CAA CENTRE
COMPUTER ASSISTED ASSESSMENT

Principles of Assessment

by Mhairi McAlpine
Robert Clark Centre for Technological Education,
University of Glasgow

edited by CAA Centre, University of Luton.

Bluepaper Number 1
February 2002

The CAA Centre TLTP Project

The Principles of Assessment is published by members of the Implementation and Evaluation of Computer-assisted Assessment consortium, a project funded by the HEFCE and DENI under phase three of the Teaching and Learning Technology Programme (TLTP). The project is led by the University of Luton and includes Glasgow, Loughborough and Oxford Brookes Universities.

Copyright

ISBN 1-904020-01-1

TABLE OF CONTENTS

PRINCIPLES OF ASSESSMENT

Introduction

There are a variety of issues that must be taken into consideration when planning an assessment strategy within higher education. It is important to start with the understanding that assessment is a form of communication. This communication can be to a variety of sources, to the students (feedback on their learning), to the lecturer (feedback on their teaching), to the curriculum designer (feedback on the curriculum), to administrators (feedback on the use of resources) and to employers (quality of job applicants).

There are five main points to consider when designing any assessment to ensure that the communication is as meaningful, useful and honest as possible.

1. The purpose of the assessment and whether the task fulfils that purpose

An essential starting point is to be aware of the reasons why you are assessing the students, and how to design an assessment that will fulfil your needs. To do this it is important to consider the decisions you are going to make, the information you need to gather to make those decisions, and what methods are the most effective for gathering that information.

2. The validity and reliability of the assessment that you are constructing

To ensure that the communication is as honest as possible it is crucial to make sure that the assessment is both valid - in that it tests a relevant skill or ability, and that it is reliable - in that the same result would be achieved if you repeated the assessment.

3. The referencing of the assessment

To make the assessment meaningful it is important to compare the candidates' abilities with a common measure. The most common comparisons made are with other candidates, with objective criteria, or with the candidates' own performance in another area. Through careful consideration of the purposes of the assessment the most appropriate reference frame should become clear.

4. The construction quality of assessment items

To ensure useful assessment, the assessment items must be constructed to an appropriate quality. Judging the quality of items can be complicated but, as a starting point, consider the difficulty level of the items. In general, a good assessment ought to be at about the difficulty level of the average candidate. Also consider how well the assessment differentiates between the candidates - to provide maximum information the assessment ought to separate out the candidates as much as possible.

5. The grading of the assessment

Grades awarded are very concise summaries of students' abilities. They are generally designed for purposes outwith the institution and, therefore, they should be clear and easily understood by a lay audience. The grading of the assessment is often related to the referencing of the assessment, and as such the two should be considered in tandem.

PURPOSE OF ASSESSMENT

Before designing any assessment, you need to ensure that what you are planning will fulfil the demands that you wish to make on it. This involves a thorough examination of your reasons for assessing: considerations may include the information that you want to get out of the task, the uses that you will put that information to, how much time and effort you are able to devote to it, what information you wish to convey to students and others. The rest of this section discusses some of the decisions that you may wish to consider.

Formative or summative

Formative assessment is designed to assist the learning process by providing feedback to the learner, which can be used to highlight areas for further study and hence improve future performance. Self and diagnostic assessment are types of formative assessment with specific purposes. Summative assessment is for progression and/or external purposes, given at the end of a course and designed to judge the students' overall performance.

Examples

Summative assessment	**IQ tests, traditional examinations, driving test.**
Formative assessment	**computer-based test which provides feedback on areas of weakness, an essay which is annotated with the lecturer's comments, but no overall mark.**

Advantages and disadvantages

Summative assessment is the most useful for those external to the educative process who wish to make decisions based on the information gathered, for example employers, institutions offering further study, the courts (in the case of a driving test). It generally provides a concise summary of a student's abilities which the general public can easily understand either as a pass/fail (driving test) or a grade (A-E; 1-7; 1st-3rd etc.). It is not however very useful for communicating complex data about a student's individual abilities - are they strong in algebra but weak in calculus for example. Formative assessment on the other hand allows the students and other interested parties to form a more detailed opinion of their abilities, which can then be used to inform further study, concentrating students' efforts on the more appropriate areas and hence improving overall performance.

