Evaluating and designing assessments for medical education: the utility formula

M Chandratilake, M Davis, G Ponnamperuma

Citation

M Chandratilake, M Davis, G Ponnamperuma. *Evaluating and designing assessments for medical education: the utility formula*. The Internet Journal of Medical Education. 2009 Volume 1 Number 1.

Abstract

As assessment serves several important purposes in medical education, it is a vital element in the training of doctors. A rigorous assessment system, therefore, is an essential requirement in enhancing quality and accountability of medical education. This can be achieved by considering the utility formula which takes in to account the validity, the reliability, the educational impact, the practicability, the cost-effectiveness and the acceptability of assessment. All utility elements of a given assessment should be satisfactorily engaged for it to be psychometrically rigorous and sustainable in a particular context. This article discusses the utility elements in relation to medical assessments and introduces some useful measures for achieving acceptable utility.

INTRODUCTION

As assessment serves several important purposes in medical education, it is a vital element in the training of doctors:

- The ultimate aim of undergraduate, postgraduate and continuing medical education is to improve the health and the health care of the population. The outcomes of all medical education programmes, in general, are focused on this aim. Assessments should accurately measure the students' or trainees' progress towards or achievement of these outcomes at different levels of their training;
- Pass / fail decisions are taken and qualifications are awarded based on assessment results. Students who perform well in assessments receive good ranks, grades and prizes. On the other hand, poorly performing students may be offered support and additional training;
- As assessments drive student learning, they are a crucial component of the teaching I learning process.1,2,3,4,5 Therefore, assessment is an important mode of communicating to students what teachers value (i.e. intended outcomes of the programme);
- The assessment results should provide students with meaningful feedback on their strengths and weaknesses. At times, students use their assessment performance as a basis for career selection.
- Similarly, assessment results should provide useful feedback to other stakeholders in the educational process such as teachers and future employers.

A rigorous assessment system, therefore, is an essential requirement in enhancing quality and accountability of medical education. Quality enhancement agencies of many countries have formally emphasised the importance of credible assessment systems in medical education.[6 7] Although students can escape from poor teaching by independent learning, they cannot escape from the effects of poor assessments; they have to pass the examination.[8]

Various assessment methods are used in both undergraduate and postgraduate medical education. Assessors need to consider some essential questions when implementing assessments. Are our assessments psychometrically sound? What is their educational impact? Are the assessments with sound psychometric properties and positive educational impact feasible and cost-effective in our own setting, and acceptable to all involved in assessments? This article discusses: the contribution of different elements, namely psychometric properties, educational impact, practicability, cost-effectiveness and acceptability to the utility value of our assessments[1]; and the practical measures for improving each aspect.

PSYCHOMETRIC PROPERTIES

The assessments are psychometrically sound if they are valid and reliable. Validity is defined as the "extent to which a test measures what is intended to be measured and nothing else".[9] Reliability is a measure of the consistency and precision with which a test measures what it is supposed to assess.[9]

1. VALIDITY OF THE ASSESSMENT

Major determinants of the validity are: assessment of what is purported to be assessed; selection of suitable assessment instruments for the purpose; and adequate representation of the curriculum in the assessment material. These aspects need to be considered before the assessment is conducted (i.e. at the planning stage). After assessments are held, however, the validity of assessments may be reviewed by quantitative analysis of results.

ASSESSMENT OF WHAT IS PURPORTED TO BE ASSESSED

The assessments should assess what is intended by the curriculum. The purpose of the course (i.e. intended educational message) is demonstrated by: the time allocated to each topic in teaching; and the level of thinking and competence/performance encouraged by the course objectives. For example, in an endocrine module, the curriculum expects the students to solve clinical problems related to common endocrine disorders, which they meet at first contact level. Accordingly, more teaching time is allocated to diabetes than pheochromocytoma, as in primary care settings the presentation of patients with the former is more frequent than the latter. When clinical-problem solving is the level of competence that is required in specific curriculum, problem-based learning is used as the main method of teaching. If the assessment mostly tests factual recall about pheochromocytoma, however, the purpose of the module is not represented in the assessment. Students, no doubt, will be driven towards memorising facts rather than solving problems, and more about pheochromocytoma than diabetes. As a result, incongruence between the time devoted to teaching (more time for diabetes), and weight (more assessment content in pheochromocytoma), and level (factual recall) assessed leads to undesired student learning. Therefore, the relative weight given to each topic in assessment should be proportionate to teaching and teaching time allocated in the planned curriculum.

