate, |
|  | i.e. Judges the logical consistency of | research report) for a given purpose. The |
|  | written material | judgements are to be based on definite criteria. |
|  | Judges the adequacy with which | These may be internal criteria (organization) or |
|  | conclusions are supported by data | external criteria (relevance to the purpose) and |
|  | Judges the value of a work by use of | the student may determine the criteria or be given |
|  | internal criteria | them. Learning outcomes in this area are highest |
|  | Judges the value of a work by use of | in the cognitive hierarchy because they contain |
|  | external standards of excellence | elements of all of the other categories, plus |
|  |  | conscious value judgments based on clearly |
|  |  | defined criteria |
| **FIGURE 6. Continued** |  |  |

Sample Military Objectives or Teaching Points



* Locate a fault in a photocopier
* Select an Observation Post position by map inspection
* Distinguish between facts and opinions
* Precis an argument or paper.
* Identify the relationship between different weapons in a fire plan.
* Determine the weak points in a defensive position
* Write an essay
* Design a leadership exercise
* Plan the layout of a defensive position
* Design a weapon
* Write an estimate for a defensive operation
* Select the most efficient solution from an array of options
* Selects the best candidate for a job
* Decide whether an attack plan is likely to succeed
* Write a trials report
* Assess an estimate for a defensive position



Issue V1.0 Aug 07 2 - 24

UNCONTROLLED COPY

**Figure 7: Sample Test Plan or Test Grid**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Teaching | Knowledge | Comprehension | Application | Total |
| Point | Questions | Questions | Questions |  |
| Lists the | 4 |  |  | 4 |
| components |  |  |  |  |
| of defence | *(a)* | *(b)* | *(b)* |  |
| Explains the |  | 4 |  | 4 |
| role of tanks in |  |  |  |  |
| FIBUA |  |  | *(b)* |  |
| Describes the | 1 | 3 |  | 4 |
| phases of an |  |  |  |  |
| attack | *(c)* |  | *(b)* |  |
| Applies the | 1 | 1 | 2 | 4 |
| principles for |  |  |  |  |
| employing |  |  |  |  |
| armour | *(c)* | *(c)* |  |  |
| Totals | 6 | 8 *(d)* | 2 | 16 |

Notes:

1. Numbers shown can relate to either no. of questions or marks available.
2. Questions at these levels would be testing above the level of training.I am not sure what this means
3. Extra questions have been added in a lower cognitive level to balance test.Or here. The point is that for a formative test you may want to test at all the levels below the final one to be achieved so that feedback can be provided as to where failure has been caused.
4. Overall test is weighted at comprehension level. If aim was for more application then marks could be weighted accordingly. Watch out for tests that are heavily weighted only on knowledge. In this case only low levels of the cognitive domain were assessed.

I would allocate numbers of questions that **could be** asked to each cell of the grid and then decide the sampling strategy for number of questions **to be** asked from that.

**STEP 2 - SELECT THE APPROPRIATE FORMAT**

4.8 **Choose items by class**. Theory test items can be classified by the manner in which the student must respond: either by selecting among available options, or by supplying an answer. **Selection items** include true or false, matching and multiple-choice questions and are also examples of **objective tests**. **Supply items** include short answer and extended answer questions and are also called **subjective tests**. In determining the best mix of items, the designer should consult the test plan, and consider the advantages and disadvantages shown in figure 8. As a general rule, supply items should be avoided unless the higher levels of learning are to be tested or the piece of information has to be recalled from memory only.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-25 |

UNCONTROLLED COPY

**Figure 8: Item type selection considerations**

|  |  |  |
| --- | --- | --- |
| Factor | Selection Items | Supply Items |
| Levels of | Useful for testing knowledge, | Useful for testing evaluation, syntheses, analysis |
| Learning | comprehension and some application. | and some application. Inefficient for testing lower |
|  | Inefficient for testing higher levels | levels. |
| Sampling | Permits larger sampling through many | Limits sampling because fewer questions can be |
|  | items | asked |
| Preparation | Preparation of unambiguous items is | Preparation is relatively easy |
|  | difficult and time consuming |  |
| Scoring | Objective, easy, quick, and consistent. | Subjective, difficult, slow and often inconsistent |
|  | Scores may be distorted by reading or |  |
|  | guessing. |  |

4.9 **Choose items by type**. After deciding which general classes of item are appropriate, the designer should choose the specific item type to be used for each question. Guidelines for selecting item types are as follows:

1. **True or False**. These questions should be avoided, because they allow a 50% margin for guessing. Also, it is difficult to state items in entirely true of false terms, without either testing trivia or providing clues to the correct response.
2. **Matching**. These items provide a compact way to test many related elements, though usually at the knowledge level (e.g. matching labels to components, or rearranging items in a correct order). With care, they can also be used to test some aspects of comprehension (e.g. matching tools to jobs, or causes to effects).
3. **Multiple-choice**. These items are quite versatile. They are often used to test at the knowledge level; but provided that a novel element is introduced in the question, they can also test comprehension and application. They permit objective scoring while presenting minimal risk from student guessing, provided that they are carefully constructed.
4. **Short Answer**. These items have the advantage or requiring that a student produce, rather than recognise, the correct answer. The required response can range form a single word to a paragraph, from a numerical expression to a worked formula, from labelling a component to constructing a diagram, and so on. Depending on the form used, these items can be used for testing knowledge, comprehension, application or analysis. However, construction of unambiguous items is rather more difficult than is commonly appreciated. Consistent marking can only be achieved if the question specifies the elements of the response required, and if detailed scoring instructions are provided.
5. **Extended answer**. These items are best used to test the higher learning

levels: analysis, synthesis and evaluation. Because these items are very difficult to score, and because drawing or writing ability can greatly influence marking, extended

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-26 |

UNCONTROLLED COPY

answer items should only be used when high level testing is clearly mandated by the test plan.

4.10 By using a test plan, the designer can select item types to meet the needs of the objective. This will ensure that time spent in item construction is properly focused.

4.11 **Test Length**. To a large extent, the number of items to be included in a theory test is a matter of practicality. The designer must balance two competing considerations:

1. Longer tests (hence more complete sampling of student knowledge), will more accurately measure the individual student’s true ability. Tests which are designed to support important discriminations between passing and failing, or in determining relative standing among students, should be longer tests.
2. Time spent in the preparation of test items, test administration and scoring will be directly proportional to the test length. Shorter theory tests may be necessary to allow more resources to be devoted to summative practical testing.

**STEP 3 – CONSTRUCT THE TEST**

4.12 **General Guidelines**. Test items of all types should conform to the following:

1. The item tests an important, practical, and relevant element of learning – a Key Learning Point (KLP). Do not test trivial knowledge of irrelevant abilities.
2. The item is simply and directly stated in terms understandable to the student. Visual aids can often clarify the problem being presented, so that text can be kept brief. Arrange question layout so it will be unnecessary for student to refer to more than one page in answering any given question.
3. The item employs positive wording whenever possible. If this cannot be done, emphasise the negative word(s).
4. The item avoids clue words or grammatical errors which might allow the student to correctly guess or discard a response, either within the item itself, or by a link with other items on the test.
5. The item is worded in a way that demands the correct answer, and only the correct answer. Phrases such as “what would you do if...” “state your understanding of ..”, or “can you explain..” permit a subjective response which may bear little relation to the desired answer.
6. If answers are to be written on the question paper allow sufficient space for all responses. This should not necessarily indicate the length of the answer required as handwriting varies.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-27 |

UNCONTROLLED COPY

1. The mark allocation for each question should be clearly displayed. If the question is looking for the student to list items then they should have some indication of the number to be listed in the stem of the question.

4.13 **Constructing True/False items** . In true/false questions the student is presented with a statement which he/she has to judge to be true or false as shown in the following example:

Example:

The RB199 engine has an auto-ignite circuit that re-ignites if it flames out.

True/False

If these are to be used, there are a few simple rules that should be followed in order to ensure that the questions are effective.

1. Each question should contain only one statement to be judged true or false.
2. Negative statements should be avoided.
3. Avoid absolutes such as never; always; all; every, and no. Statements involving such words are likely to be incorrect as exceptions can be found. Similarly avoid statements containing some; any; generally, which are most frequently true.
4. Statements should be short so as to be clear and easily understood.

4.14 **Constructing matching items**. Matching questions can be presented in many forms, but generally consist of a set of directions followed by two related lists, as shown in the following example:

Q1. In the space provided to left of each appointment in the first column, write the letter of the established rank from the second column. Each rank may be used once, more than once, or not at all.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Appointment | |  |  | Rank | |
|  |  |  |  |  |  |
| \_\_\_\_\_\_ 1. | Adjutant | | A. Colonel | | |
| \_\_\_\_\_\_ 2. | SO2 | |  | B. Lieutenant Colonel | |
| \_\_\_\_\_\_ 3. | CO | | C. Major | | |
| \_\_\_\_\_\_ 4. | Coy Comd | |  | D. Captain | |
|  |  |  |  | E. Lieutenant | |

In constructing matching items, the following rules should be observed:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-28 |

UNCONTROLLED COPY

1. Each list must contain items that are, within the context of the training, linked by a single theme. Each item on the response list should be a possible match for each item on the other list; the basis for the match should be clearly stated in the directions.
2. The two lists should be of unequal length, to prevent students gaining marks by process of elimination.
3. To minimise the amount of “searching” required of the student:
   1. Keep items brief, and limit columns to four to seven items.
   2. List the items in the response column in a logical order.
   3. Place the entire question on one page.

4.15. **Construct multiple -choice items**. A multiple-choice item consists of a ***stem*** and several ***options***. One of the options is correct, while the others serve as ***distracters*** for the less knowledgeable students. This is the most common type of objective test item used in assessment.

For example:

Q2. An easily identified point on the ground, at which Liaison Officers from flanking units or formations meet, is a: *(stem)*

|  |  |  |  |
| --- | --- | --- | --- |
| *(Options)* | a. | Check point | *(Distracter)* |
|  | b. | Contact point | *(Answer)* |
|  | c. | Coordination point | *(Distracter)* |
|  | d. | Concentration point | *(Distracter)* |

Answer: \_\_\_\_\_\_

In constructing multiple-choice items, the following should be observed:

1. The problem or question should be clearly stated in the stem, so that the student is able to anticipate the correct answer from reading the stems alone.
2. An item should have only one correct or clearly best response. Distracters should mislead less knowledgeable students, but must also be clearly implausible to knowledgeable students. Avoid choices that are trivial, implausible or obviously wrong. Wrong answers should represent common errors or misconceptions. Include at least 4 options.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-29 |

UNCONTROLLED COPY

POOR EXAMPLE

An important qualification for a lorry driver is that he/she possesses:

1. A knowledge of the history of transportation.
2. Friends in the Motor Pool
3. A licence to drive a lorry
4. A good speaking voice

BETTER EXAMPLE

An important qualification for a lorry driver is that he/she possesses:

* 1. No accidents against his/her name.
  2. No physical deficiencies
  3. A licence to drive a lorry
  4. Good vision without glasses

1. As much of the item as possible should be included in the stem. This provides economy in presentation and cuts out unnecessary reading.
2. Avoid “none of the above“ which only requires the student to recognise wrong answers, rather than demonstrating knowledge of the correct answers. Also avoid “ all of the above”, or combinations of responses such as “b and c”. All these items compromise validity by placing emphasis on reading comprehension, and are often used as poor substitutes for more plausible distracters.
3. Ensure that each of the options is grammatically consistent with the stem. The designer should check this by reading through the stem, followed by each of the options, separately. A mismatch between singular and plural forms, or misuse of the articles “a” and “an”, can reveal the distracters.

POOR EXAMPLE

On a map a trail is shown by a :

* 1. Single broken line.
  2. Two parallel broken lines
  3. Dots and dashes.
  4. Two parallel solid lines

1. Ensure that answers appear randomly among the options, throughout the test as a whole. Avoid any pattern which might aid guessing, such as over reliance on the “b” position, or a tendency for the answer to be longer than the distracters.
2. Be careful of the clue words, or similar words, which link the stem and the answer. Such words can, however, be used to strengthen distracters. Avoid using absolute words such as all, none, always or never as absolutes are rare.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-30 |

UNCONTROLLED COPY

1. Test designers should aim to design questions that test not just knowledge of facts but also the application of the knowledge into practical scenarios. Multiple choice questions can be effectively used to test application by making students select the correct action.

GOOD EXAMPLE

On a night patrol you are caught in the light of a flare from an enemy plane. What immediate action should you take?

1. Run for the nearest cover.
2. Drop to the ground.
3. Remain where you are.
4. Shoot at the plane.

4.16 **Construct short answer items**. Short answer items can be expressed as questions requiring separate answers, or as sentences requiring completion by filling in blanks. Because unambiguous sentence completion items are very difficult to construct, it is recommended that designers favour the short answer approach. For example:

AMBIGUOUS EXAMPLE

Q 3. When withdrawing by day, troops in contact must be prepared to use (fire and movement). This can be done by withdrawing the forwards elements through those (in depth), or by forward troops conducting this (fire and movement) themselves.

Note that eliminating, or altering the position of, any of the blanks in this example gives clues to the desired answers.

BETTER EXAMPLE

Q 3. When withdrawing by day, troops in contact must be prepared to use fire and movement. State two ways in which this manoeuvre can be carried out.

*The second question is clearer, and provides a better test of the student comprehension of the principle involved, even though the term “ fire and movement” is provided within the questions.*

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-31 |

UNCONTROLLED COPY

Guidelines for the construction of short answer items are as follows:

1. State clearly and definitely the form of answer required, using as many statements as necessary to do so. The performance required of the student should match the learning level to be tested; “describe” for example, would generally demand higher learning than “list”.
2. If calculations are required, the student should be told how much work must be shown.
3. For any question requiring a numerical response, the degree of accuracy and unit of expression should be indicated; e.g. “ to the nearest metre”.
4. Attempt to word the item as a direct question, rather than as an incomplete statement. If the completion form cannot be avoided, arrange the statement so that the blank appears at (or near) the end – this precaution will aid clarity.

4.17 **Construct extended answer items**. Extended answer items involve more freedom of expression than short answer items. They have the somewhat false attraction of seeming easy to construct, but require painstaking and time-consuming scoring procedures if some measure of reliability is to be gained. Examples of extended answer questions are:

Q5. Explain how the principles of war, Surprise and Concentration of Force, were reflected in Coalition ground operations in Kuwait and Iraq in February 1991.

Q6. Write an appreciation of the situation for a brigade defence

of....

Q7. Draw a circuit diagram that will produce the following output...

Q8. Write a computer program that will print a table that lists....

Guidelines for the construction of extended answer questions are as follows:

1. Use extended answer items only to test students’ abilities at the analysis, synthesis, and evaluation levels; and only when practical tests or alternative item formats are unsuitable.
2. Avoid undue emphasis on speed, writing or drawing ability, or other factors not specified in the training standard.
3. Have all students answer the same question(s). In educational settings,

students are sometimes allowed to choose from among several alternative questions.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-32 |

UNCONTROLLED COPY

For testing achievement in military training, this approach compromises test validity and is inappropriate.

1. Provide as much structure for student responses as possible. Broadly dated questions may elicit broad answers that do not necessarily demonstrate the required learning. Give additional information concerning the scope of the answer, how it should be presented, and the manner in which it will be scored.

4.18 **Compile test items**. After all test items have been constructed, the designer should compile them as follows:

1. Arrange the test in a logical sequence. Items should be grouped by associated Training or Enabling Objective, and this distribution should be checked against the test plan/test grid. It is recommended that each section of the test is headed by the related performance statement – this helps the student (and scorers) focus on the task at hand.
2. Within each section of the test, sequence items in ascending order of difficulty. As much as possible, keep similar item types grouped together, so that instructions for answering can be kept to a minimum. For example:

EO1 – Define terms used in mine warfare.

Matching question (knowledge level)

Multi-choice questions (comprehension level)

Short answer questions (application level)

EO2 – Detect mines

Short answer questions (comprehension/application level)

Extended answer question (analysis level)

EO3 – Describe mine clearance equipment

Multi-choice questions (knowledge level)

Short answer questions (comprehension level)

Multi-choice questions (application level)

1. Read through the assembled items to check that construction guidelines have been observed. Ensure no item provides an answer or clue to another item.

**STEP 4 – PREPARE TEST INSTRUCTIONS**

4.19 At this point, the designer has organised a set of valid test items. To carry that validity forward into a usable test paper, comprehensive instructions must be prepared for those who will use it: students, administrators and scorers.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-33 |

UNCONTROLLED COPY

4.20 **Prepare general instructions for the student**. General instructions to be read by (or to) the student should include the following:

1. The purpose of the test. Usually this can be based on the performance statement for the related objective; e.g., “During this test, you will be required to [performance statement]”. The student should also be told, in common terms, whether the test is formative or summative (e.g., the implications of test failure).
2. A brief description of the scope and content of the test. This should include a brief overview of the layout of the test paper: the number of sections, pages and questions should be indicated, and students should be directed to check that their own test paper has been correctly reproduced and collated.
3. Time allowed for the test, including details of any pauses or intermediate deadlines.
4. Resources available or denied. This must be consistent with the conditions of the objective, and might refer to:
   1. Reference material (manuals, notes, job aids).
   2. Tools and equipment for writing, drawing, or calculating.
   3. What type of questions may be asked of the administrator, and when.
5. The standard of achievement required to pass the test. In the case of objective tests, directions should also inform students whether they would be penalised for incorrect answers.
6. How and where the student is to be identified on the test paper or answer

sheet.