Appropriate use

Formative assessment is most appropriate where the results are to be used internally by those involved in the learning process (students, lecturers, learning support etc.), whilst summative assessment is most appropriate to succinctly communicate students' abilities to external interested parties.

Formal or informal

Formal assessments are where the students are aware that the task that they are doing is for assessment purposes. With informal assessment the judgements are integrated with other tasks.

Examples

Formal assessments	**examinations, coursework essays, theses.**
Informal assessments	**lecturer notes taken during a practical, tape-recordings of class French conversations, audit trails of the use of computer-based learning and self-assessment tests.**

Advantages and disadvantages

Formal assessments are perceived to be 'fairer'. Criteria tend to be more explicit and have less room for bias. Students know they are to be assessed and behave accordingly. However, such assessments can induce stress sometimes causing students to perform less well; others may cram and perform well, but without deep understanding. Informal assessments can reduce stress, and give a more valid view of students' abilities, however some students may feel 'cheated' out of their chance to shine (see section on Validity and Reliability, page 11). There can also be problems with hidden prejudices and stereotypes influencing the judgement of the assessor when informal assessments are used.

Appropriate use

For summative assessment, formal is most frequent, while for formative and diagnostic assessment, informal is more common. Where there is evidence of high examination stress, or where a formal exam would be so artificial that it would challenge the validity of the assessment, summative informal assessment is desirable. Formal assessment, however, can have motivational effects. If students are unmotivated, early formal assessment may be useful to encourage achievement.

Final or continuous

Final (terminal) assessment is that which takes place only at the end of a course while continuous assessment is scattered throughout the course.

Examples

Final assessments	**the traditional 'finals' assessment, where the result of 3 years' study is assessed over a period of a few days.**
Continuous assessment	**the more modern form of modular assessment, where judgements are made at the end of each field of study.**

Advantages and disadvantages

The primary advantage of final assessment is that it is simple to organise and condenses the assessment process into a short space of time. This means, however, the timing of the examination becomes of great importance. Illness at an unfortunate time can unduly influence the result. Furthermore final assessment cannot be used for formative purposes. The main advantages of continuous assessment are that both students and lecturers obtain feedback from the process which can then be used to improve teaching and learning, and the final result is based on evidence gathered over the span of the learning period. Disadvantages include the increased workload inherent with this mode of assessment, and difficulties associated with students from different backgrounds tackling the same material and being assessed in exactly the same way.

Example

Imagine for example Maths 101, a course attended by engineers and philosophers as well as mathematicians. The mathematicians may well score better in the assessment than the other students despite the fact that they had extracted the relevant parts from the course and integrated them into their field of study. Under continuous assessment, they would be assessed as mathematicians rather than as philosophers and engineers.

Appropriate use

Final assessment may be appropriate where there is evidence that learning each new field of study contributes to the understanding of every other, and hence learning can only be assessed as a complete whole rather than as constituent parts. Continuous assessment is more appropriate where student feedback is required and a series of pieces of information are required across the course to build up a picture of students' abilities. Computer-assisted assessment (CAA) can provide a powerful means of continuous assessment, providing rapid and detailed feedback to students and academics about the learning process.

Process or product

With the rapidly changing nature of modern society, increased emphasis is being placed on skills and abilities rather than knowledge. It is therefore important to consider whether you wish to assess the product of student learning, or the process undertaken.