Factual knowledge is a prerequisite for effective problem solving.[10] However, 'in real professional practice, factual knowledge is mostly not a goal itself, but only a single aspect of solving professional problems'.[11] One of the important principles of recent curricular changes in undergraduate medical education is the promotion of higher order thinking.[12] The role of assessments in encouraging higher order thinking is vital.[13]

Bloom's taxonomy[14] categorises knowledge into six levels: recall; comprehension; application; analysis; synthesis; and evaluation. The assessment of recall and comprehension of knowledge is essential, but if only recall and comprehension are tested, lower order thinking will be promoted. In contrast, higher order thinking is encouraged by assessing the knowledge at application; analysis; synthesis; and evaluation levels. Context-free questions, i.e. questions that are not based on practical / clinical scenarios, encourage the consideration of simple answers, e.g. yes/no.[15] The promotion and assessment of higher order thinking can be achieved by introducing context-rich questions, i.e. questions based on patient, practical or clinical scenarios, for knowledge assessments[11,16](Box 1).

Figure 1

Box I — Context free and context rich questions

E.g. An assessor needs to test knowledge on urinary tract infection. It can be tested as either context-free or context-rich Single-Best-Answer MCQ formats.

Context free format

What is the most appropriate investigation of detecting urinary tract abnormalities in toddlers?

a. Antegrade pyelogram
b. CT scan
c. DMSA scan
d. KUB
e. IVU

The same options list can be followed for *context-rich format*

A 2-year-old girl who has had a febrile illness and a proven urinary tract infection on two prior occasions has been diagnosed once again with pyelonephritis.

Which is the single most appropriate investigation to perform?

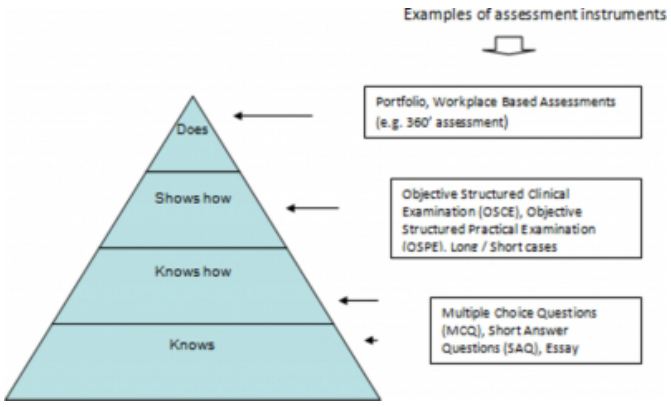
(Correct answer – c)

SUITABILITY OF ASSESSMENT INSTRUMENTS

Miller describes four levels of assessment: knows; knows how; shows how (competence); and does (performance) (Figure 1).[17] Suitability of the assessment instrument(s) can be determined by relating the objectives or outcomes assessed to the different levels of Miller's pyramid. Assessors, therefore, may require an assessment 'tool kit' rather than a single instrument to assess every thing they need to assess.

Figure 2

Figure 1 — Examples of assessment instruments for assessing different levels of Miller's pyramid



The use of multiple assessment instruments enhances both validity and reliability of results.[1] The students also perceive more satisfaction and motivation with the use of multiple assessment instruments than with the use of a single instrument.[15]

Some assessment instruments possess more than one format; e.g. single best response and extended matching items formats in Multiple Choice Questions (MCQs). The appropriate format should be chosen considering the content to be assessed, the training and experience of the assessor, and the psychometric properties (validity and reliability) of each format.