1. Instructions for the maintenance of discipline during the test (silence, movement control etc).
2. The action to be taken when the student completes the test. This will normally include steps to recover all test materials, to preserve the validity of the test for later use with subsequent courses.

4.21 **Prepare specific instructions for the student**. In addition to general instructions, students should be given specific instructions for the completion of each section of the test: how a response is to be given, and how a response already recorded can be changed.

Usually, a single set of instructions can precede each group of similarly formatted items (e.g. multi-choice, short answer). The marks available for each question should also be clearly indicated. Instructions should be accompanied by examples, which can sometimes be conveniently displayed on separate sheets or foldout pages.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-34 |

UNCONTROLLED COPY

4.22 **Prepare instructions for the administrator**. The manner in which the test situation is presented to students may have unforeseen effects on the validity and reliability of the test. Bad administration can bias results and scores can vary for reasons other than genuine differences in ability. To avoid such effects, administrators should be given clear and firm guidance for the following:

1. Purpose, brief description, and time allowed (as in sub paras 20a-c above).
2. Security of test materials (to avoid compromising the random

nature of the test).

1. How to set up for the test, including
   1. Number of administrators required.
   2. Reference material.
   3. Tools and equipment (including spares).
   4. Supervision for cheating.
2. How and when instructions for the student should be presented.
3. How much assistance, if any, may be given by the administrator. Clarification of instructions should be allowed, but not in the form of hints to students.
4. Control of student and third party movement around the test location, as necessary to protect the validity of the test and to avoid distractions.
5. Action to be taken upon completion of each student’s test, and at the end of the test session. This might include clearing up, collecting and securing test materials, and so on.

4.23 **Prepare instructions for the scorer**. Often, the same person can carry out test administration and scoring. However, it is good practice to write a set of instructions for each. Directions for scoring should include:

1. A ***test key*** that lists the correct answers. For objective tests, this should be presented in a form which will most conveniently enable the score to compare student responses with the required answers; for example, a correctly completed test sheet, or answer slip which can be placed alongside a student’s paper.
2. For extended answer items, the key should be in the form of an ideal answer, with the crucial elements indicated. When items demand particularly complex solutions, or permit alternative solutions, consider either or both of the following:
   1. Training for scorers, which might include discussion to agree a common approach (i.e. a form of standardisation).



|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-35 |

UNCONTROLLED COPY

* 1. Use of two or more independent scorers who contribute to a combined score. Whilst time consuming, this method significantly improves the validity of the final score (i.e. a form of moderation).

1. A method by which, where the student has responded incorrectly, correct answers are to be indicated on student papers. For example, the correct option of a missed selection item could be marked with a highlighter pen. When the students review the test, this will help emphasise the correct over the incorrect, enhancing the formative effect. Testing can and should be a learning opportunity for both trainees and instructors. This learning opportunity is missed without good feedback. Trainees should have marked scripts and scores given back to them as soon as possible after the test. The original test paper could be designed with feedback in mind and have a space for the scorer to put remarks for the students. If marked papers are to be shown to students, then the scorer should be told this in advance of the marking process.
2. Direction for awarding of scores for individual items, and for the test overall. It should be sufficient to award a single point for each selection item, and for each crucial element of a supply item. Issues of relative importance, difficulty, weighting, and overall test standard have been accounted for in Step 1 – Determine Test Content and Standard (paras 5 – 8 above). In preparing this direction, the designer should consider the following:
   1. The application of a mathematical formula to “correct” students’ scores for the effects of guessing is not recommended. Students are unlikely to guess blindly, and well-crafted items will have already discriminated between students who firmly know the answer and those who don’t.
   2. If grades are to be used, the number of categories should be

minimised. For example, use of the letter grades A, B, C, D and E implies a degree of discrimination within a test which many already be suspect; adding a “+” and “-“ values to these, creating 15 categories, would rarely be justified. In general, the shorter the test, the fewer the grading categories which can be validly applied.

**STEP 5 – PILOT THE TEST**

4.24 Before a test is put into service, the designer should confirm that the test appears to be valid, and that the test instructions are clear. The following is recommended:

1. **Check range and relevance**. Have a knowledgeable colleague scrutinise the complete test, with reference to the related objective, making note of:
   1. Typographical errors.
   2. Incorrect or confusing terminology
   3. Ambiguous instructions.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-36 |

UNCONTROLLED COPY

* 1. Any apparent mismatch between the objective and test.

1. **Check reliability across scorers**. Try out the test with one person acting as the student and several others, without collusion, scoring the test. Compare the scores to determine how closely they match. Where items are scored differently, determine whether this can be attributed to the test or test instructions and, if so, how it can be corrected.
2. **Check reliability across students**. Try out the test with several persons (the more typical of the student group, the better), who have been briefed to present both satisfactory and unsatisfactory performances, acting as students. Determine the extent to which the scoring correctly discriminates between the behaviours. At the same time, confirm administration details. The time required to complete the test should also be recorded and the results analysed to identify remaining faults.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-37 |

UNCONTROLLED COPY

**SECTION 5 – EVALUATING ASSESSMENTS**

**GENERAL**

5.1 Test evaluation is an important aspect of the piloting stage of test design and the Internal Validation of training. Responsibility for the monitoring and revision of tests should be assigned as discussed in Section 2, paras 10 and 11.

5.2 Even well designed and carefully piloted tests must be periodically evaluated, within the context of the overall strategy for assessment. Because tests are used as tools for important assessment decisions, confidence in tests must be renewed on a regular basis.

5.3 Test evaluation confirms the effectiveness of the assessment plan, checks the validity of each of the tests employed in support of that plan and ensures that test results are properly interpreted. A check sheet which outlines the steps that can be undertaken in evaluating assessment can be found at Annex A to this section.

**STEP 1 – EVALUATE THE ASSESSMENT STRATEGY**

5.4 **Review the assessment strategy.** The assessment strategy should be reviewed to confirm that all essential elements (discussed in Section 2, para 12) are present, and that the strategy is consistent with the concepts described in section 2 of this handbook. As a minimum, the strategy should call for a practical summative test for each Training Objective, supplemented as necessary by summative tests (practical or theory) to measure important enabling performances. There should also be policy for comprehensive formative testing.

5.5 **Assemble test records.** The requirement for well-maintained test records is an important element of the assessment strategy. The following records are essential to test evaluation:

1. A summary of formative test results. This serves as evidence that a system of formative assessment is in place.
2. A summary of summative test results, showing a frequency count of student test performance. An example of a useful format is shown at Figure 9. A summary such as this can help to identify objectives, which need further scrutiny. The problem, if there is one, may extend beyond the realm of testing. For example, the records in Figure 9 might focus attention on Enabling Objective 2.2, which seems to have a high failure rate. Further investigation might reveal any of the following:
   1. The test lacks validity.
   2. The instruction has not matched the test.
   3. Better instructional method/media could be applied, e.g., more practise or more formative testing beforehand.
   4. The objective needs to be clarified.
   5. This is a valid test of a difficult objective!

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-38 |

UNCONTROLLED COPY

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TO or EO | Course No \_\_\_\_\_ | | | Course No \_\_\_\_\_ | | |
|  |  |  |  |  |  |
|  | Passed | Passed | Not | Passed | Passed | Not |
|  | 1st try | re-sit | passed | 1st try | re-sit | passed |
| TO1– | 19 | 1 |  | 18 | 1 | 1 |
| Set up antenna |  |  |  |  |  |  |
| EO 2.2 – | 12 | 6 | 2 | 11 | 7 | 1 |
| Establish comms |  |  |  |  |  |  |
| TO2– | 17 | 1 |  | 16 | 2 |  |
| Operate the radio |  |  |  |  |  |  |

**Figure 9: Sample summary of (summative) test results.**

**STEP 2 – EVALUATE PRACTICAL TESTS**

5.6 In evaluating each practical test for validity, the following procedure should be followed (see also Section 3):

1. **Determine reliability:**
   1. Does the test make maximum use of checklist items, while including rating scale and anecdotal record items only where necessary?
   2. Are items on the test sheet brief, unambiguous and structured consistently? (Scorers should be consulted for comment).
   3. Does the test include clear instructions for the student, the administrator and the scorer? (Students, administrators and scorers should be consulted).
2. **Determine relevance:**
   1. Does the test situation match the performance, conditions and standard stated in the training or enabling objective?
   2. Does the test check ***only*** those processes and/or products essential to successful performance?
3. **Determine range:**
   1. Does the test check ***all*** the processes and/or products essential to successful performance (by sampling if necessary)?

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-39 |

UNCONTROLLED COPY

1. Are there any indications that sampling, if used, has been or could be compromised

**STEP 3 – EVALUATE THEORY TESTS**

5.7 **Select appropriate techniques.** The methods appropriate for evaluating theory tests differ between norm-referenced and criterion-referenced tests. Norm-referenced theory tests, which would (by definition) produce a normal distribution of results, can be evaluated by a number of statistical techniques. These techniques are not applicable to criterion-referenced tests.

5.8 **Evaluate norm-referenced theory tests.** Before undertaking the somewhat painstaking process described below, the evaluator should first confirm that the norm-referenced theory test is being used for an appropriate purpose (see Figure 5 in section 2, page 8).

1. **Determine reliability.** A useful measure of reliability is the ***correlation coefficient***. A summary of how this is accomplished can be found at Annex B to this section.
2. **Determine relevance.** First, check to confirm that the test has been

constructed in accordance with a test plan (see Figure 7, section 3, page 5). If so, individual test items should be analysed to determine the ***facility value*** (how easy or difficult is the item?) and the ***index of discrimination*** (how well does the item discriminate between students with high and low achievement overall?). Multi-choice items can be further analysed to determine the effectiveness of distracters. Details of how to accomplish this can be found at Annex C to this Section.

1. **Determine range:**
   1. Does the test check all the facts, concepts and principles indicated on the test plan (by sampling if necessary)?
   2. Are there any indications that sampling, if used, has been or could be compromised?

5.9 **Evaluate criterion-referenced theory tests.**A criterion-referenced theory test is designed to measure achievement of a theory-based enabling objective **. *Following effective instruction*** most, if not all, of the students should pass most, if not all, of theitems. For this reason, calculation of a correlation coefficient for the test is not likely to be meaningful. Similarly, the item analysis described in sub-para 8b above is also inappropriate. The validity of such tests can be estimated by reference to the enabling objective and to the test plan, as follows:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-40 |

UNCONTROLLED COPY

1. **Determine reliability:**
   1. Within the constraints of the test plan, are objective items (matching, multi-choice) used and subjective items (short answer, extended answer) avoided?
   2. Are test items worded unambiguously? (Students should be consulted).
   3. Are instructions for students, administrators and scorers clear and comprehensive? (Students, administrators and scorer should be consulted).
2. **Determine relevance:**
   1. Do test items require only performances, which are consistent with the enabling objective?
   2. Do test items measure the levels of learning identified in the test plan?
   3. Is the pass mark high enough to demonstrate achievement of the objective?
3. **Determine range:**
   1. Does the test check *all* the facts, concepts and principles indicated by the test plan (by sampling if necessary)?
   2. Are there any indications that sampling, if used, has been or could be compromised?

**STEP 4 – INTERPRET TEST RESULTS**

5.10 **Uses.** A test must be evaluated, and found to be valid, before the results of that test can be properly interpreted. Once validity has been established, test results can be used for all of the following:

1. To check, on the effectiveness of course design and instruction. Where many students miss a test item, an oversight or error in instruction might be signified. Where such difficulties are repeated over several courses, a need for a change in course design (content, sequence, methods or media) is often indicated.
2. To help, build a more flexible testing system. Items from valid tests can be added to an ***item bank***– a pool or items of proven effectiveness. Items can be drawn from the pool to meet the requirements of a test plan, thus enhancing the random nature of the test. Alternatively, a valid test can serve as a model for construction of
3. test of ***equivalent form***– a set of similar items, which can form the basis of a test, which will parallel the established test. Either of these approaches can be helpful in developing tests to be used for re-sits.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-41 |

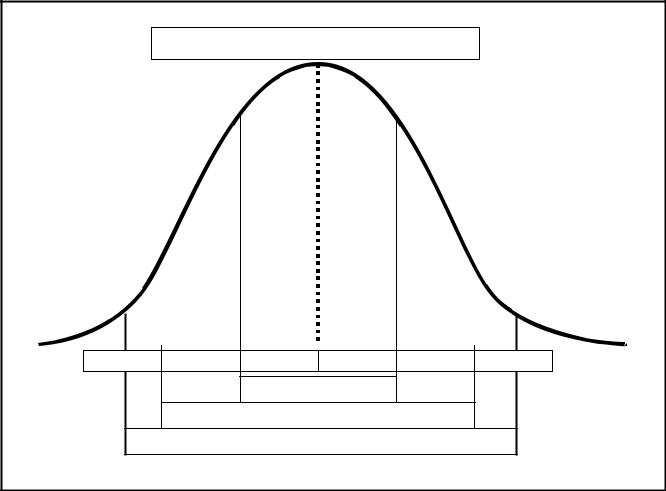
UNCONTROLLED COPY

1. To guide decisions about students. A valid test can help discriminate between those students who have achieved the objective, and those who have not (criterion-referenced test); or can help determine the standing of students in relation to one another (norm-referenced test).

5.11 **Scores in context.** A test score such as “63 points” bears little meaning in isolation. A frame of reference needs to be established which will enable analysis. The nature of that frame depends upon the nature of the test: norm-referenced or criterion-referenced.

5.12 **Interpret Norm-referenced Tests.** The results of norm-referenced tests should be considered, *and reported,* in comparison to scores for the student group. This can be done by using the following, ***descriptive statistics***:

1. **Central tendency** – a measure of the extent to which the scores cluster together. Typically, this is measured by the arithmetic, ***mean*** (average) score. Though the ***mode*** (most frequent score) or ***median*** (middle of the ordered scores) might also be used. This subject is further discussed at Annex D to this section.
2. **Scatter** – a measure of the extent, to which scores are spread apart. This might be recorded as a simple ***range*** of scores (high and low values for the group), but a more useful measure is the ***standard deviation***. The calculation of standard deviation is explained is at Annex E to this Section. Standard deviation can also be derived from many hand calculators, or from almost any computer statistics software. The standard deviation value provides a measure of scatter, which can be directly related to the mean 68% (a constant) of the group scores fall into the interval defined by the mean value ± the standard deviation value. A constant 95% of the scores fall into the interval defined by two standard deviations about the mean as shown in figure 10.



Mean score = 0 SD Units

-3SD -2SD -1SD 1SD 2SD 3SD

68%

95%

99%

**Figure 10: Spread of Scores about the Mean**

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-42 |

UNCONTROLLED COPY

Example interpretation of standard deviation:

Scores for a group of 50 students have a calculated mean (**x** ) of 45 points, and a calculated standard deviation (SD) of 15.

1. ±SD = 45±15 = 60, 30

Therefore, 34 students (68% of 50) scored between 30 and 60 points (inclusive).

1. ±2SD = 45±30 = 75, 15

Therefore, 48 students (95% of 50) scored between 15 and 75 points (inclusive).

A student who scored 63 on this test would be in the top 16% of the group (by definition, 32% scored outside the interval defined by the first standard deviation [30 to 60 points], 16% scored above it and 16% below).

Note: The scores used throughout this example are ***raw scores*** – taken directly from the test papers – they need not be convert to percentages to be meaningful.

1. **Score standardisation**. Calculation of central tendency and scatter is independent of the scale of marks used for the test. Since different tests may have different scales for raw scores, direct comparison of test results can be misleading (for example, 43/50 vs 43/100). A simple conversion of scores to a percentage may not be helpful, because in norm-referenced testing, the relative standing of a student score is best viewed in relation to the distribution of all student scores. A common scale, which takes the overall distribution of scores into account, is that of standard scores (usually called z scores or t scores). The calculation of ***standard scores*** is detailed at Annex E to this section.
2. **Combining results**. Once test scores are standardised, it is tempting to weight, add or average such scores, in the hopes of producing a total or overall mark. Such an approach is almost always ill conceived, and ***should be avoided****.* Military training and testing are based on achievement of discrete objectives, and the combination of scores from different objectives is usually not more meaningful than adding “apples and oranges”. Standard scores should be used for comparison of scores only.
3. When distributions are badly skewed and z scores suspect as a method of scaling, a better method of assessing an individual’s relative attainment is to use Stanines. Stanine, which is short for Standard Nine, is a numbered convention which represents the conversion of raw scores data onto a Normal Distribution profile. The

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-43 |

UNCONTROLLED COPY

convention is based on the ranking of the raw data within specified percentage frequency intervals. The percentage frequency intervals are represented by the Stanine Band shown in the table below, where the highest performer is in band 9 and the lowest in band 1. This forces the distribution of scores found (the empirical distribution) whatever its shape, into a normal distribution with a mean of 5 and a SD of 2. A Stanine grade of 5 is thus equivalent to a Z score of 0; 7 is equivalent to a z score of +1: 2 is equivalent to a z score of –1.5 etc. This is further discussed at Annex F to this section along with some other approaches to **Grading**.