Examples

Product driven	**(e.g. in French) computerised objective test of recently taught vocabulary, an essay question on an area of study.**
Process driven	**(e.g. in French) computerised objective test of unfamiliar vocabulary aided by an online French dictionary, research of an unfamiliar area.**

Advantages and disadvantages

Product-driven assessments are usually easier to create, as the assessment criteria tend to be more tangible. They can also be more easily summarised. Process-based assessments however can give more useful information about skills, and can highlight to students the importance of learning generalised techniques rather than specific knowledge. Some students do see process-based assessments as unfair 'How are we supposed to write an essay on Quarks when you haven't taught us about them?'. Therefore, the reasons for assessing in this manner, the criteria that will be applied, and what will be considered evidence must be explained carefully to students.

Appropriate use

Process-based assessments are best where the learning is skill or ability-based, while product-based assessments are best where the knowledge content of the course is fundamental. Most assessments are mixtures of the two forms. The balance is critical in ensuring that the assessment is fit for the purpose.

Convergent or divergent

Convergent assessments are those which have one correct answer that the student is trying to reach. Divergent assessments appreciate a range of answers based on informed opinion and analysis.

Examples

Convergent assessment computerised objective test; there is only one correct answer.

Divergent assessment essay type questions; no one correct answer, but rather an overall measure of quality.

Advantages and disadvantages

Convergent assessments are generally easier to mark - both by automated and human means. They tend to be quicker to deliver and give more specific and directed feedback to individuals and can also provide greater curricular coverage. However, they can be limited in scope and can occasionally degenerate into a 'quiz' of facts about the area of study. There is also a temptation to only test that which is easily translated into a convergent form.

CAA is an increasingly common form of convergent assessment. Computers offer particular advantages in extending the scope and authenticity of convergent assessments, however good questions and tests require skilled design and construction. Divergent assessments can be more authentic, and make it easier to assess higher cognitive skills. However, they can be time consuming to set and mark. They also require greater marking skill than convergent assessments, this can involve training markers and/or detailing criteria.

Appropriate use

Where knowledge is the primary issue, convergent assessments can be very useful. Because of their wide curricular coverage, they can be very important in formative assessment to quickly and effectively highlight areas of weakness and gaps in students' knowledge. Where there is a core of knowledge that is a fundamental base for the study of the subject, convergent assessment can be an effective way of ensuring that it is in place. CAA is often used to provide broad and rapid assessment of students' knowledge, it can effectively identify gaps in students' knowledge using statistical analysis and reports. Divergent assessments by contrast are best suited when there may be a difference of opinion based on interpretation. This is most obvious in an area such as literary criticism, but can also be important in a medical diagnosis for example. A divergent assessment - requiring students to explain the basis for their diagnosis - can check students' reasoning, and uncover misapprehensions that they might be under. It also allows for valid diagnoses that may not have occurred to the question setter.

VALIDITY AND RELIABILITY

Although validity and reliability are separate entities it makes sense to consider them together because jointly they define the overall quality of assessment. Conventional wisdom suggests that a valid test must always be reliable, although this is being challenged in some quarters. A valid assessment is one which measures that which it purports to measure, while a reliable assessment is one where the same results are gained time after time.

Validity

A valid assessment is one which measures that which it is supposed to measure. For example, a Physics assessment which insisted that answers had to be written in German would not be a valid assessment as there is a good chance that you would be testing students' knowledge of German rather than their abilities in Physics. It is important when designing an assessment that you consider whether it does actually assess what you intend it to. There are several different types of validity and it is worth considering each of these in turn.

Curricular (content) validity

The first overarching form is curricular validity - ensuring that the aims of the curriculum are in keeping with what the students need to know. Without curricular validity, not only is the assessment quality doubtful, but calls into question the quality of the whole course. Ensuring curricular validity means ensuring that the learning objectives for the course are closely related to the desirable outcomes of a successful student.

Construct validity

Construct validity is essentially how closely the assessment relates to the domain that you wish to assess. Most assessments require broadly based skills beyond the subject domain (e.g. the ability to read questions involving technical terminology, to construct an essay, even the ability to turn up to the exam hall on time). Some of these skills can be validly included as part of the assessment as they could be considered to be implicit criteria within the learning objectives, while other skills may not be. For example, a CAA which required a high level of information technology skills would be inappropriate if you were testing students' ability to read geological maps. Ensuring construct validity means ensuring that the assessment content is closely related to the learning objectives of the course.