SAMPLING OF THE CURRICULUM FOR ASSESSMENT

One measure of ensuring validity is adequate sampling of the curriculum for the assessment (i.e. the assessment content should be representative of the curriculum content).[18]

a) Representativeness

As assessment drives learning,[1,4,19] the representation of each topic and each curriculum objective in assessments sends a clear educational message to the students about the topics and outcomes they should master. Therefore, the sample of curriculum content in the assessment should represent the whole curriculum and this is a primary requirement of content validity.[2,12,20]

Before assessment, the assessment contents should be plotted against the planned objectives (this is often referred to as "blueprinting").[20] In the assessment blueprint, the columns represent the course outcomes or objectives and the rows represent the teaching/learning topics. This process helps assessors sample all topics and outcomes/objectives in the assessment materials, establishing the content validity of the assessment.[2]

The number of questions focused on assessing different topics and objectives in an assessment vary in congruence with the relative emphasis given to each topic and objective in the curriculum. No topic or objective/outcome, however, should be left out, as the assessment material should be a representative sample of the course content.

b) Technical accuracy

The questions formulated to assess the sampled content should not contain technical errors. For example, a grammatically incorrect MCQ may not assess the students' knowledge of the intended topic, as the students may not understand what is being asked. Frequently observed technical flaws in relation to MCQs include: use of absolute (e.g. using must, typical in MCQ) and frequency (e.g. using often, sometimes in MCQ) terms; and spelling and grammar mistakes.[5] Technical flaws confuse students and directly affect the students' marks, reducing the validity of the assessment.[18] Therefore, they should be eliminated in constructing any type of assessment question.

Quantitative analysis of marks

Based on the performance of students, calculating difficulty and discrimination indices, and correlation of marks may provide validity evidence.

a) The difficulty and discrimination indices

The difficulty of a test item and its discrimination power (DP) could provide supportive evidence for validity of examinations.[21,22] The difficulty index (DI) is the proportion of candidates that passes a test item (e.g. single question in a single-best-answer type MCQ paper). It is calculated by dividing the number of candidates who passed the test item by the number who sat the examination. Thus a high DI (e.g. 0.9) may indicate an easy item and a low DI (e.g. 0.1) may indicate a hard item. The DP is the ability of a test item to distinguish between high and low performers. For example, to demonstrate high DI, students who are more competent in clinical skills (high performers) should score higher than the students who are less competent (low performers) in an OSCE station designed for the assessment of history taking skills (test item). In calculating the DP of a test item, the candidates are ranked by descending order of their marks for the whole examination. The number of candidates in upper third and lower third of the list who correctly answered the

item is calculated. The proportion of candidates who have correctly answered the item in the lower third is subtracted from the proportion of their counterparts in the upper third. The DP should be positive. A negative DP requires investigation.

Most of the medical undergraduate assessments have either criterion-referenced components (passing or failing is based on the standard achieved) or norm-referenced components (passing a percentage of candidates after ranking them based on their performance), or both. If the DI of a test item is low, the test setter may be able to observe that: the item assesses content outside the curriculum; the teaching / learning of the content area has taken place ineffectively; the item is technically flawed; or the students have not learnt the topic represented by the item.[23] Obviously, DI of a norm-referenced examination should be high in order to discriminate between high and low performers. Although the intention of a criterion-referenced test is not discrimination between high and low performers, the discrimination index still has a value.[23] An item with a negative discrimination index (i.e. more low performers answering correctly than high performers) usually denotes a technical flaw, a mistake (e.g. wrong answer), or mis-key.

A DP near to zero together with a high DI in a criterion-referenced test may indicate the effectiveness of the teaching / learning of the content area related to the item (i.e. both high and low performers have mastered the topic).

b) Correlation coefficients

In an examination, assessors may use different assessment instruments to assess different levels of Miller's pyramid. Supportive evidence for the use of an appropriate instrument for a specified level may be obtained by correlating students' marks (using a Pearson correlation) for different assessment instruments. The correlation of marks of two instruments which assess the same level (e.g. MCQ and SAQ assessing 'knows' level) should be higher than the correlation coefficient of marks of instruments assessing different levels (e.g. MCQ assessing knows and OSCE assessing shows how).