STANINE BAND FREQUENCY DISTRIBUTION

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Band Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Score | 4% | 7% | 12% | 17% | 20% | 17% | 12% | 7% | 4% |
| Frequency |  |  |  |  |  |  |  |  |  |
| Distribution |  |  |  |  |  |  |  |  |  |

5.13 **Interpret Criterion-referenced Tests.** Scores for valid criterion- referenced tests can be analysed in a straightforward manner. The main (and arguably sole) concern is whether the student passed the test, thereby demonstrating achievement of the objective. The use of the descriptive statistics discussed in para 12 is invalid. In some cases, there may be justification for noting achievement of an additional, higher criterion, which is taken to demonstrate exceptional ability. Where such higher standards are used, they should be carefully applied as follows:

1. The higher standard should in no way affect the pass standard.
2. The higher standard must, like the pass standard, be a fixed measure of performance, set before the test is conducted. It should also be selected such that it reflects performance, which would be exceptional on the job.
3. Achievement of the higher standard should be independent of achievement of other students in the group. All, some, or none of the students should, in theory, be able to achieve the higher standard.

**SECTION SUMMARY**

5.14 Within the context of the assessment strategy, all tests should be periodically evaluated. This evaluation is concerned first with establishing that tests are valid, and second, that test results, are properly interpreted.

5.15 Test scores must be viewed in context. For norm-referenced tests, this context can be defined by the use of descriptive statistics. For criterion-referenced tests, the context is set by the standard fixed within each test. A shorthand guide for the statistical tools covered in the section can be found at Annex G.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-44 |

UNCONTROLLED COPY

**ANNEX A TO**

**SECTION 5**

**EVALUATING ASSESSMENTS CHECKSHEET**

When evaluating assessments it is useful to have a checklist showing the criteria against which the evaluation can be made. Below is an example checklist that you may find useful.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **EVALUATING AN ASSESSMENT STRATEGY** | | **Yes** | **No** | **Comments** |
| 1 | Is there an entry testing policy? | |  |  | If no move |
|  |  |  |  |  | to 4 |
| 2 | Is the content of the entry standard clearly | |  |  |  |
|  | defined? |  |  |  |  |
| 3 | If a student fails the entry test is the RTU or At | |  |  |  |
|  | Risk decision clearly stated? |  |  |  |  |
| 4 | Does it have a concept for formative testing? | |  |  | If no move |
|  |  |  |  |  | to 7 |
| 5 | Is the responsibility for the conduct of formative | |  |  |  |
|  | tests clearly defined? |  |  |  |  |
| 6 | If a student fails any formative tests is there a | |  |  |  |
|  | policy for re-training? |  |  |  |  |
| 7 | Does it have a policy for summative testing? | |  |  | If no move |
|  |  |  |  |  | to 13 |
| 8 | Are all TOs summatively tested? | |  |  |  |
| 9 | Are practical subjects tested by practical | |  |  |  |
|  | assessments? |  |  |  |  |
| 10a | Are theory tests used as supporting evidence of | |  |  |  |
|  | underpinning knowledge? |  |  |  |  |
| 10b | Are theory tests used as supporting evidence of | |  |  |  |
|  | for critical areas? |  |  |  |  |
| 11 | Are summative tests criterion – referenced? | |  |  |  |
| 12a | If a student fails a summative test | |  |  |  |
|  | is there an opportunity for re-training? | |  |  |  |
| 12b | If a student fails a summative test | |  |  |  |
|  | is there an opportunity for re-testing? | |  |  |  |
| 13 | Is there a policy for the constitution of a SRB? | |  |  |  |
| 14 | Does the course have a clear policy for | |  |  |  |
|  | determining course pass or fail? | |  |  |  |
| 15 | Does the course have a clearly stated grading | |  |  |  |
|  | policy? |  |  |  |  |
| 16 | Does it have a policy for maintenance of test | |  |  |  |
|  | records? |  |  |  |  |
| 17 | Is there evidence of a moderation policy? | |  |  |  |
| 18 | Is there evidence of a standardisation policy? | |  |  |  |
| 19 | Is there an appeals procedure? | |  |  |  |
| 20 | Is there an indication of the resources needed for | |  |  |  |
| Issue V1.0 Aug 07 | | 2-45 |  |  |  |

UNCONTROLLED COPY

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | assessment? | |  |  |  |
| 21 | Is it a valid assessment strategy? | |  |  |  |
| 22 | Will it produce reliable assessment? | |  |  |  |
|  |  |  |  |  |  |
|  | **EVALUATING A TEST SPECIFICATION** |  | **Yes** | **No** | **Comments** |
| 23 | Are all formative and summative tests listed? |  |  |  |  |
| 24 | Do the tests cover the range of the course |  |  |  |  |
|  | content? |  |  |  |  |
| 25 | Are all critical elements of the course tested? |  |  |  |  |
|  | **For each test does it state:** |  |  |  |  |
| 26 | Format of test – Practical/theory? |  |  |  |  |
| 27 | Type of test - Formative / Summative? |  |  |  |  |
| 28 | Time allowed? |  |  |  |  |
| 29 | Content – TOs, EOs being assessed? |  |  |  |  |
| 30 | Reference details? |  |  |  |  |
| 31 | Marking details? |  |  |  |  |
| 32 | Pass/Fail criteria? |  |  |  |  |
| 33 | Consequence of failure? |  |  |  |  |
| 35 | Is each test a valid assessment of content of the |  |  |  |  |
|  | TOs? |  |  |  |  |
| 36 | Will each test produce reliable assessment ? |  |  |  |  |
|  |  |  |  |  |  |
|  | **EVALUATING A PRACTICAL TEST** |  | **Yes** | **No** | **Comments** |
|  | **EVALUATING SCORERS SHEET** |  |  |  |  |
| 37 | Does it make use of checklists? |  |  |  | If no move |
|  |  |  |  |  | to 45 |
| 38 | Does the checklist: state each item by referring to |  |  |  |  |
|  | a single, observable performance? |  |  |  |  |
| 39 | Does the checklist: describe in positive actions |  |  |  |  |
|  | what the student must do to achieve a |  |  |  |  |
|  | satisfactory rating? |  |  |  |  |
| 40 | Does the checklist: structure all items |  |  |  |  |
|  | consistently? |  |  |  |  |
| 41 | Does the checklist: clearly show critical areas? |  |  |  |  |
| 42 | Is it clear whether the order or sequence of the |  |  |  |  |
|  | tasks is significant? |  |  |  |  |
| 42 | If yes Can the score record whether the correct |  |  |  |  |
| a | sequence was followed? |  |  |  |  |
| 43 | Does the checklist: keep each item descriptor |  |  |  |  |
|  | short? |  |  |  |  |
| 44 | Does the checklist: allow the marker to make |  |  |  |  |
|  | comments against each item? |  |  |  |  |
| 45 | Does the checklist: allow the marker to complete |  |  |  |  |
|  | the checklist easily? |  |  |  |  |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-46 |

UNCONTROLLED COPY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 46 | Does it have a rating scale? |  |  | If no move |
|  |  |  |  | to 53 |
| 47 | Is it clear if it is rating process or product? |  |  |  |
| 48 | Is the scale used logical? |  |  |  |
| 49 | Does the scale use between 3 and 7 points? |  |  |  |
| 50 | Does each scale rate one single dimension of the |  |  |  |
|  | performance? |  |  |  |
| 51 | Does it give clear descriptions of how to score? |  |  |  |
| 52 | Are items consistently structured? |  |  |  |
| 53 | Does it have anecdotal records? |  |  | If no move |
|  |  |  |  | to 57 |
| 54 | Is it clear what is being assessed (behavior or |  |  |  |
|  | attitudes)? |  |  |  |
| 55 | Is it clear what the scorer has to write down? |  |  |  |
| 56 | Is it clear of the place and weight of this |  |  |  |
|  | assessment in the overall pass/fail criteria? |  |  |  |
| 57 | Does it have general instructions for scorer? |  |  | If no move |
|  |  |  |  | to 68 |
| 58 | Is there guidance for observation of student? |  |  |  |
| 59 | Does it clearly state how/when cues for student? |  |  |  |
|  | action may be given? |  |  |  |
| 60 | Does it clearly state whether the scorer can |  |  |  |
|  | assist the student? |  |  |  |
| 61 | Does it clearly state the circumstances when test |  |  |  |
|  | may be interrupted (safety?)? |  |  |  |
| 62 | Does it clearly state how to complete test sheet? |  |  |  |
| 63 | Is there space for the student details to be |  |  |  |
|  | recorded? |  |  |  |
| 64 | Is there space for the scorers details to be |  |  |  |
|  | recorded? |  |  |  |
| 65 | Is there space for the assessment result to be |  |  |  |
|  | recorded? |  |  |  |
| 66 | Are the pass/ fail criteria clearly stated? |  |  |  |
| 67 | Does it clearly state action to take upon? |  |  |  |
|  | completion of test? |  |  |  |
| 68 | **DOES IT HAVE STUDENT TEST** |  |  | If no move |
|  | **INSTRUCTIONS?** |  |  | to 75 |
| 69 | Does it include the purpose of the test? |  |  |  |
| 70 | Does it include the outline content or format of |  |  |  |
|  | the tests? |  |  |  |
| 71 | Does it state the time allowed? |  |  |  |
| 72 | Does it state what resources are available and |  |  |  |
|  | denied? |  |  |  |
| 73 | Does it state the standard of achievement |  |  |  |
|  | expected? |  |  |  |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-47 |

UNCONTROLLED COPY

1. Does it state what should be done on completion of the test?

|  |  |  |
| --- | --- | --- |
| 75 **DOES IT HAVE ADMIN INSTRUCTIONS?** | | If no move |
|  |  | to 85 |

1. Does it include the purpose of the test?
2. Does it include the outline content or format of the test?
3. Does it state the time allowed?
4. Does it state how to set up test: (number of administrators, scorers, reference material, tools, equipment, supervision, safety hygiene considerations. rubbish disposal)?
5. Does it state how/when to present instructions to students?
6. Does it state how/when to cue for student action?
7. Does it state how much assistance authorised to student?
8. Does it state how to control student and third party movement in test site?
9. Does it state the action upon to be taken on student completion of test?
10. **IS IT A VALID ASSESSMENT OF CONTENT OF THE TOs?**
11. **WILL THIS TEST PRODUCE RELIABLE ASSESSMENT?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **EVALUATING A THEORY TEST** | **Yes** |  | **No** | **Comments** |  |
| 87 |  | **DOES THE TEST PAPER HAVE** |  |  |  | If no move |  |
|  |  | **INSTRUCTIONS FOR THE STUDENT** |  |  |  | to 94 |  |

1. Does it include the purpose of the test?
2. Does it include the outline content or sections of the test?
3. Does it state the time allowed?
4. Does it state what resources are available or denied?
5. Does it state the standard of achievement expected?
6. Does it state what should be done on completion of the test?
7. Does it have somewhere for the student to put their name?
8. Is the paper clearly laid out?
9. Is the paper divided into sections?
10. Is there enough space for answers?

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-48 |

UNCONTROLLED COPY

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 98 |  | Does the test paper have instructions for the | | |  |  |  |
|  |  | student to complete each item? | | |  |  |  |
| 99 |  | Are they told what to do to correct an answer? | | |  |  |  |
| 100 |  | Are the marks available for each item stated? | | |  |  |  |
| 101 |  | Are selection items used for simple knowledge | | |  |  |  |
|  |  | questions? | | |  |  |  |
| 102 |  | Are matching questions well constructed? | | |  |  |  |
| 103 |  | Are multi-choice questions well constructed? | | |  |  |  |
| 104 |  | Are true/false items well constructed? | | |  |  |  |
| 105 |  | Are short answer questions well constructed? | | |  |  |  |
| 106 | Are word/sentence completion questions well | | | |  |  |  |
|  |  | constructed? | | |  |  |  |
| 107 |  | Are extended answer/ essay questions only | | |  |  |  |
|  |  | used for the higher domains? | | |  |  |  |
|  |  |  | | |  |  |  |
| 108 |  | **DOES IT HAVE INSTRUCTIONS FOR THE** | | |  |  | If no move |
|  |  | **TEST ADMINSTRATOR?** |  |  |  |  | to 117 |
| 109 |  | Does it include the purpose of the test? | | |  |  |  |
| 110 |  | Does it include the outline content or format of | | |  |  |  |
|  |  | the test – i.e. no of pages? | | |  |  |  |
| 111 | Does it state the time allowed? | | | |  |  |  |
| 112 |  | Does it state how to set up test: (reference | | |  |  |  |
|  |  | material, tools, equipment, supervision, rubbish | | |  |  |  |
|  |  | disposal)? | | |  |  |  |
| 113 |  | Does it state how/when to present instructions to | | |  |  |  |
|  |  | students? | | |  |  |  |
| 114 |  | Does it state how much assistance is authorised | | |  |  |  |
|  |  | to student? | | |  |  |  |
| 115 |  | Does it state how to control student movement? | | |  |  |  |
| 116 | Does it state the action upon to be taken on | | | |  |  |  |
|  |  | student completion of test? | | |  |  |  |
| 117 |  | **DOES IT HAVE INSTRUCTIONS FOR THE** | | |  |  | If no move |
|  |  | **MARKER?** | | |  |  | to 124 |
| 118 |  | Does it state how marks are to be shown on the | | |  |  |  |
|  |  | test paper and/or space for markers comments? | | |  |  |  |
| 119 |  | Does it contain a test key that lists correct | | |  |  |  |
|  |  | answers? | | |  |  |  |
| 120 |  | For supply questions are the crucial | | |  |  |  |
|  |  | words/elements indicated and the allocation of | | |  |  |  |
|  |  | marks clear? | | |  |  |  |
| 121 |  | Does it have a method for correct answers to be | | |  |  |  |
|  |  | indicated on the test paper for feedback | | |  |  |  |
|  |  | purposes? | | |  |  |  |
| 122 | Does it state where and how the final mark is to | | | |  |  |  |
|  |  | be recorded? | | |  |  |  |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-49 |

UNCONTROLLED COPY

1. Does it clearly explain the pass/fail criteria?
2. **IS IT A VALID ASSESSMENT OF CONTENT OF THE TOs?**
3. **WILL EACH TEST PRODUCE RELIABILE ASSESSMENT?**

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-50 |

UNCONTROLLED COPY

**ANNEX B TO**

**SECTION 5**

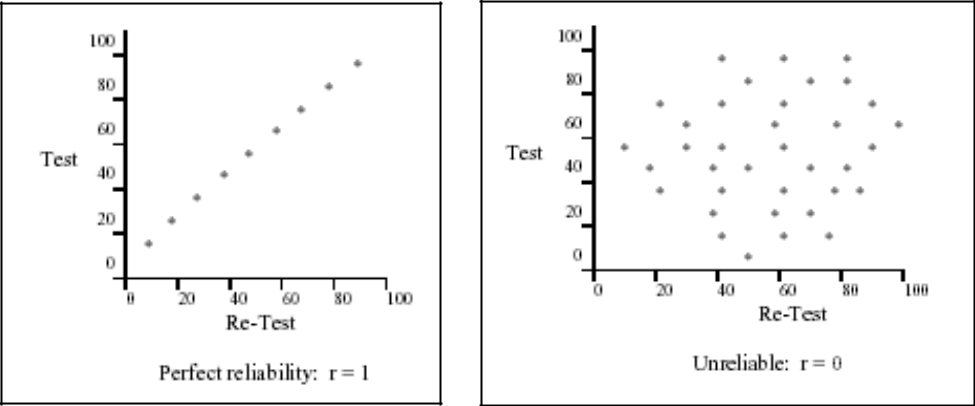
DETERMINING RELIABLITY

1. Test Designers are responsible for ensuring the reliability of assessment processes and the test itself. Reliability of a test is the consistency with which a test measures whatever it is supposed to measure. The reliability of any test used in training should be established. If the

measuring instrument itself is unreliable then clearly no inferences should be drawn from the results. In a training situation test reliability is best measured in one of two ways.

* 1. **Test - Re-test**. The test is given twice to the same trainees and their results compared.
  2. **Split-half**. The test is given and the results on one half of the test are compared with the results on the other half.

1. **Correlation Coefficient**. The reliability of a test is measured by the correlation coefficient (r). This can take values between -1 and +1. When using it for reliability only the positive area is of concern: +1 means perfect reliability; 0 means no reliability. In a test with high reliability one would expect Trainee A who scored a high mark on the test today, or on half the test, to score a high mark on the test next week, or the other half of the test. One should aim at a reliability of +0.7 or better. If the coefficient falls below this the test is not sufficiently reliable. There are a number of methods of calculating a correlation coefficient. Explanation of mathematical methods such as the Pearson Correlation Coefficient and shorthand guide to the calculations used can be found at Annex E to Section 5. An alternative method is to use a Scattergram.
2. **Scattergrams**. Results from the test - re-test or split half method may be plotted on a graph and visual inspection may indicate how reliable the test is:



1. **Pearson’s Quadrant Method**. If a scattergram has been drawn it is easy to apply Pearson’s Quadrant method to gain an estimate of the reliability index for a test.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-51 |

UNCONTROLLED COPY

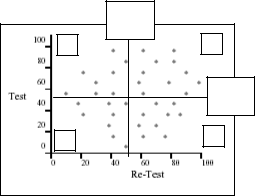
1. The method involves the following steps.
   1. Calculate the average score in test 1.
   2. Draw a line on the scattergram axis that corresponds to the average score for that test.
   3. Repeat for the second test.

Your scattergram has now been divided into 4 quadrants

1. Label the top right Quadrant – A, the top Left B, the bottom right C and the Bottom left D.
2. Count the number of points in each quadrant.
3. (Multiply the number in A by the number in D) and then divide that by the number in (B times the number in C.)

This will give you a value called Q.

1. Look up the corresponding value of r using the table called Pearson’s estimates of reliability (r) for various values of ad/bc (Q).