Predictive validity

Predictive validity suggests that predictions made on the basis of the assessment results will be valid. For example you might predict that someone who scored an A in Biology (at A-level) might perform better in a degree course in Biology than someone who failed. If that is the case, then the assessment can be considered to have predictive validity. This type of validity is most important when the primary purpose of the assessment is selective. Ensuring predictive validity means ensuring that the performance of a student on the assessment is closely related to their future performance on the predicted measure.

Reliability

A reliable assessment consistently gives the same results under identical circumstances. A physics assessment which gave the same candidate three different grades on three consecutive occasions, without any substantive change in the candidate's abilities in-between, would not be a reliable assessment. It is important when designing an assessment that you consider whether the results achieved will be consistent. There are several different ways of measuring reliability.

Test-retest reliability

Test-retest reliability is the correlation between candidates' attempts at the same test. Where there is little test-retest reliability, the people who did well first time round may not do well second time round. Obviously this is an important consideration as it suggests that some element of the measure may be due to chance rather than actual skills, ability and knowledge.

Parallel forms reliability

Parallel forms reliability is the correlation between candidates' attempts at two tests which are supposed to be identical. Where this type of reliability is lacking, there is evidence that the tests are testing different things; suggesting that one or both are not testing the pre-defined knowledge and skills - or domain- intended.

Internal consistency

The internal consistency of a test is essentially a pseudo-measure of reliability. Most of the time we do not have the luxury of constructing two separate tests, or testing the students twice. Internal consistency is designed to measure what would have happened had we done that. It is essentially the correlations between the test items. It can be thought of as an estimate of the correlation between the test that was actually delivered, and all of the other possible tests that might have been constructed using those items.

REFERENCING

The referencing of an assessment is the basis of the judgement. There are three main ways of referencing: against peers (norm-related referencing), whereby the judgement is essentially a comparison between the student and other people; against objective criteria (criterion referencing) where the judgement is a comparison between the student's abilities and the contents of a pre-defined domain; and against the student her/himself (ipsotive referencing) where the judgement is a comparison of the student's performance on one area as against prior performance, or performance on other areas.

Norm-related referencing

Norm-related referencing is the comparison of individuals with their peers. This was popular through the mid-20th century, but has become rather unfashionable in modern testing. It can be useful for selective purposes (e.g. for the distribution of a scholarship to the 5 best students, or extra tuition to the 5 which are struggling most), but gives little information about the actual abilities of the candidates.

Norm referencing

Classic norm referencing involves delivering a test to a representative sample of the type of students that you wish to assess, and developing norms based on the results. These norms are then used to grade subsequent groups of students. This can lead to anomalies where the group on which the norm was based becomes very different from the group that is currently taking the examination. This type of referencing is normally credited with maintaining standards across time however, as the curriculum and intake changes, these will not be reflected in the assessment leading to unreliable results.

Cohort referencing

Cohort referencing is similar to norm referencing, however, it takes the subgroup of candidates attempting the assessment as its base-line. Under this type of referencing, the highest results are given to students who attain the best marks relative to their peers who also took the assessment at the time. Unless you can be confident that the intake remains unchanged, this makes for unreliable comparisons across student groups, particularly where the cohort is small. Attainment of a high grade can be as dependent on the performance of the other students taking the assessment as on your own performance.

Criterion referencing

Criterion referencing is a comparison of an individual with pre-defined criteria. It can be used for both formative and summative purposes, both highlighting areas of weakness and determining whether candidates have achieved an acceptable level in the areas they are expected to know about. Results can often be misinterpreted, particularly by those who are more familiar with the older, norm (related) referencing. It must be made clear to users of the assessment data that the criteria for success is performance against learning objectives, rather than performance against other students.