1. Example

B

D

Ave

Test 2

A

Ave

Test 1

C

Number of points in A = 11

Number of points in B = 8

Number of points in C = 9

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-52 |

UNCONTROLLED COPY

|  |  |
| --- | --- |
| Number of points in D = 8 |  |
| A x D = 88 |  |
| B x C = 72 therefore | Q=1.2 |

From table r = 0.07 Test show very little reliability – almost 0.

**PEARSON'S ESTIMATES OF RELIABILITY (r) FOR VARIOUS VALUES OF ad/bc (Q)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **NO CORRELATION** | | | |  |  |  |  |  |  |  |  |  |  |
| ad/bc | | |  |  |  | ad/bc |  |  |  |  | ad/bc |  |  |  |
|  | |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Values Between | | |  | r = |  | Values Between | |  | r = |  | Values Between | |  | r = |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.00 |  | 1.00 |  | 0.00 |  | 2.49 | 2.55 |  | 0.35 |  | 8.50 | 8.90 |  | 0.70 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.01 |  | 1.03 |  | 0.01 |  | 2.56 | 2.63 |  | 0.36 |  | 8.91 | 9.35 |  | 0.71 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.04 |  | 1.06 |  | 0.02 |  | 2.64 | 2.71 |  | 0.37 |  | 9.36 | 9.82 |  | 0.72 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.07 |  | 1.08 |  | 0.03 |  | 2.72 | 2.79 |  | 0.38 |  | 9.83 | 10.33 |  | 0.73 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.09 |  | 1.11 |  | 0.04 |  | 2.8 | 2.87 |  | 0.39 |  | 10.34 | 10.90 |  | 0.74 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.12 |  | 1.14 |  | 0.05 |  | 2.88 | 2.96 |  | 0.40 |  | 10.91 | 11.51 |  | 0.75 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.15 |  | 1.17 |  | 0.06 |  | 2.97 | 3.05 |  | 0.41 |  | 11.52 | 12.16 |  | 0.76 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.18 |  | 1.20 |  | 0.07 |  | 3.06 | 3.14 |  | 0.42 |  | 12.17 | 12.89 |  | 0.77 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.21 |  | 1.23 |  | 0.08 |  | 3.15 | 3.24 |  | 0.43 |  | 12.90 | 13.70 |  | 0.78 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.24 |  | 1.27 |  | 0.09 |  | 3.25 | 3.34 |  | 0.44 |  | 13.71 | 14.58 |  | 0.79 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.28 |  | 1.30 |  | 0.10 |  | 3.35 | 3.45 |  | 0.45 |  | 14.59 | 15.57 |  | 0.80 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.31 |  | 1.33 |  | 0.11 |  | 3.46 | 3.56 |  | 0.46 |  | 15.58 | 16.65 |  | 0.81 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.34 |  | 1.37 |  | 0.12 |  | 3.57 | 3.68 |  | 0.47 |  | 16.66 | 17.88 |  | 0.82 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.38 |  | 1.40 |  | 0.13 |  | 3.69 | 3.8 |  | 0.48 |  | 17.89 | 19.28 |  | 0.83 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.41 |  | 1.44 |  | 0.14 |  | 3.81 | 3.92 |  | 0.49 |  | 19.29 | 20.85 |  | 0.84 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.45 |  | 1.48 |  | 0.15 |  | 3.93 | 4.06 |  | 0.50 |  | 20.86 | 22.68 |  | 0.85 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.49 |  | 1.52 |  | 0.16 |  | 4.07 | 4.2 |  | 0.51 |  | 22.69 | 24.76 |  | 0.86 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.53 |  | 1.56 |  | 0.17 |  | 4.21 | 4.34 |  | 0.52 |  | 24.77 | 27.22 |  | 0.87 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.57 |  | 1.60 |  | 0.18 |  | 4.35 | 4.49 |  | 0.53 |  | 27.23 | 30.09 |  | 0.88 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.61 |  | 1.64 |  | 0.19 |  | 4.5 | 4.66 |  | 0.54 |  | 30.10 | 33.60 |  | 0.89 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.65 |  | 1.69 |  | 0.20 |  | 4.67 | 4.82 |  | 0.55 |  | 33.61 | 37.79 |  | 0.90 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.70 |  | 1.73 |  | 0.21 |  | 4.83 | 4.99 |  | 0.56 |  | 37.80 | 43.06 |  | 0.91 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.74 |  | 1.78 |  | 0.22 |  | 5 | 5.18 |  | 0.57 |  | 43.07 | 49.83 |  | 0.92 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.79 |  | 1.83 |  | 0.23 |  | 5.19 | 5.38 |  | 0.58 |  | 49.84 | 58.70 |  | 0.93 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.84 |  | 1.88 |  | 0.24 |  | 5.39 | 5.59 |  | 0.59 |  | 58.80 | 70.95 |  | 0.94 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.89 |  | 1.93 |  | 0.25 |  | 5.6 | 5.8 |  | 0.60 |  | 70.96 | 89.01 |  | 0.95 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.94 |  | 1.98 |  | 0.26 |  | 5.81 | 6.03 |  | 0.61 |  | 89.02 | 117.54 |  | 0.96 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.99 |  | 2.04 |  | 0.27 |  | 6.04 | 6.28 |  | 0.62 |  | 117.55 | 169.67 |  | 0.97 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.05 |  | 2.10 |  | 0.28 |  | 6.29 | 6.54 |  | 0.63 |  | 169.68 | 293.11 |  | 0.98 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-53 |

UNCONTROLLED COPY

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2.11 | 2.15 |  | 0.29 |  | 6.55 | 6.81 |  | 0.64 |  | 293.12 | 923.97 | 0.99 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.16 | 2.22 |  | 0.30 |  | 6.82 | 7.1 |  | 0.65 |  | 923.98 | Infinity | 1.00 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.23 | 2.29 |  | 0.31 |  | 7.11 | 7.42 |  | 0.66 |  | **PERFECT** |  |  |
|  |  |  |  |  |  |  |  |  |  | **CORRELATION** | |  |
| 2.29 | 2.34 |  | 0.32 |  | 7.43 | 7.75 |  | 0.67 |  |  | **b** | **a** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.35 | 2.41 |  | 0.33 |  | 7.76 | 8.11 |  | 0.68 |  |  | **d** | **c** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
| 2.42 | 2.48 |  | 0.34 |  | 8.12 | 8.49 |  | 0.69 |  | Compute | bc/ad if it is | larger than |
|  |  |  |  |  |  |  |  |  |  | ad/bd in this case r will be | | |
|  |  |  |  |  |  |  |  |  |  | negative i.e. inverse correlation | | |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-54 |

UNCONTROLLED COPY

**ANNEX C TO**

**SECTION 5**

**TEST ITEM ANALYSIS**

**INTRODUCTION**

1. **Background**. Multiple choice testing is an objective form of written test which has the advantages of high reliability and ease of marking. However, writing of multiple choice test items which correctly discriminate the better from the poorer trainee and which tests at the required level of difficulty is not as easy as it may seem. Too often the construction of multiple choice test items provides clues, distracters fail to distract or discriminate inversely. The test designer should therefore pilot multiple-choice tests and conduct an analysis of he results to confirm the validity of the test items.
2. **Scope**. This annex will discuss the purpose of item analysis, procedure for analysis of multiple choice test items, item analysis statistics, improving the effectiveness of distracters and, briefly, the application of item analysis to short answer type questions. Detail on construction of multiple choice test items is provided in Section 4 of this Handbook

**PURPOSE OF TEST ITEM ANALYSIS**

1. Item analysis refers to an analysis of responses made to multiple choice test items. It is a process that should be carried out during the pilot of a test, on a sample of the target population, to determine whether each item is acceptable or not for inclusion in a test for the whole target population. If an item is proved effective then it should be 'banked' for later use.
2. When determining whether a test item is playing its part in contributing to an accurate assessment of ability or performance there are three aspects to be considered:
   1. The facility of the item.
   2. The discriminating power of the item.
   3. The power of the distracters.
3. On each occasion that an item is used it should be examined critically for its facility and discriminating power.
4. **Preparation for Item Analysis**. When carrying out an item analysis on a pilot or normal course multiple-choice test, it is useful to prepare a test item analysis matrix for the test.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-55 |

UNCONTROLLED COPY

**ITEM ANALYSIS PROCEDURE**

1. Prepare an item analysis matrix before the test. If item analysis is a regular occurrence it is worthwhile having matrix blanks already prepared with room for 50 test items and 50 testees. Better still, it could be a TEMPLATE on a computer!
2. Mark the test. If trainees record their responses on an answer sheet then the marking load is eased by the preparation of a marking template indicating the correct responses.
3. After a test, rank scripts in order of total test score with the highest score on top.
4. Enter responses in the appropriate row in the matrix. Take each script in turn starting with the highest overall score and:
   1. Enter name or identifying number.
   2. Look at the response to each test item and record as:
      1. Correct responses - leave blank.
      2. No response - enter 'O'.
      3. Incorrect response - enter incorrect response letter or letters.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | TEST ITEM NUMBER | | | | |  |  |  | **Total** | **Correct** |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** |  | **8** | **9** | **10** | **Error** |  |
|  | **C** | **A** | **B** | **B** | **A** | **D** | **C** |  | **B** | **D** | **C** |  |  |
| **Atkin** |  |  | A |  |  | A |  |  |  |  |  | 2 | 8 |
| **Tar** |  |  | A |  |  | B |  |  |  | B |  | 3 | 7 |
| **Blogg** | A |  | A |  |  | B |  |  |  | B |  | 4 | 6 |
| **Smith** | A |  |  | A |  | C |  |  |  | B | A | 5 | 5 |
| **Jones** | A |  |  | C |  | A |  |  | C | C | B | 6 | 4 |
| **Bell** | A |  |  | D |  | B |  |  | A | C | D | 6 | 4 |
| **Bond** | A |  |  | A |  | C |  |  | D | C | B | 6 | 4 |
| **FV** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **ID** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Okay?** |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. Scan the Results. Some apparent test or training defects may well be identified at this stage. This analysis is not realistic as there are only 10 test items and 7 trainees, but it will serve for demonstration purposes. Note the following:
   1. **Item 1**. Only two have correctly answered the item. Of more interest is why the less able on the test opted for alternative A.
   2. **Item 3**. The more able trainees on the test overall all got this item wrong and went for the same option whereas the less able overall got the item correct. This

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-56 |

UNCONTROLLED COPY

may mean the way the item is worded at present misleads the trainees who appear to know more about the subject. It may well need to be changed.

* 1. **Item 5**. All trainees have got this right. This may well be because of good instruction but the item needs checking to ensure that there is no clue to the right answer in the item.
  2. **Item 6**. All trainees have got this item wrong. Does it indicate a poorly worded item or does it indicate poor instruction?

1. This initial scanning of the results may indicate possible problem areas with the test items themselves or with the training. This can only be checked by investigating the test items and what instruction the trainees received. There will also be items which do not at first sight appear to have problems, but some method is required to establish whether the items are acceptable or not. Also the response rate may change over time. The methods used are the calculation of facility values (FV) and index of discrimination (ID) for each item.
2. An FV and ID are entered on the matrix and a decision taken whether the item is:
   1. Acceptable and can be banked.
   2. Unacceptable and should be rewritten or discarded.

**ITEM ANALYSIS STATISTICS - FACILITY VALUE (FV)**

1. The FV of a test item is a numerical expression of how easy the item proved to be to the group tested as a whole. It is expressed as a value between 0 and 1. An FV of 0 signifies a very difficult item with no correct responses, and 1 indicates a very easy item with all correct responses.
2. Calculation. An FV is calculated by dividing the number of testees that responded correctly to the test item by the total number that sat the test. The value is expressed correct to 2 decimal places:

FV = Number of correct responses

Number of group tested

1. **Example**. A question (item) with 4 alternative responses is posed to 40 trainees. These results are obtained:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Response | A | B | C | D | Total |
| Totals | 20 | 10 | 3 | 7 | 40 |

If B is the correct response then the facility value of the item is:

FV = 10/40

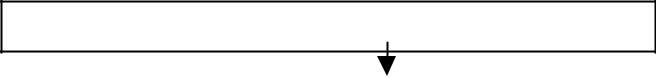
|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-57 |

UNCONTROLLED COPY

* + 0.25

1. Significance. There is no clear answer to the question of what is acceptable as a facility value, as it depends upon the purpose of the test which the item forms a part. The item in the example is, however, of doubtful value since there is a 1-in-4 (0.25) chance of a testee guessing the correct response and this is equal to its facility value. This item cannot be used confidently in a test. If, however, the item had 5 alternative responses, a guessing factor of 0.20 (1-in-5) and a facility value of 0.25, then the item could be used, but with caution.
2. As a general rule and assuming that instruction is effective, achievement tests should consist of items of a high facility value 0.70 to 1.00. Selection tests, however, require items in the medium range of facility 0.30 and 0.70, so as to discriminate between candidates.
3. FV value for items can be entered on the test analysis matrix.
4. If items have low FVs, the following three questions should be asked:

|  |  |
| --- | --- |
| Is the item testing something they should know? | NO Discard item |
|  |  |
| YES |  |



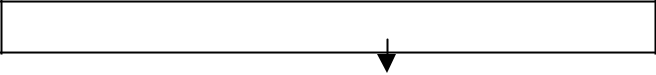
Is the item badly phrased?

NO

YES Rewrite & retest item



Is the standard higher than in TOs & EOs



NO

Examine training on that part of a course for:

* Appropriate methods & media
* Accuracy

YES Rewrite & retest item



**INDEX OF DISCRIMINATION (ID)**

1. The ID of a multiple choice test item is a numerical statement of how well an item discriminates between the better and the poorer testees. It is expressed as a value between +1 and -1 where +1 signifies that all the high scorers on the test responded correctly and all the low scorers responded incorrectly to the item. The reverse, where high scorers are incorrect and low scorers correct, produces a value of -1 for the item.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-58 |

UNCONTROLLED COPY

1. Procedure. These steps outline a simple but effective procedure for obtaining the ID of test items when dealing with small groups (15-50). Four test papers are used to illustrate these steps:
   1. Arrange the test papers in rank order from the highest score to the lowest score. This should already have been recorded on the item analysis matrix.
   2. Select approximately 1/3 of the papers with the highest scores and call this the upper group (13 papers). Select the same number of papers with the lowest scores and call this the lower group. Set aside the middle group of papers (14). It helps to draw lines across the matrix to divide the groups. When dealing with groups of more than 40 trainees, use the results of the top 27% and bottom 27% of the group to calculate the ID of each test item.
   3. For each item subtract the proportion of trainees responding correctly in the lower group from the proportion responding correctly in the upper group.
2. **Example**. A question with 4 alternative responses is posed to 40 trainees. The results shown are obtained, where B is the correct answer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response | A | **B** | C | D |
| Upper Group | 6 | 6 |  | 1 |
| Middle Group | 10 | 2 |  | 2 |
| Lower Group | 4 | 2 | 3 | 4 |
| Totals | 20 | 10 | 3 | 7 |

Total



13



14



13



40



|  |  |  |
| --- | --- | --- |
| Proportion of upper group correct = | | 6/13 |
| Proportion of lower group correct = | | 2/13 |
| Index of Discrimination = | 6/13 -2/13 |  |
| ID = | 4/13 |  |
| ID = | 0.31 |  |

1. Significance. An item should be rejected if the ID is between -1 and 0 since in the former case it is discriminating against the upper group of scorers on the test, and in the latter case it is failing to discriminate between the upper and lower groups of scorers.

**RELATIONSHIP BETWEEN FACILITY VALUE AND DISCRIMINATION**

1. When constructing an objective test, the selection or rejection of an item will depend on whether its FV and ID are within a specified band of values appropriate to the type of test.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-59 |

UNCONTROLLED COPY

1. Items that have a high FV have a low ID, since most of a group will have answered items correctly. The difference between upper and lower groups of scorers will be small.
2. Items that have a low FV may have an ID in the range 0.1 to 0.9.

a. **Example 1**. A question with alternative responses is posed to 30 trainees. The results are as shown in the next table where C is the correct response:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response | A | B | **C** | D |
| Upper Group |  |  | 9 | 1 |
| Middle Group | 1 |  | 8 | 1 |
| Lower Group | 1 | 1 | 8 |  |
| Totals | 2 | 1 | 25 | 2 |

Total



10



10



10



30



|  |  |  |
| --- | --- | --- |
| Number of students who got it correct | =25 |  |
| Number of students on test | =30 |  |
| Facility Value = 25/30 | = 0.83 | High |
| Proportion of upper group correct | = 9/10 |  |
| Proportion of lower group correct | = 8/10 |  |
| Index of Discrimination | = 9/10 – 8/10 | |
| ID | = 1/10 |  |
| ID | = 0.1 | Low |

1. **Example 2**. A question with 4 alternative responses is posed to 30 trainees. The results are as shown where D is the correct response:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response | A | B | C | **D** |
| Upper Group | 1 | 1 |  | 8 |
| Middle Group | 1 | 8 | 1 |  |
| Lower Group | 3 | 6 | 1 |  |
| Totals | 5 | 15 | 2 | 8 |

Total



10



10



10



30



|  |  |  |
| --- | --- | --- |
| Number of students who got it correct | =8 |  |
| Number of students on test | =30 |  |
| Facility Value = 8/30 | = 0.27 | low |
| Proportion of upper group correct | = 8/10 |  |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-60 |

UNCONTROLLED COPY

Proportion of lower group correct

= 0/10

Index of Discrimination

ID

ID

= 8/10 – 8/10

= 8/10

= 0.8 High

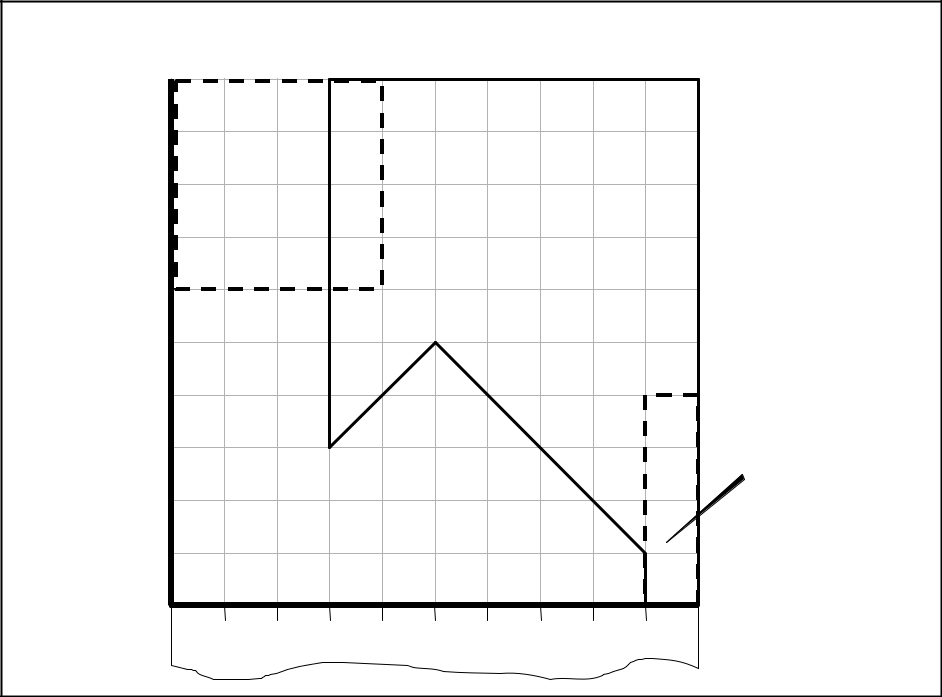
1. This table lists the minimum values of the ID recommended for different facility

bands:

|  |  |
| --- | --- |
| Facility | Index of |
| Value = | Discrimination |
|  | (Minimum) |
| 0.9 | 0.1 |
| 0.8 | 0.2 |
| 0.7 | 0.3 |
| 0.6 | 0.4 |
| 0.5 | 0.5 |
| 0.4 | 0.4 |
| 0.3 | 0.3 |

1. To read values from the table the FV should be rounded to one decimal place 0.8 means between 0.75 and 0.84.
2. If the ID falls below the value given in the table the item is faulty and needs rewriting. After the rewrite it should be re-tested. Items with IDs that are greater than those in the table are good items and should be kept in the test.
3. The table is a useful guide to what items are immediately acceptable and which items require to be re-examined and rewritten.
4. Figure 5-C-1 provides a diagrammatical guide to the decision making required when considering the FV and ID of multiple-choice items.

**Figure 5-C-1: Multiple Choice Items FV/ID Categories**

****

***Multiple Choice Items FV/ID Categories***

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** 1.0 | Possible | |  |  |  |  |  |  |  |  |  |
| 0.9 |  |  |  |  |  |  |  |  |  |
| zone for extreme | | |  | **Accept for** | | | |  |  |  |
| 0.8 |  |  |  |  |
| selection | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | tests |  |  |  | **achievement** | | | | |  |  |
| 0.7 |  |  |  |  |  |  |
|  |  |  |  | **tests** | |  |  |  |  |  |
| 0.6 |  |  |  |  |  |  |  |  | Negative ID values |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | are not shown |
| 0.5 |  |  |  |  |  |  |  |  |  |  | here but would be |
|  |  |  |  |  |  |  |  |  |  | re-examined |
| **EFFECTIVENESS**0.4 **OF DISTRACTERS** | | | | | |  |  |  |  |  |  |
| 0.3 |  |  |  |  |  |  |  |  |  |  | **Critical ?** |
| 0.2 |  |  |  |  |  |  |  |  |  |  |
| **Re-examine** | | | | | |  |  |  |  |  |
| Issue V1.0 Aug 007.1 |  |  |  |  |  |
|  |  |  |  | 2-61 | |  |  |  |  |  |
| 0 |  |  |  |  |  |  |  |  |  |  | **FV** |
| 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |

UNCONTROLLED COPY

1. The distracters in a multiple choice item should be so appealing to the trainees who lack the knowledge called for by the item, that they select distracters in preference to the correct answer.
2. Things that can be done to increase the plausibility and attractiveness of distracters

are:

* 1. Use the common trainee misconceptions, or errors, as distracters.
  2. State the alternatives in the language of the trainee.
  3. Make the distracters similar to the correct answer in both length and complexity of wording.
  4. Use extraneous clues in the distracters, such as stereotyped phrasing and verbal associations with the stem of the item. Do not, however, overuse these clues, Beware of trick questions.
  5. Make the alternatives uniform, but beware of fine discriminations that are educationally insignificant.

1. Effectiveness. The effectiveness of distracters can be determined by comparing the number of trainees in the upper and lower groups who selected each incorrect alternative. A good distracter will attract more trainees from the lower group than from the upper group.
2. **Example**. A question with 4 alternative responses is posed to 30 trainees. The results are shown in this table and the correct answer was B.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response | A | **B** | C | D |
| Upper Group |  | 6 | 3 | 1 |
| Middle Group |  | 4 | 2 | 4 |
| Lower Group |  | 2 | 2 | 6 |
| Totals | 0 | 12 | 8 | 8 |

Total



10



10



10



30



Alternative D is functioning effectively. C is poor, however, since it attracted more testees from the upper than the lower group. A is ineffective since it attracted no one.

**SHORT ANSWER QUESTIONS**

1. An item analysis on the short answer test provides less information than an item analysis in a multiple-choice test but it can still provide useful information. The procedure is similar to a multiple choice item analysis but changes have to be made to the matrix and the calculations of FV and ID. These steps should be followed:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-62 |

UNCONTROLLED COPY

1. Prepare test item matrix.
2. Mark test.
3. Rank scripts.
4. Enter on matrix actual score for each item.
5. Calculate FV/ID or mean mark for each test item.
6. Decide what questions are acceptable and/or indicate problems with training, those items that need rewriting and those that should be discarded.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-63 |

UNCONTROLLED COPY

**ANNEX D TO**

**SECTION 5**

**CALCULATING AVERAGES AND SUMMARISING DATA**

**Arithmetic Mean**

1. Average is what is usually meant as arithmetic mean. It is a measure of central tendency, expressing how scores cluster about a central point. It is calculated by dividing the sum of all the scores in the test by the number of scores in the series or group.
2. The average mark or arithmetic mean of a class of students gives a useful reference point to compare the performance of an individual student against the group as a whole. It can also be used as a crude estimate of the difficulty of the test.
3. An individual score of 70 % is not as meaningful as “top have the class” and neither statement is as meaningful as “Scored 70 %; test range 10 %; class average 60 %”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Calculating the Mean using Frequencies** | | |  |  |  |
| Arithmetic Mean = | Sum of all the scores |  |  | = ∑ | *fx* |
| *x* |
| *n* |
|  | Number of scores in the series | |  |  |

Note: the Greek symbol **Σ** is used to indicate ‘the sum of’ and the *x* symbol is used to indicate the ‘arithmetic mean’ while the letter ‘n’ is used to indicate the number of scores in the series.

1. The table below gives the scores of 40 students on a test that has a maximum score

of 10.

|  |  |  |
| --- | --- | --- |
| **Scores (x)** | **Frequency (f)** | **Frequency multiplied by Score (fx)** |
| 10 | 1 | 10 |
| 9 | 2 | 18 |
| 8 | 3 | 24 |
| 7 | 4 | 28 |
| 6 | 5 | 30 |
| 5 | 8 | 40 |
| 4 | 7 | 28 |
| 3 | 3 | 9 |
| 2 | 3 | 6 |
| 1 | 3 | 3 |
| 0 | 1 | 0 |
| TOTALS | 40 | 196 |
|  | n = 40 | Σ fx = 196 |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-64 |

UNCONTROLLED COPY

1. From the table we can calculate the Arithmetic Mean.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Arithmetic Mean = | | | | |  |  | Sum of all the scores | |  |
|  |  |  |  |  |  | Number of scores in the series | | | |
|  |  | = ∑ | *fx* | = | 196 | | = | 4.9 |  |
|  | *x* |
|  |  | 40 | |  |
|  |  |  | *n* | |  |  |  |

6. Therefore the arithmetic mean or average mark is 4.9

Simple Mean Method

7. It is only necessary to calculate the frequency column of the table when a graph is to be plotted. Other wise it is quicker to add all the scores together and then divide by the number of scores.

*x* = ∑ *nx*

8. Note: We have multiplied and divided scores to obtain the arithmetic mean this implies that we have been using interval data. Strictly speaking this type of data will only be derived from objective tests with many short items and we should only calculate arithmetic means for such tests.

**Median**

1. An alternative form of average is the median or middle score. This is that score which divides the group of all the scores in half.
2. The median is the correct form of average to use on ordinal data.

For example:

* 1. Work out the median of the scores 48, 55, 64, 71, 82, 90 and 91.
  2. We arrange the scores in numerical order and then chose the middle score.
  3. As there are 7 scores the middle score or median is the fourth score.
  4. The median is 71.

1. Another example:
   1. Work out the median of the scores 22, 34, 48, 52, 56, 60, 68 and 74.
   2. We arrange the scores in numerical order and then chose the middle score, as there are 8 scores there is no middle score.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-65 |

UNCONTROLLED COPY

1. We therefore chose the median at a point halfway between the two scores that bracket the middle position.
2. The two middle scores are 52 and 56 and the middle score or median is 54.

**Mode**

1. A mode is the most frequently occurring score. However, if a set of values occur with each frequency there is no mode. If there are two sets of values with the same frequency, the distribution is bi-modal.

Percentage Pass Rate

1. The simplest method of summarising data is to express the results as a percentage pass rate. The percentage pass rate is the number of passes divided by the class total. Continuous monitoring of these pass rates will highlight those courses that require attention.
2. For example: during the formative or summative testing of a course a high failure rate is noticed (and assuming the testing regime is focused correctly) then the tutor or course facilitator can investigate several areas. Among these are:
   1. Student selection to establish that the students do in fact meet the initial trainee specification.
   2. Course material to confirm:
      * Availability.
      * Readability (Flesch Reading Ease or Flesch-Kincaid Grade Level are useful aids here).
      * Pitch (is the level of instruction based on student experience as well as underpinning knowledge levels).
   3. Teaching methods and media.
   4. Course scheduling to confirm there is no physical or mental overload (sports, parties or a similar material taught) prior to any of the lessons leading to the formative test or the formative test itself.
3. The table below gives an example of calculating a group percentage pass rate. The pass mark for the test is 90%.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student** |  | **Result** | | **Student** |  |  | **Result** | |
|  | **Mark** |  | **Passed?** |  |  | **Mark** |  | **Passed?** |
| Barbara | 10 |  | Yes | Adam | 9 |  |  | Yes |
| Debra | 10 |  | Yes | Edward | 9 |  |  | Yes |
| Harriet | 10 |  | Yes | Francis | 9 |  |  | Yes |
| Irene | 10 |  | Yes | James | 9 |  |  | Yes |
| Kelly | 10 |  | Yes | Michael | 9 |  |  | Yes |
| Linda | 10 |  | Yes | Neil | 9 |  |  | Yes |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-66 |

UNCONTROLLED COPY

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pauline | 10 | Yes | Robert | 9 | Yes |
| Sarah | 9 | Yes | Victor | 9 | Yes |
| Tricia | 9 | Yes | Charles | 8 | No |
| Ursula | 9 | Yes | Oscar | 7 | No |
| Wendy | 9 | Yes | Quentin | 4 | No |
| Yvonne | 9 | Yes | George | 3 | No |
| Summary | 24 Students | of which 20 | Passed the test. | |  |

1. The percentage pass rate is therefore 20 passes divided by the class total of 24 and then multiplied by 100 to give a percentage pass rate of 83 %.
2. If only one additional student had failed to pass the test the percentage pass rate would have been 79 %. This slight variability can be due to factors over which we can have no control such as a slight cold in one or more students i.e. blind chance.
3. Therefore percentage pass rates cannot be used in isolation to establish the effectiveness of the course. On a well designed and mature course, the possibility that a unique circumstance has affected the results must be considered before unwarranted and possibly unwanted investigations are made. For the same reason, when piloting a new course great care must be taken to avoid misinterpretation of percentage pass rates.
4. By far the best method, whenever possible, is to collect percentage pass rates from previous courses and examine the data for a distinct trend. This would normally require at least 96 students before such a trend analysis could be carried out with confidence in the result.
5. Although percentage pass rates are easy to calculate and then understand they have the disadvantage of wasting potentially useful information. They fail to show the relationship between the scores and how often (the frequency) each score is attained.

Gain Ratios

1. Raw scores obtained by students at the end of a course of training do not necessarily reflect the effect of the course on the students. A high score at the end of the course may indicate either that the course was effective or that a high score could have been obtained without the course.
2. Wherever possible the gain for each student should be calculated. This implies testing the student before and after each course.
3. To obtain the students’ gain score each student is tested before and after the course, using the same test on each occasion, the change in student score is called the gain score. This can be written as:

GAIN = (POST TEST SCORE) – (PRE TEST SCORE)

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-67 |

UNCONTROLLED COPY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 24. | Unfortunately this gain score depends upon the marking scale used: | | | |
|  |  |  |  |  |
|  |  | **STUDENT A** | **STUDENT B** |
| MARKING SCALE | | 0 to 100 | 0 to 25 |  |
| **PRE TEST** | | 40 | 10 |  |
| **POST TEST** | | 80 | 20 |  |
| **GAIN** |  | 40 | 10 |  |
| 25. | Here both students have doubled their pre test scores yet the gain scores appear to | | | |
| be different. Student A gains 40 and student B gains only 10. | | | | |
| 26. | The more satisfactory method is to compare what the student has gained from the | | | |

course with the maximum gain that they could have made. This gives a ratio that can then be converted into a percentage gain.

27. The equation normally takes the following form:

|  |  |  |
| --- | --- | --- |
| GAIN RATIO = | ACTUAL GAIN | x 100 % |
|  | MAXIMUM POSSIBLE GAIN |  |

Which can be expanded into:

GAIN RATIO = (POST TEST SCORE) – (PRE TEST SCORE) x 100 % (POSSIBLE) – (PRE TEST SCORE)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | **STUDENT A** | **STUDENT B** |
|  | MARKING SCALE |  | 0 to 100 | 0 to 25 |
|  |  | |  |  |
|  | **PRE TEST** | | 40 | 10 |
|  | **POST TEST** | | 80 | 20 |
|  | **GAIN** | | 40 | 10 |

1. Using the table above we can calculate the gain ratio for each student as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Student A gain ratio = |  | 80 - 40 x 100 % | | = |  | 40 x 100% | | = | 66 %. |
|  |  | 100 – 40 |  |  |  | 60 |  |  |  |
| Student B gain ratio = |  | 20 - 10 x 100 % | | = |  | 10 x 100% | | = | 66 %. |
|  |  | 25-10 |  |  |  | 15 |  |  |  |

1. The gain ratio for each student is therefore 66 %.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-68 |

UNCONTROLLED COPY

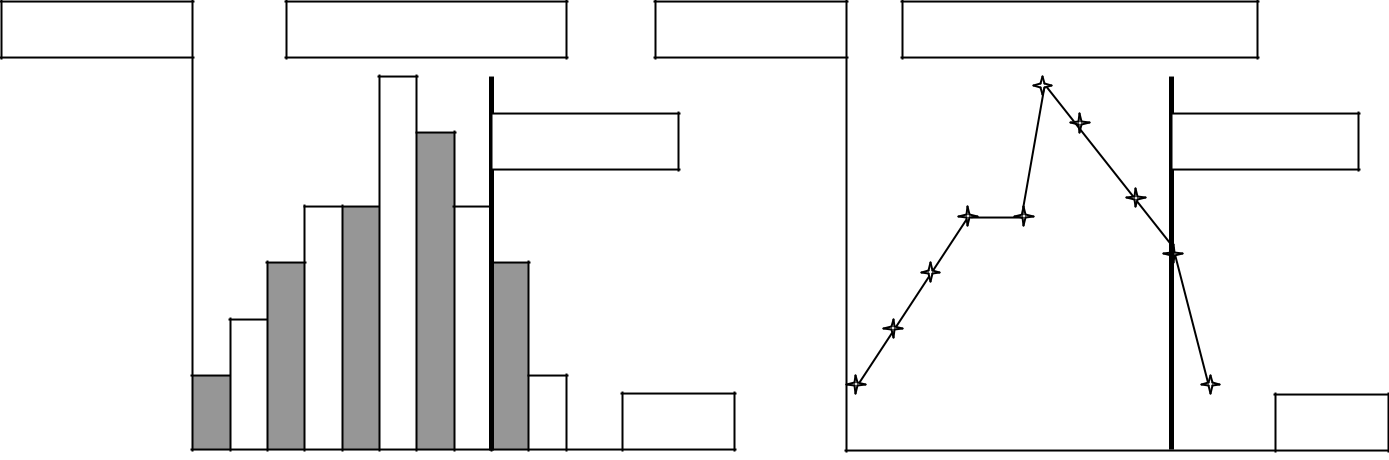
1. The average gain ratio of a group of students is a more accurate measure than the percentage pass rate for the purposes of monitoring training effectiveness. Since it can demonstrate the effects of training and automatically compensates for differing pre – course abilities.
2. Low individual gain ratios detect those who did not benefit from the course because they are well above the trainee specification or unfortunately did not respond in a positive manner to the instruction.
3. If, a large proportion of the group have low gain ratios then there is a need to consider introducing modules at different levels or at different lengths to cover the course material.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-69 |

UNCONTROLLED COPY

**Graphical Representation**

1. There are two simple ways of showing relationship and frequency.
   * The Histogram or bar graph, results are represented as a series of solid bars.
   * The Frequency Polygon, results are individually plotted and then connected by lines.
2. Examples of both are given below.