Ipsotive referencing

Ipsotive referencing is a comparison of an individual against him/herself. Although generally unsuitable for selective purposes, Ipsotive referencing can be extremely useful for diagnostic or formative purposes.

Relative ipsotive referencing

Relative ipsotive referencing is the comparison of a person's performance in one sub-domain compared with others, regardless of overall performance. With this type of referencing, students are pointed towards their weakest areas regardless of what their overall abilities in the subject might be. This has the advantage of pinpointing areas for students to work on without the complacency that a good grade engenders, nor the despondency of a weak one. Students are also focused on their own performance rather than on the performance of those around them.

Time dependent ipsotive referencing

Time dependent ipsotive referencing is the comparison of a student's performance over time. In this type of referencing students are encouraged to improve on their past performance on graded questions testing related domain areas. This allows students to see their progressing skills, abilities and knowledge, and harnesses the competitive spirit to positive advantage.

CONSTRUCTION QUALITY

Obviously the construction quality of the assessment items is fundamental in ensuring that the test will provide the information needed. A poorly constructed assessment may provide only partial, or even misleading, information. There are various reasons why questions may be poorly constructed and various ways to improve them. The main indicators of question quality, difficulty and discrimination, are described below. In essence it is essential to ensure that the questions are of appropriate difficulty for the students that you wish to examine and whether they will discriminate adequately between strong and weak students.

As you move along the continuum from objectively marked short answer questions (e.g. multiple choice) to subjective, extended assessments (e.g. theses, projects), the quality of the construction shifts from the question itself to the mark scheme. A particular advantage of CAA is the ability to automatically generate the main indicators of question quality.

Difficulty (facility)

The difficulty of a question (or mark point) can be thought of as the proportion of students who get the question correct. In order that students are separated out as much as possible it is desirable for assessments overall to have a difficulty level of about 0.5 - so that the mean mark is roughly half of the marks available.

Where one question in an assessment carries a high proportion of the marks (e.g. a 25 mark essay question on a paper worth 40 marks), it is desirable for the difficulty level of that question to be close to 0.5. In contrast where an individual question is worth a lower proportion of the marks, it is quite acceptable for it to have a higher or lower facility value.

Where a test is comprised of many questions, each worth a low proportion of the total marks available, it is desirable to have questions which vary in difficulty, so that candidates at all points of the ability stratum may be fully tested. It is, however, undesirable for questions to have facility values above 0.85 or below 0.15.[1], because at this level they are contributing little to overall measurement. The closer the questions come to having a facility value of 0.5, the more they are contributing to the measurement of the candidates.

Example

Imagine 50 candidates taking a 40 mark computerised multiple choice test where the questions are arranged in difficulty order. If the facility value of all the items was 0.5 you might expect the 25 strongest candidates to get 40 while the 25 weakest candidates get 0 (assuming high discrimination). Where there is a range of facility values across the items, you are more likely to get a range of marks, as students fail to achieve on questions which are too difficult for them.

Note

[1] Although on occasion this can be justified for reasons of curricular coverage or criterion referencing.

Discrimination

Discrimination is a measure of how well the question distinguishes between students - and thus how much information the question is providing. There are several methods used to calculate the discrimination of a question, the most common being the Pearson product-moment correlation between the question and total score. This measure assumes unidimensionality[1]. Where this is not the case and the test is designed to examine more than one content area or skill, it may be better to use the correlation between the question and the total of other questions within the same domain as a measure of discrimination.

Being essentially a correlation, question discrimination can vary from +1.0 (where there is a perfect relationship between those who score high marks on the question and those who score high marks on the test) to -1.0 (where there is a perfect *inverse* relationship between those scoring high marks on the question and on the test overall).

In general question discrimination should be positive, unless there is good reason to suppose that the assumption of unidimensionality has been violated. In such a case, question discrimination should be positive within the sub-domain that the question tests, or (if it is the only question representing the sub-domain) with another more representative indicator of performance.