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Frequency |  |  |  | **Histogram 1** | | | | Frequency | |
| 6 |  |  |  |  |  |  |  |  | 6 |
| 5 |  |  |  |  |  |  | **Pass** |  | 5 |
| 4 |  |  |  |  |  |  |  |  | 4 |
| 3 |  |  |  |  |  |  |  |  | 3 |
| 2 |  |  |  |  |  |  |  |  | 2 |
| 1 |  |  |  |  |  |  |  | Mark | 1 |
|  |  |  |  |  |  |  |  |  |
| 1 | 2 | 3 | 4 | 5 | 6 7 | 8 | 9 10 |  |  |

1. In both graphs:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Frequency Polygon 1** | | | | | | | |
|  |  |  |  |  |  |  |  |  | **Pass** |
|  |  |  |  |  |  |  |  |  | Mark |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

* The horizontal axis along the base represents the number of marks awarded.
* The vertical axis on the left represents the number of trainees who were awarded a particular mark.
  1. The shape of the graph is a valuable piece of information. We can see that the mark with the highest frequency is 6 and the majority of the marks are clustered about this figure. Thus 6 students attained 6 marks; 5 students attained 7 marks and so on. If the pass mark was 90 % (9 out of 10) then only 4 students (the 3 students gained 9 marks plus 1 student who gained 10 marks) passed the test. The heavy dashed line on each graph indicates the pass mark.
  2. This information can still be presented as the result of calculation. The percentage pass rate is 4 / 33, which gives a percentage pass rate of 12%.
  3. The illustrated situation can occur for a number of reasons, two are listed below:

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-70 |

UNCONTROLLED COPY

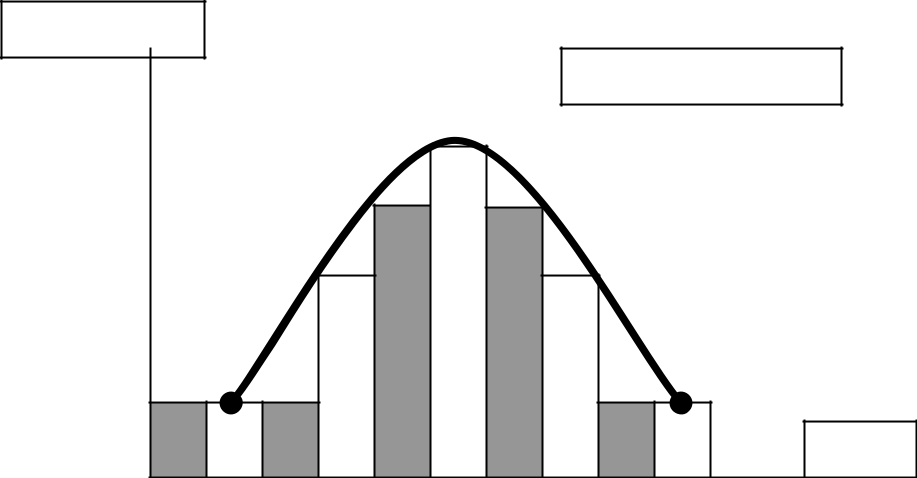
* + The test was too difficult or questions were ambiguous, allowing misinterpretation of what was required for a complete answer.
  + The students have been inadequately trained.

1. Table 3 presents another example. This time the pass mark is 70%. The information is presented in Table 3 and in Histogram 2.

**TABLE 3**

**MARKS AWARDED TO 24 STUDENTS AS THE RESULT OF A TEST**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student** |  | **Result** | | **Student** |  |  | **Result** | |
|  | **Mark** |  | **Passed?** |  |  | **Mark** |  | **Passed?** |
| Barbara | 100 |  | Yes | Ursula | 80 |  |  | Yes |
| Debra | 95 |  | Yes | Wendy | 80 |  |  | Yes |
| Harriet | 90 |  | Yes | Francis | 75 |  |  | Yes |
| Irene | 90 |  | Yes | James | 75 |  |  | Yes |
| Yvonne | 90 |  | Yes | Michael | 75 |  |  | Yes |
| Kelly | 85 |  | Yes | Neil | 75 |  |  | Yes |
| Linda | 85 |  | Yes | Charles | 70 |  |  | Yes |
| Pauline | 85 |  | Yes | Robert | 70 |  |  | Yes |
| Sarah | 85 |  | Yes | Victor | 70 |  |  | Yes |
| Adam | 80 |  | Yes | Oscar | 65 |  |  | No |
| Edward | 80 |  | Yes | Quentin | 60 |  |  | No |
| Tricia | 80 |  | Yes | George | 55 |  |  | No |



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Frequency |  |  |  |  |  |
| 6 |  |  |  |  | **Histogram 2** |
| 5 |  |  |  |  |  |
| 4 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 1 |  |  |  |  |  |
|  |  |  |  |  | Mark |
| 55 | 60 | 65 70 | 75 80 | 85 90 | 95 100 |

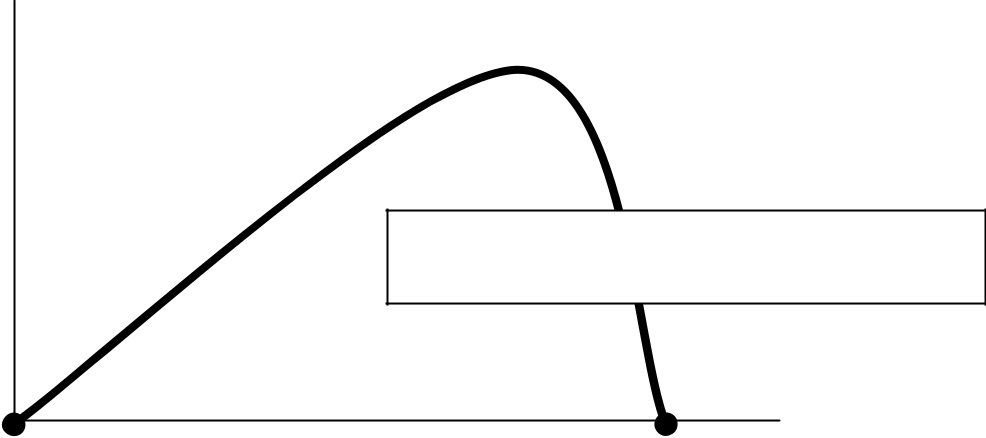
1. We can see that the majority of the scores are clustered about the 80 % point. With only 3 scores below 70 %. The percentage pass rate is therefore 21 / 24, which equals 87.5 %.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-71 |

UNCONTROLLED COPY

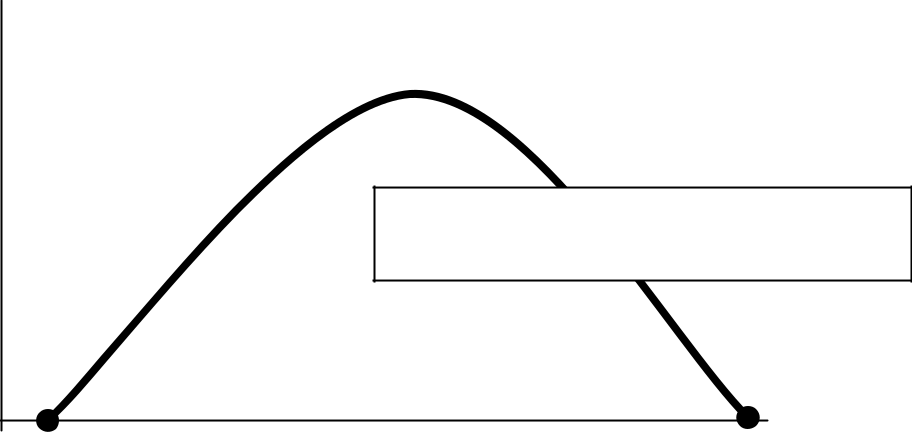
**Distribution Curves**

1. The shape of the graph is approximately ‘Bell shaped’ this is illustrated in Histogram
2. with a bold line. This characteristic shape is commonly called the normal distribution.
3. There are 3 main forms that this shape can take they are illustrated below.



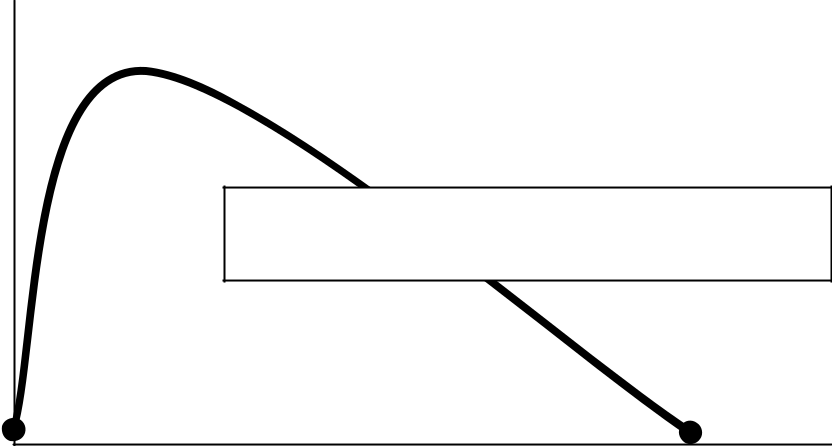
**FREQUENCY POLYGON**

**SKEWED DISTRIBUTION - RIGHT**

****

**FREQUENCY POLYGON**

**NORMAL DISTRIBUTION**

****

**SKEWED DISTRIBUTION - LEFT**

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-72 |

UNCONTROLLED COPY

1. The normal distribution is the basis for most statistics and indeed the majority of grading and selection tests are designed to give a distribution of scores in exactly the same way.
2. However, this normal distribution shape is only likely to occur in tests of ability for which people have not been trained. The reason for this is that people respond to training in different ways based on their aptitude and intelligence. Their personal characteristics are said to skew or distort the normal distribution.
3. Two examples of this distortion are shown on the previous page.
4. The first frequency polygon is skewed to the right. The majority of people gain high scores. Either the test is too easy or the people have been deliberately over trained; on a safety issue perhaps.
5. The second frequency polygon has no skewing and the third is skewed to the left. The majority of people have achieved low scores. Either the test is too difficult or, the test is designed to select only the top few students for further training or advancement.
6. On any occasion that the graphical representation of test results cannot be readily explained then the instruction up to and including the test must be examined on order to ensure quality instruction has taken place.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-73 |

UNCONTROLLED COPY

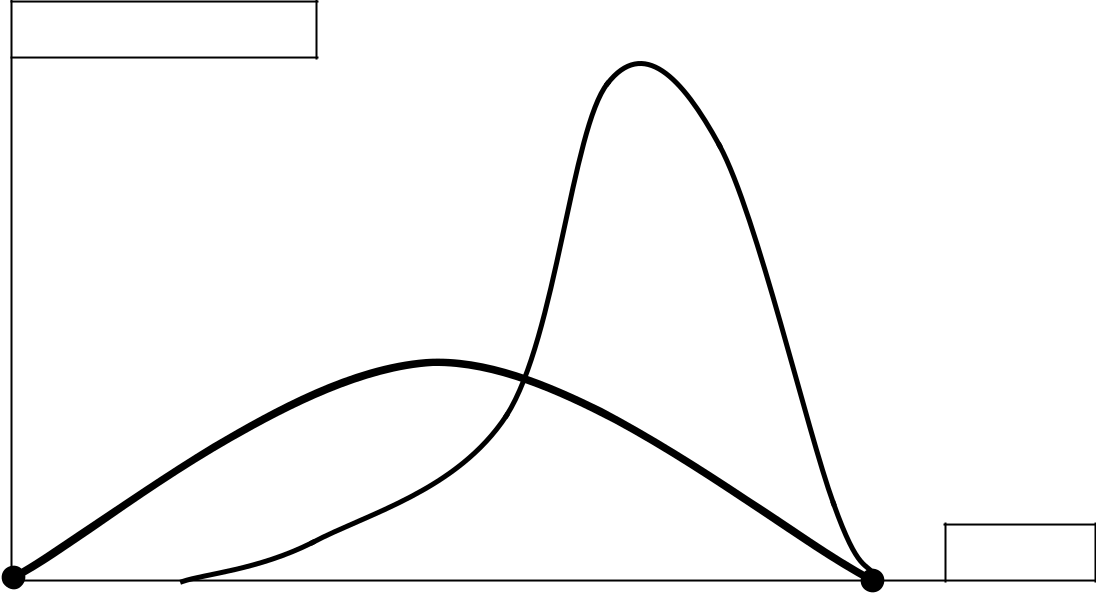
**ANNEX E TO**

**SECTION 5**

**STANDARD DEVIATION, Z SCORES AND T SCORES**

1. Some tests spread students scores out more than others. This is illustrated below using the bold line. The spread or range of scores is 100 and there is very little clustering of scores, that is, the central bulge of the frequency polygon is ill defined and the whole graph is almost flat.
2. Compare this with the dashed line that shows the expected distribution.

**Graph 1**

****

|  |  |
| --- | --- |
| **7** | **FREQUENCY** |
|  |
| **6** |  |
| **5** |  |
| **4** |  |
| **3** |  |
| **2** |  |
| **1** |  |
| **0** | **MARK** |
|  |

**0** **10** **20** **30 40 50 60 70 80 90** **100**

1. Whenever a test of training or learning has been given, some summarising data should be calculated. Three basic measures are required:
   * Range – test discrimination.
   * Average mark – allows a comparison of individuals and the group as a whole.
   * Score distribution – a measure of test variability.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-74 |

UNCONTROLLED COPY

**Range**

1. The range of scores is the difference between the highest and lowest scores in the group. It demonstrates the potential that a particular test has to discriminate between individuals in a class of students.
2. A broad range may discriminate because:
   * The students vary in ability.
   * The training has been uneven.
   * The test is designed to discriminate.
3. A narrow range is often preferable in a summative test at the end of instruction.

**Arithmetic Mean**

1. This is covered along with other types of averages in Annex D

**Distribution**

1. A simple definition of distribution is test variability or the amount that scores scatter about the arithmetic mean. The simplest way of finding the distribution of scores about the mean (scatter) is to plot the scores on a graph.

**FREQUENCY**

**5**

**4**

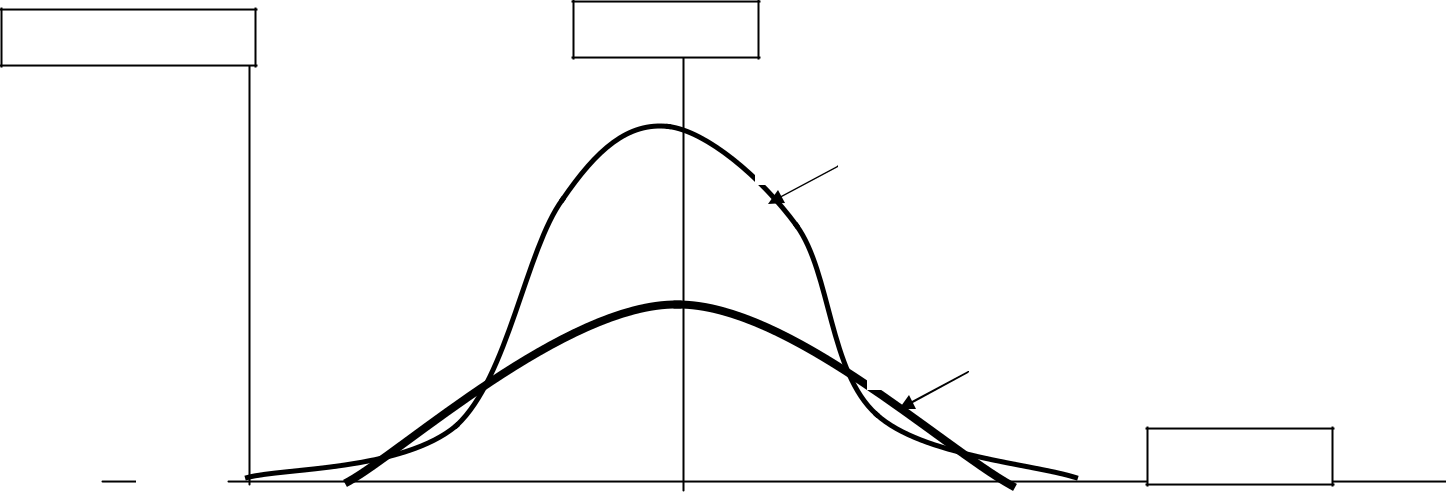
**3**

**2**

**1**

**0**

**Graph 2**

****

**MEAN**

**TEST A**

**TEST B**

**MARK**

**0** **10** **20** **30** **40 50 60 70 80 90 100**

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-75 |

UNCONTROLLED COPY

1. Both sets of scores have the same range of marks and the same mean. They differ only in the amount of scatter or score distribution about the mean.
2. It is more difficult to get a high mark in test A than it is in test B because test A has a smaller distribution about the mean and allows ‘less freedom of movement away from the mean’.

**Standard Deviation**

1. It is possible to calculate the amount of deviation in a set of test results. The measure is the Standard Deviation (SD), calculated from the mean of the group of test results.
2. The steps in the calculation of the SD are as follows:
   1. Calculate the mean. *x*
   2. Subtract the mean from each score to determine each deviation. (x - *x* ).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| c. | Square the deviation for each score. ( x - | | |  | ) 2 |
| *x* |
| d. | Sum the squared deviations. Σ ( x - |  | ) 2 | | |
| *x* |

1. The SD is then given by:

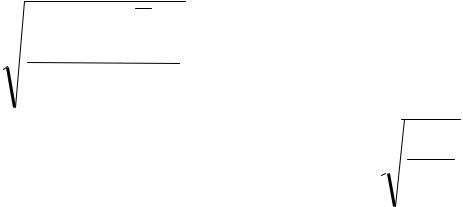
*SD*= ∑( *x* − *x* )2



*n* − 1

1. An example of this is shown below. The number of scores n is 8 and the mean *x* is calculated as 104 / 8 = 13.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Raw Scores** | **Deviation** | | | | **Deviation squared** | | | | | |
|  |  |  | **(x)** | **( x -** |  |  | **)** | **( x -** | | |  |  | **) 2** |
|  |  |  | *x* | *x* |
|  |  |  | **19** | **6** | |  |  | **36** | | | | |  |
|  |  |  | **17** | **4** | |  |  | **16** | | | | |  |
|  |  |  | **15** | **2** | |  |  | **4** | | | |  |  |
|  |  |  | **14** | **1** | |  |  | **1** | | | |  |  |
|  |  |  | **13** | **0** | |  |  | **0** | | | |  |  |
|  |  |  | **11** | **- 2** | |  |  | **4** | | | |  |  |
|  |  |  | **8** | **- 5** | |  |  | **25** | | | | |  |
|  |  |  | **7** | **- 6** | |  |  | **36** | | | | |  |
| Results |  | **Σ x = 104** | |  |  |  |  | **Σ ( x -** |  | **) 2** | | | **= 122** |
|  |  |  |  | *x* |
| *SD* | | = | ∑ ( *x* − *x*) 2 |  |  |  |  |  |  |  |  |  |  |
| *n* − 1 |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | 122 |  |  |  |  |  |  |
|  |  |  |  | *SD*= | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  | 7 |  |  |  |  |  |  |
| Issue V1.0 Aug 07 | | | | 2-76 | | |  |  |  |  |  |  |  |



UNCONTROLLED COPY

|  |  |
| --- | --- |
| Entering the data from the table: | = 4.17 |

1. The Standard Deviation is 4.17
2. The SD is an accurate summary of the amount of scatter in a set of test scores. A large SD means a lot of scatter about the mean. A small SD means very little scatter about the mean. See test B and test A in graph 2 for respective examples.
3. The SD tells us more about the test than the mean and range. It can also:
   * Produce a standardised scale if ability is being assessed from more than one test.
   * Compare performance from two different attempts using the same tests.
   * When the test is used regularly the SD can be used as a benchmark to judge training performance.

**Standard scores- Z scores**

1. The anomalies indicated earlier in paragraph 21 and demonstrated in graph 2 arise because the underlying parameters for each test are different. One technique has been developed to convert the individual marks or scores on each test to a common scale.
2. The standard score (usually called Z scores) establishes the arithmetic mean as 0 and uses the SD as a unit of measurement.
3. The z score for each student is calculated by

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Z score = Deviation of score from mean *Z* | | | = | *x* | − | *x* |
|  |
|  | Standard deviation |  |  | *SD* | | |

**

1. The following table illustrates the method. The following set of scores have a mean of
2. and a SD of 4.

|  |  |  |
| --- | --- | --- |
| **Score** | **Deviation** | **Z Score** |
| 16 | 6 | + 1.5 |
| 15 | 5 | + 1.25 |
| 12 | 2 | + 0.5 |
| 11 | 1 | + 0.25 |
| 10 | 0 | 0 |
| 9 | - 1 | - 0.25 |
| 8 | - 2 | - 0.5 |
| 5 | - 5 | - 1.25 |
| 4 | - 6 | - 1.5 |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-77 |

UNCONTROLLED COPY

1. By converting test scores to this standard scale we have overcome the problem of different means, each test now has a mean of zero.
2. However people with no statistical training do not readily understand Z scores. Telling a student that they have passed a test with a score of + 1.25 would cause confusion at least; imagine if another student’s score was stated as – 1.25.

**T Scores**

1. The T score converts the Z Score into a more readily appreciated form. The most satisfactory scale has a mean of 70 and an SD of 10. This will give a range of marks between 40 and 100.
2. Using this scale, to convert a Z score to a T score we must
   1. Multiply the z score by 10
   2. Add 70 to the result
3. For example: if the Z Score is 2.1 then:
   1. Z score = 2.1
   2. 2.1 x 10 = 21
   3. 21+70=91
   4. T score = 91
4. And again: this time using a negative z Score of – 1.5
   1. Z score = – 1.5
   2. – 1.5 x 10 = – 15
   3. –15+70=55
   4. T score = 55

**Summary**

1. The aim must always be to provide accurate analysis of data. This is often time consuming and by inference expensive to produce. The expense can always be justified by thinking of the whole process as one of quality control. A shorthand guide to calculating the z score can be found at Annex G.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-78 |

UNCONTROLLED COPY

**ANNEX F TO**

**SECTION 5**

**STANINES AND GRADING**

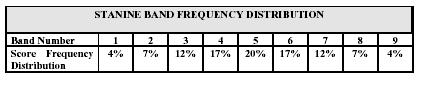
**STANINES**

1. When distributions are badly skewed and Z scores suspect as a method of scaling, a better method of assessing an individual's relative attainment is to use Stanines.
2. Stanine, which is short for Standard Nine, is a numbered convention that represents the conversion of raw data onto a Normal Distribution profile. The convention is based on the ranking of raw data within specified percentage frequency intervals. The percentage frequency intervals

are represented by the Stanine Band shown in the table below, where the highest performer is in band 9 and the lowest in band 1. This forces the distribution of scores found (the empirical distribution), whatever its shape, into a normal distribution with a mean of 5 and an SD of 2. A stanine grade of 5 is thus equivalent to a Z score of 0; 7 is equivalent to a Z score of +1; 2

is equivalent to a Z score of -1.5, etc.

1. **An Example**. The following table shows 100 scores obtained on a test of training. They are rank ordered and split into 9 stanine intervals.



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Score | Freq | Stanine | Interval | Score | Freq | Stanine | Interval |
| (x) |  |  |  | (x) |  |  |  |
| 109 | 1 | 9 | >102 | 84 | 2 | 4 | 80-84 |
| 105 | 1 | (4%) |  | 83 | 5 | (17%) |  |
| 103 | 2 |  |  | 82 | 3 |  |  |
|  |  |  |  | 81 | 5 |  |  |
| 102 | 1 | 8 | 97-102 |  |  |
| 101 | 1 | (7%) |  | 80 | 2 |  |  |
| 99 | 2 |  |  |  |  |  |  |
|  |  | 79 | 1 | 3 | 72-79 |
| 98 | 3 |  |  | 78 | 2 | (12%) |  |
|  |  |  |  | 76 | 3 |  |  |
| 96 | 3 | 7 | 94-96 |  |  |
| 95 | 3 | (12%) |  | 75 | 3 |  |  |
| 94 | 3 |  |  | 74 | 1 |  |  |
|  |  |  |  | 73 | 1 |  |  |
| 93 | 3 | 6 | 89-93 |  |  |
| 92 | 4 | (17%) |  | 72 | 1 |  |  |
| 91 | 4 |  |  |  |  |  |  |
|  |  | 71 | 2 | 2 | 60-71 |
| 90 | 4 |  |  | 70 | 2 | (7%) |  |

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-79 |

UNCONTROLLED COPY

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 89 | 2 |  |  | 65 | 1 |  |  |
| 88 | 5 | 5 | 85-88 | 63 | 2 |  |  |
| 87 | 2 | (20%) |  | 58 | 1 | 1 | <60 |
| 86 | 7 |  |  | 56 | 1 | (4%) |  |
| 85 | 6 |  |  | 50 | 1 |  |  |
|  |  |  |  | 46 | 1 |  |  |

1. Because Stanines are a standardised system of scoring, the grades obtained can be added and compared. The great advantage of stanines is that they convert any distribution to an even scale. The decision whether to use stanines or Z scores is usually based on visual inspection of the distributions. To be absolutely correct one must compare the distribution with the same mean and standard deviation. Z scores should be used when the distribution is accepted as being normal.
2. **Standardised Tests**. Once a test has been standardised in this way on about 50 results it is no longer necessary to rank all the scores to allocate stanines. A raw score can be converted to a stanine simply by fitting it into the interval for that particular test. After a second 50 results the whole group should be re-ranked and new intervals calculated. It is unlikely that the interval will change much after this unless the training given is reorganised. The greatest advantage of the stanine system is that an individual score cannot only be compared with the scores obtained by trainees on the same course but also with the previous 50 scores (e.g. previous five courses if 10 on each course). Therefore a trainee with a stanine score of 9 is not only in the top 4% the course, but also of the previous courses that have been used to standardise the scores.
3. **Calculation of Stanines**. Because the stanine distribution is based on a mean of 5 and a SD of 2, a stanine can also be calculated directly from the Z-Score by times the z score by 2 and adding 5.
4. If the calculated value is less than 1 then it is specified as Band 1 as it is in the lowest 4%; conversely, any calculated number greater than 9 is specified as Band 9 as it is in the highest 4%, etc.
5. **Limitations**. The use of Stanines is limited as it is only suitable for analysis of large numbers and can only provide trends in mark intervals. Therefore, their use is not recommended for detailed analysis when data is sufficiently normally distributed for T-Scores to be calculated.

**GRADING**

1. Quite often grades rather than marks are given for a test. When marking essays giving grades is usually as precise a method of scoring as is realistic. This is also true of assessment of qualities such as leadership.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-80 |

UNCONTROLLED COPY

1. Gradings can only compared if the underlying scale is standardised. It is impossible, for instance, to compare a grade B on one course with a grade B on another is one does not know the proportion of trainees to each grade.
2. A suggested distribution is:

Grade A B C D E

5% 25% 40% 25% 5%

1. This forces the distribution in the shape of a normal distribution.
2. Grades from such a scale can then be added or compared. It is however, **essential** to keep a constant check on instructors giving the grades. If they deviatefrom the percentages laid down the scale collapses and the grades become meaningless. One ought to note however, that the shape will not stabilise until at least 50 students have been graded.
3. This system of grading has a number of other advantages:
   1. the rather more popular distribution of 10,20,40,20 10 (although easier to use) does not give a normal distribution.
   2. 5 % for the top grade is realistic; one does not give the top grade to one in ten trainees.
   3. the centre points of the 5% and 25% groups can be used for weighting as they fall conveniently near whole numbers.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-81 |

UNCONTROLLED COPY

**ANNEX G TO**

**SECTION 5**

**SHORTHAND TO STATISTICS FOR TEST ANALYSIS**

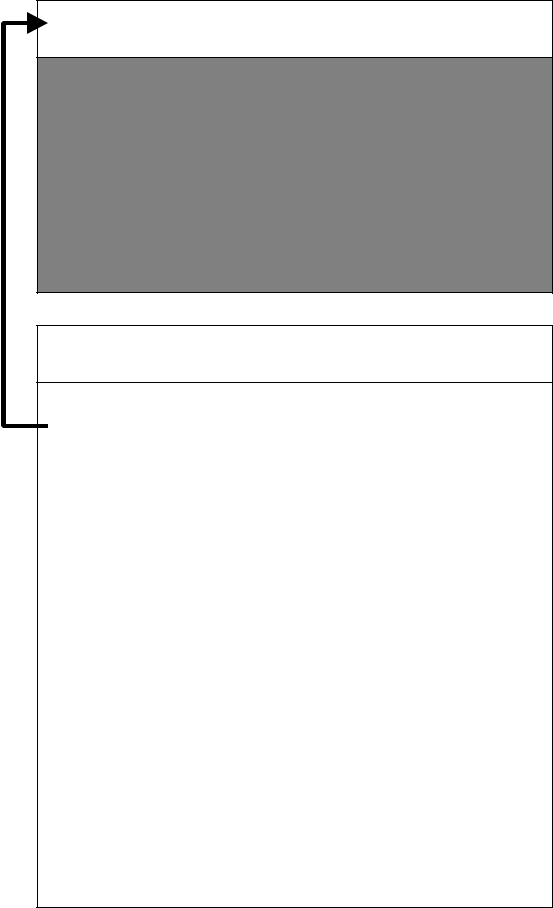
**THE LANGUAGE OF STATISTICS**

1. “x” represents a value or score eg scores of 55, 53, 52 can be represented as x1, x2, x3 etc.
2. “n” stands for the number of items or scores.
3. ∑ stands for the sum of, ∑ x so means all the scores added together.

4 *x* stands for the arithmetic mean.

1. σ or σ n-1 stands for standard deviation.
2. “**r**” stands for correlation.

**The Arithmetic Mean**

****

1. Add up the set of scores (∑x)
2. Count up the number of scores (n)
3. Divide Step 1 by Step 2 to give the mean

**Standard Deviation (SD)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 |  | W |  |  | for(⌧)the score | |  |
|  | 1 |  |  | ork out the mean ( ) for the scores | | | | |
|  | 2 |  |  |  |  |  | it |  |
|  |  | For each score work out how much it | | | | |  |
|  | 2 |  | d |  |  | - | ⌧) |  |
|  |  |  | differs from the mean (x - | | | | ) |  |
|  | 3 |  | S | e | result-⌧from) S ep 2 | | | 2 |
|  | 3 |  | Square each (x - | | | ) result from Step | | 2 |
|  | 4 |  |  |  |  |  | 3 |  |
|  | 4 |  | Add up all the results from Step 3 | | | | |  |
|  | 5 |  |  |  |  |  | - |  |
|  | 5 |  | Work out the degrees of freedom (n-1) | | | | | |
|  | 6 |  |  |  | 4 | 5 |  |  |
|  | 6 |  | Divide Step 4 by Step 5 | | | |  |  |
|  | 7 |  |  |  |  |  | 6 |  |
|  |  | Find the square root of Step 6 to give | | | | |  |
|  | 7 |  | the standard deviation (SD) | | | | |  |
|  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Standard Deviation (SD)** a

measure of the scatter of the results about the mean

**Standardised Scores(z-score)** a way of comparing sets of results which have different means, scores and SDs

*z = deviation from mean standard deviation*

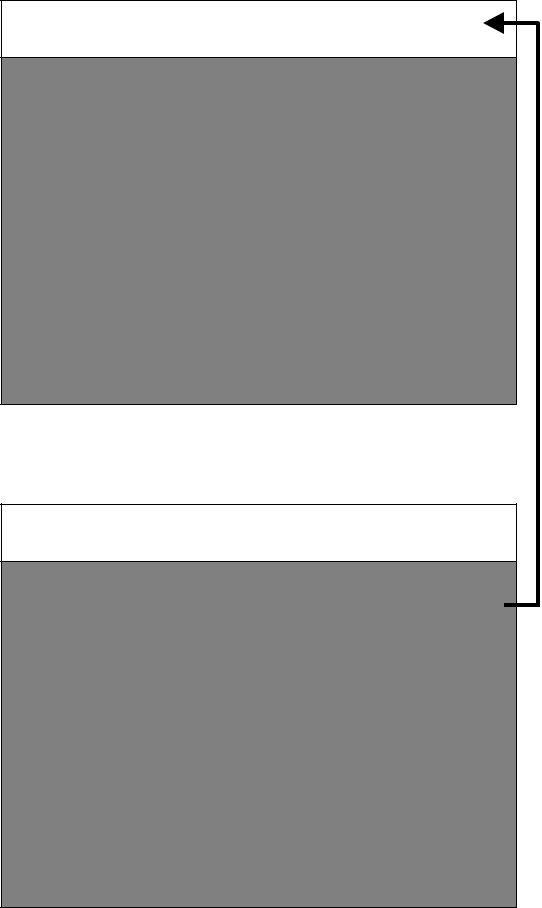
**t-scores** do the same but in a more understandable way. They assume a mean of 70 and SD of 10, so to make a z-score into

1. t-score multiply by

10 and add 70: *t= (zx10)+70*

**Correlation (r)** is a measure of how alike two sets of scores are. It is used to determine whether a test is reliable (likely to produce the same set of results each time it is used)

**Standardised Scores (Z-scores)**

****

1. Work out the mean for the scores
2. For each score work*x* out how much it differs from the mean
3. Work out the standard deviation (SD) for each set of scores
4. Divide Step 2 by Step 3 to give the standardised score (z-score)

**Correlation Coefficient (r)**

1. Convert the scores (x) into standardised form (z-score)
2. Multiply the z-scores together and add them up
3. Work out the degrees of freedom (n-1)
4. Divide Step 2 by Step 3 to give the correlation coefficient, r

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-82 |

UNCONTROLLED COPY

**SECTION 6 – GLOSSARY OF TESTING TERMS**

**Accuracy:** The degree of freedom from error or the degree of conformity to a standard.

**Achievement:** A measurement of what a person knows or can do after training.

**Aptitude:** the natural ability to acquire and utilise specific skills and knowledge.

**Assessment:** Essentially a measurement process of the learning that has either taken place or can take place. Usually measured against stated learning outcomes.