Negative question discrimination with a valid criterion should always be regarded as suspect, however, there is no upper limit for this statistic: the higher the correlation, the better the question discrimination, the better the question. In general values below 0.2 are weak, and values above 0.4 are desirable. It should be noted that questions with lower maximum marks, and those with extreme difficulty levels have less potential for variance than those with higher maximum marks, and hence are likely to have a lower discrimination.

Note

[1] all of the questions are testing a single content area or skill.

GRADING

Grading involves comparing a student's performance with a pre-defined set of standards. The two types of grading most commonly in use are norm referenced grading, where the candidate's performance is compared to other people who are considered to have 'set the standard', and mastery learning where the candidate's performance is compared with a set of learning objectives.

In practice most types of grading involve combining the two types.

Norm referenced grading

Classic norm referenced grading is practised, for example, in IQ tests - where the tests have been calibrated on a pre-defined group, and subsequent test takers' performances are compared to those norms. Within education this type of grading has become rare, however, its ghost still informs many tacit grading beliefs. It has the advantage of seeming to be stable across time - although, of course, as the curriculum and groups of test-takers change this is not necessarily the case.

Most academic departments practise some form of related cohort referenced grading. It would be unusual for a department to willingly dispense first-class degrees to all of its students in one year, even if there were some quirk of the intake which meant that the students for that year were, indeed, unusually able.

Mastery learning

Mastery learning is the grading system generally applied with criterion referencing - and can perhaps best be seen in action within the driving test. Within grading based on mastery learning there are a number of criteria in which the learner must demonstrate competence, with minor allowances. Within mastery learning, there is no concept of marks as such, merely a pass/fail judgement - either the learner has mastered the area, or they have not.

This type of grading does not sit well with the traditional 3 point degree grading structure often in use for undergraduate degrees, although there is more acceptance of this model within post-graduate certification. The model is, however, widely used within professional and vocational examining.

Combining the two

Both mastery learning and norm-referenced grading are generally unsuitable for the assessment needs of a modern undergraduate higher education course. Such courses are generally based on learning criteria, and as such, tend toward a mastery approach to grading. However, this is generally unsuitable for the summative purposes that the assessment is often put to (although it is notable that within medicine, final grading is merely pass/fail).

Some attempts have thus been made to combine the desirable features of the two approaches - the reference group independence of mastery learning with the usefulness of grades. There are two main ways in which this is attempted. Firstly, by establishing levels of competence within the criteria. For example, if the criteria was to understand Brownian motion, a minimally competent learner might have knowledge of the concept and be able to apply it in limited circumstances, while a fully competent learner might be able to apply it in a wider range of contexts. The second attempt at combining the two generally involves the aggregation of criteria - a minimally competent physicist may have mastery of five out of seven areas of physics, a competent physicist have six, and a fully competent physicist have mastery of all seven.

SELECTED BIBLIOGRAPHY

Bond, L.A. (1996) 'Norm-and Criterion-Referenced Testing,' *Practical Assessment, Research and Evaluation*. Vol.5 No. 2 (Also available online: http://ericae.net/pare/getvn.asp?v=5&n=2)

Brown, G.I., Bull, J. and Pendlebury, M.(1997) *Assessing Student Learning in Higher Education*, Routledge, London.

Crocker, A. C. (1971) *Statistics for the Teacher,* Penguin, London.

Lloyd-Jones, R. and Bray, E. (1985) *Assessment – From Principles to Action*, Macmillan Educational Limited, London.

Palomba, C.A. and Banta, T.W. (1999) *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education,* Jossey-Bass, San Fransisco.

Wolfe, A. (1995) *Competence-based Assessment*, Open University Press, Milton Keynes.

Published by:

CAA Centre
University of Luton
Park Square
Luton
Bedfordshire
LU1 3JU

Telephone: +44 (0) 01582 489214
Fax: +44 (0) 01582 743400
Web site: http://caacentre.ac.uk

ISBN: 1-904020-01-1