* predictive testing: attempts to assess what the student might learn given suitable tuition.
* attainment testing: attempts to measure what the learner knows or can do, and is usually related to the syllabus of a course the learner has followed.

**Assignment:** Learners are presented with a topic/subject or problem for an in-depth analysis.

**Benchmark measures:** A set of measurements (metrics) that is used to establish goals for performance improvements. These are often derived from other firms that display "Best In Class" performance.

**Checklist:** A tool used to ensure that the important actions or steps in the performance of a task have been taken. If the steps have to be performed in a particular order, then it should be noted on the checklist.

**Closed question:** a question in a test, or programmed learning sequence, in which a unique answer is required, and where there is no scope for divergent thinking or qualification.

**Competence:** The ability to perform the activities within an occupation or function to the standards expected in employment.

**Competency:** (1) Areas of personal capability that enable people to perform successfully in their jobs by completing task effectively. A competency can be knowledge, attitudes, skills, values, or personal values. Competency can be acquired through talent, experience, or training. (2) Competency comprises the specification of knowledge and skill and the application of that knowledge and skill to the standard of performance required in employment.

**Competency-based instruction:** Instruction that is organised around a set of learning objectives based upon the knowledge, skills and attitudes required to perform a set of skills called competencies. Evaluation of student success is based on competent performance of the skills. Normative measurement is specifically excluded from competency-based instruction.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-83 |

UNCONTROLLED COPY

**Completion item:** A test component requiring the completion of a statement, phrase, or concept. Also called a **Supply** item

**Confirmation:** Giving each learner knowledge of the results of each exercise throughout the instructional process. This reinforces or rewards the learner during the entire learning situation.

**Constructed response test item:** An examination item requiring recall or completion as opposed to recognition (e.g., drawing a diagram; filling in a form; labelling the parts of a piece of equipment; writing a sentence, paragraph, or essay).

**Content validity:** An assessment that attests that a product will produce the desired results. One example in military training compares objectives, tests, and materials to ensure that they track with each other and, thus, may be expected to produce the desired results.

**Continuous assessment:** Where assessment is carried out at frequent intervals during the course. It is often part of the summative assessment.

**Criterion:** The standard by which something is measured. In training, the task or learning objective standard is the measure of student performance. In test validation, it is the standard against which test instruments are correlated to indicate the accuracy with which they predict human performance in some specific area. In evaluation it is the measure used to determine the adequacy of a product, process, or behaviour.

**Criterion-referenced instruction:** Testing of the objectives as a learner progresses through the course of instruction. Learners pass or fail depending upon their attainment of the objectives and NOT in accordance with their rank or standing among peers.

**Criterion-referenced objective:** An objective with prescribed levels of performance. Each criterion-referenced objective contains a behaviour (task statement), condition (available equipment, checklists, and governing directives, or the situation requiring the task), and a standard (regulation, operating instruction) for the task.

**Critical task:** A task that, if not accomplished to the specified standard, results in a serious adverse effect upon mission accomplishment, survivability, or safety. Critical tasks must be trained.

**Dependent testing:** Requiring student mastery of skills and knowledge in one learning objective before testing skills and knowledge in another learning objective.

**Developmental testing:** Progressive testing of an instructional system before implementation to evaluate the effectiveness of the course. It consists of three phases:

* individual testing - done with an individual learner to evaluate sequence and grammatical structure or problem areas in the script while the material is still in draft format.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-84 |

UNCONTROLLED COPY

* group tryout - done with individual learners to evaluate the final format after revisions are made from the individual testing.
* operational tryout - conducted by course personnel in the actual training situation before implementation into the system.

**Differential feedback:** Test response feedback specific to the multiple-choice answer selected by the student.

**Differential weighting:** Assigning different weights (marks) to different questions or options in a multi-choice item in order to give some reward for partly correct answers.

**Discrimination:** The ability to choose between two closely related responses to a specific stimulus.

**Distracter:** In testing, incorrect answers provided as choices in multiple choice or matching type test items.

**Drill:** A standardised technique or procedure that prepares students to execute critical collective tasks in an instinctive and spontaneous manner. The drill includes the method(s) by which it is trained.

**Drill and practice:** Upgraded verifications of comprehension of enabling objectives (e.g., questions, exercises, and problems). A method of instruction characterised by systematic repetition of concepts, examples, and practice problems. An ungraded practice quiz.

**Entry behaviours:** Specific competencies or skills a learner must have mastered before entering a given instructional activity.

**Entry test:** Contains items based on the prerequisites that the intended learners must have mastered in order to begin a course.

**Evaluation:** The process of gathering information in order to make good decisions. It is broader than testing, and includes both subjective (opinion) input and objective (fact) input. Evaluation can take many forms including memorisation tests, portfolio assessment, and self-reflection. There are at least six major reasons for evaluating training, each requiring a different type of evaluation:

* Improve the instruction (formative evaluation)
* Promote individual growth and self-evaluation (evaluation by both trainer and learner)
* Assess the degree of demonstrated achievement (summative evaluation attained by the teacher)
* Diagnose future learning needs (of both trainer and learner)
* Enhance one's sense of merit or worth (learner)

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-85 |

UNCONTROLLED COPY

* Identify or clarify desired behaviours (trainer).

**Evaluation hierarchy (four levels of evaluation model):** Donald Kirkpatrick identified the evaluation model most widely recognised today in corporate training organisations. The Kirkpatrick Model addresses the four fundamental behaviour changes that occur as a result of training.

* Level one is how participants feel about training (reaction). This level is often measured with attitude questionnaires.
* Level two determines if people memorised the material (learning). This is often accomplished with pre-testing and post-testing.
* Level three answers the question, "Do people use the information on the job?" This level addresses transference of new skills to the jobs (behaviour change). This is often accomplished by observation.
* Level four measures the training effectiveness, "What result has the training achieved?" This broad category is concerned with the impact of the program on the wider community (results).

**Evaluation instrument:** A test or other measuring device used to determine achievement (go and no-go) or the relative standing of an individual or group or a test objective (i.e., attitude, behaviour, performance objective, and other attributes). Evaluation instruments include tests, rating forms, inventories, and standard interviews.

**Facility value (FV):** a numerical expression of how easy/difficult students find a test item.

**Formative testing:** testing that is carried out during an activity, performance, project or course to determine effectiveness/quality at a time when remedial action can still be taken.

**Go / no go:** Two parameters are possible: Go - passes the performance standards or No Go - fails the performance standards. The evaluation criterion whereby a student cannot partially pass.

**Index of discrimination (ID):** a numerical expression of how well a multi-choice item distinguishes between students of different abilities.

**Item bank:** a collection of test questions, together with their characteristics, facility values and discrimination indices, dates and conditions of previous use which may be used for automatic construction of tests.

**Lickert Scale:** A way of generating a quantitative value (numerical) to a qualitative questionnaire (e.g. poor, fair, good, very good, excellent). Sometimes used on end of course evaluation. (smile sheets) For an ascending five point scale incremental values are assigned to each category and a mean figure for all the responses is calculated. (via the sum of the products of the categories' assigned value times the number of respondents for that category, divided by the total number of respondents) Example: Total number of

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-86 |

UNCONTROLLED COPY

respondents=25, assigned values are; poor=1, fair=2, good=3, very good=4, excellent=5; respondents selecting following categories are; good=9, very good=10, excellent=6. The quantitative mean = ((9\*3)+(10\*4)+(6\*5))/25=3.9

**Mastery:** Meeting all of the specified minimum requirements for a specific performance.

**Measurement:** A tool used to provide feedback to the learner and the trainer to determine where the learner is in relation to the ultimate goal or objective.

**Metrics:** A set of numerical performance measurements related to a goal or a set of goals.

**Multi-choice test item:** A test item containing a question in the form of a stem and several alternatives, which are possible answers to the question posed.

**Objective testing:** Testing that eliminates any subjective bias caused by the judgement of the marker, by limiting choices to fixed alternatives (selection items).

**Open question:** a question in a test, programmed learning sequence, in which many acceptable answers or outcomes are possible rather than just one single correct solution or outcome.

**Performance checklist:** The breakdown of a learning objective into elements that must be correctly performed to determine whether each learner satisfactorily meets the performance standards described in the learning objective.

**Performance evaluation tools:** Competency tests that allow the trainer to profile the student's proficiency and identify weak areas so that training can be efficiently planned for the areas of most critical need.

**Performance evidence:** Evidence that the individual can actually achieve the standard, meeting all the specified performance criteria, is essential for a judgement of competence. Knowledge evidence alone is not enough. Performance evidence may be naturally occurring or the product of special assessment arrangements such as simulations, projects and assignments.

**Performance exercise** A proficiency (criterion-referenced) test used to evaluate mastery of a task as specified by the criterion-referenced objective.

**Performance measures:** The actions that can be objectively observed and measured to determine if a task performer has performed the task to the prescribed standard.

**Practical exercise:** A technique used during a training session that permits students to acquire and practice the knowledge, skills, and attitudes necessary to successfully perform one or more training objectives.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-87 |

UNCONTROLLED COPY

**Predictive validity:** The extent to which the test or expert opinion predicts how well students will actually perform on the job.

**Quiz:** A short test administered by the instructor to measure achievement on material recently taught or on any small, newly completed unit of work.

**Ranking**: Arranging scores, measures, items, individuals in order of merit or magnitude.

**Raw score**: A score that has not so far been modified by mathematical operations such as scaling or standardisation. Also known as an observed score.

**Reliability:** Yielding comparable results each time. In examinations, reliability is consistency; the same result is achieved on successive trials (see “test reliability”).

**Remediation:** Supplemental course materials to correct a learner's understanding or to reinforce the learning objective.

**Selection Items.** Questions where the student has to select the correct answer from a group of correct and incorrect responses. EG Multiple choice and matching questions. Often called **Objective** tests.

**Standards:** Describes the criterion or standards of performance that must be attained. An established norm against which measurements are compared. The time allowed to perform a task including the quality and quantity of work to be produced.

**Standardisation:** Transformation of a set of raw scores so that they have a mean and standard deviation.

**Stem:** The part of a test item that asks a question.

**Summative testing:** Evaluation that is carried out at the conclusion or completion of a performance, activity or project in order to provide data for determining the overall effectiveness of a student. A test that is used to determine if a trainee has achieved mastery of a training objective or enabling objective.

**Supply Items.** Questions where the student has to supply the answer. e.g. missing word, short answer or essay questions. Often called **subjective** tests as the answers may need to be interpreted.

**Subjective testing:** Testing that relies upon the judgement of the marker.

**Test:** A device or technique used to measure the performance, skill level, or knowledge of a learner on a specific subject matter. It usually involves quantification of results -- a number that represents an ability or characteristic of the person being tested.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-88 |

UNCONTROLLED COPY

**Test fidelity:** The degree to which the test resembles the actual task performed. The closer the resemblance, the higher the fidelity of the test.

**Test, instructional:** Any device/technique used to measure the performance, skill level and knowledge of an individual. See appropriate types listed below:

* **Achievement test**. A test for measuring an individual's attainment of knowledge/skills as the result of specific teaching or training.
* **Aptitude test.** A test or battery of tests designed to show a person's capacity for a particular type of behaviour in a single field or in several related fields.
* **Attainment test.** A test that measures what a person has actually achieved in a particular area (skill) at the time of testing; often used as a pre- or post-test.
* **Comparative test**. A test given at the completion of a major section of a course and, as required, at completion of a course to measure whether the student has mastered the course learning objectives.
* **Criterion-referenced test**. A test that measures to what extent a trainee has performed the learning objective to the established standard. Performance can be assessed by comparing the measurement against the prescribed criterion or set of criteria - the training objective standard as a "go" / "no-go" or simply to record the degree of achievement against the standard. It is scored based upon absolute standards, such as job competency, rather than upon relative standards, such as class standings.
* **Diagnostic test**. A test used to measure performance against a criterion and to identify specific areas of weakness or strength in individual knowledge and skills.
* **End-of-course comprehensive test**. An end-of-course test, administered to all initial entry students prior to graduation, designed to ensure a high probability that students can perform all critical tasks taught in the course. It provides feedback on the need for both reinforcement training and course revisions.
* **Entry test**. A test designed to determine if a student already possesses certain knowledge or skills needed as a prerequisite before undertaking new instruction.
* **Field test**. Tryout of any training course on a representative sample of the student target population to gather data on the effectiveness of instruction in regard to error rates, criterion test performance, and time to complete the course.
* **Heuristic test**. Heuristic or discovery tests will present problem-solving simulations that emulate the on-the-job environment. These tests present the student with stimulus information that is inadequate, incomplete, ambiguous, or irrelevant to the simulated environment. The student will be required to synthesise knowledge and apply training received in order to solve the job performance simulation.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-89 |

UNCONTROLLED COPY

* **Job performance test**. A test used to determine whether an individual can perform a job. It may include all job performance measures for a job or a subset of the job performance measures.
* **Knowledge test**. A test that measures the achievement of theory supporting skill through the use of test items written at the appropriate knowledge and training levels.
* **Multiple-choice test**. A type of selection test in which the student is asked to choose for each test item the answer(s) that is most correct.
* **Non-language test**. Identical to the definition for "non-verbal test".
* **Non-verbal test**. A test that requires little or no speaking, reading, or understanding of language on the part of the examinee either in connection with comprehending directions or making responses. Directions may be given pictorially or in pantomime. Also called "non-language test”.
* **Norm-referenced test**. A test that ranks a student in relation to the performance of other students in contrast to criterion-referenced testing wherein a student is measured against a prescribed performance standard.
* **Objective test**. A test whose scoring requires no human judgement.
* **Performance test**. An evaluation of the actual performance of the task or learning objective using the conditions under which it will be performed and the absolute standards for acceptable performance.
* **Post-test**. A test administered to a student upon completion of a course or unit of instruction to measure learning achieved and to assess whether a student has mastered the objectives of the course or unit of instruction.
* **Power test**. A test in which items are usually arranged in order of increasing difficulty and in which examinees are given all the time they need to complete as many items as they possibly can.
* **Pre-test**. A test administered to a student prior to entry into a course or unit of instruction to determine the technical skills (entering behaviours) the student already possesses in a given subject. Often used to identify portions of the instruction the student can bypass.
* **Proficiency test**. A test designed to measure a student's capabilities in terms of the job. It measures both psychomotor and cognitive skills. A performance test is sometimes understood to mean a skill demonstration, while a proficiency test is understood to be a comprehensive procedure used to examine the student's capability to do what the job requires.
* **Progress test**. A short test administered throughout a course to evaluate student progress. It is administered at strategic points in a course to determine the degree to which students are accomplishing the learning or enabling objectives. Also called “within-course test”.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-90 |

UNCONTROLLED COPY

* **Qualifying test**. A test administered to determine whether a student is qualified for a task that the student has been selected or trained for, or for which the student is being considered. A qualifying test may also be applied to tests used for selecting personnel for training, although the usage is not so common.
* **Quiz**. A short test administered by the instructor to measure achievement on material recently taught or on any small, newly completed unit of work
* **Recall test**. A type of objective test in which the subject is required to supply missing items of information (usually words, phrases or numbers) to complete statements relating to material previously encountered or learned.
* **Retention test**. A test administered some time after the completion of a course or programme of instruction in order to determine the extent to which acquired knowledge or skills have been retained by the learner.
* **Simple gaming test**. Presents the student with fill-in-the-blank, multiple choice, matching, completion, and true/false test items formatted and presented in a gaming style.
* **Simulated part-task performance test**. Measures critical sub-sets of job task performance. Simulated part-task performance tests should meet the same construction criteria as simulated performance tests.
* **Simulated performance test**. A performance-based two-dimensional simulation of the job performance required. A synthetic performance test.
* **Simulation performance test**. A test that measures the student's ability to meet training objectives by performing whole tasks or parts of tasks using simulators or simulations.
* **Speed test**. A test in which the time limit is set so that almost no one can finish all the items or tasks making up the test.
* **Survey test**. A criterion-referenced test used prior to the development of an instructional system. It is administered to a sample of prospective students to determine what knowledge and skills should be put into the course of instruction. Also called "Threshold Knowledge Test (TKT)".
* **Threshold Knowledge Test** (TKT). Identical to the definition for "survey test".
* **True-false test**. A type of selection or alternate-response test in which the student indicates whether each of a number of statements is true or false.
* **Verbal test**. Any test involving language. In general usage, the term is restricted to those tests in which the questions and responses are mainly expressed in language or which use language to a substantial degree.
* **Within-course test**. Identical to the definition for progress test.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-91 |

UNCONTROLLED COPY

* **Written test**. A test in which an individual demonstrates their capabilities by responding to written test items. It is not usually a performance test, and hence is usually a measure of supporting knowledge rather than skills.

**Test item analysis:** The process of evaluating single test items by any of several methods. This usually involves the determination of how well an individual item separates examinees, its relative difficulty value, and its correlation with some criterion of measurement.

**Test reliability:** The degree to which a test/test item gives consistent results each time it is used i.e. to what extent will it produce the same measurement when it is repeated.

**Test validity:** The degree to which a test measures what it was designed to measure.

**Validity:** The degree to which a test measures what it is intended to measure. Although there are several types of validity and different classification schemes for describing validity there are two major types of validity that test developers must be concerned with, they are content-related and criterion-related validity.

|  |  |
| --- | --- |
| Issue V1.0 Aug 07 | 2-92 |

UNCONTROLLED COPY