2. THE RELIABILITY OF ASSESSMENT RESULTS

Reliability indicates the ability of an assessment result to be replicated given the same or similar conditions. Assessment is a measurement. As in all measurements, assessment results may not be always consistent (i.e. reliable) due to measurement errors.[23,24] Exam questions and examiners either individually or in combination may contribute to measurement error.

The reliability of assessment results can be estimated using Classical Test Theory (CTT) and Generalisability Theory (GT).[24] Both these theories examine the variance of scores.

ESTIMATING RELIABILITY USING CTT

A widely used reliability measure that uses the CTT as its basis is the Alpha coefficient (AC). AC is a value between zero and one (0-1), which can be calculated using statistical software like SPSS. For example, an AC of 0.8 means that the reproducibility is 80% and the total measurement error is 20%. However, CTT cannot be used to identify the sources of error (i.e. what contributes to the 20% of error in the example above) and their relative magnitudes, as in CTT the error is identified as a single entity.[24]

ESTIMATING RELIABILITY USING GT

In GT, the G-coefficient (value between 0 – 1) also indicates the reliability of results. Different sources (e.g. items/stations: raters,) can be responsible for the error component. The assessors would want to know not only the magnitude of the overall error but also the source(s) of error and their individual magnitude.[24] GT can be used to identify the sources of error and quantify their contribution to the total error, as GT analyses the variance. [24] It also gives provisions to identify how to minimize the error and what is needed to achieve results that are sufficiently reliable. G-coefficient can be calculated using statistical software packages such as GENOVA.

In both CTT and GT, a value of more than 0.8 is considered acceptable reliability. However, in high stake examinations, some assessment authorities (e.g. Postgraduate Medical Education and Training Board) recommend the achievement of 0.9. The evidence of reliability estimated by these statistical methods, however, should always be interpreted against the backdrop of the validity of the assessment. The reliability values have no meaning with poor validity.

EDUCATIONAL IMPACT

The educational message, i.e. the educationally desirable direction that teachers expect the students to follow, conveyed to the student by the assessment is referred to as educational impact. Citing many authors, van der Vleuten points out that the "assessment programme has tremendous impact on learners and students do whatever they are tested

on and are not likely to do what they are not tested on".[1] Although more time is allocated for learning clinical skills in wards, if students are assessed on recalling facts using a MCQ examination, they have a propensity to read books and notes in a library. Conversely, they will learn clinical skills, spending more time in clinical skills centres or wards, if their clinical skills are assessed using an OSCE.[25] Therefore, the assessments should reflect the educationally desirable direction expressed in the curriculum outcomes.

It is true that high validity, reliability and positive educational impact enhance the rigor of assessments. However, the psychometric properties and educational impact of assessments should be balanced with the practicability and the cost-effectiveness of using an assessment instrument in a given context, and its acceptability to people involved in the assessment process (e.g. exam setters, examiners, examinees).[1]

PRACTICABILITY

Strategies to improve validity (e.g. the use of the OSCE to assess skills) and reliability (e.g. testing with as many observers and cases or situations as possible) may not be feasible for many reasons. Ram et al,[26] in their evaluation of using video observations for the assessment of general practitioners, identified that feasibility issues were related to the cost, availability of equipment, time, recruitment of patients and assessors, and manpower necessary to develop infrastructure. Psychometric rigor may be very important in some high stake assessments (e.g. final year undergraduate examination, national board examinations). But feasibility may be equally important for iterative in-training assessments.[27] Therefore, at times, a compromise of psychometric rigor, to a certain extent, may be necessary for the assessment system to be practicable. For example, the number of summative examinations can be reduced when the number of formative examinations is increased provided that the formative exams follow the same format as the summative examinations. Because formative assessments may not warrant such strict psychometric rigor as summative assessments, this approach may help mobilise the existing resources and make psychometrically rigourous summative examinations practicable.

COST-EFFECTIVENESS

In practice, the cost of assessment is a compromise between the information elicited and the resources required by the examination.[1] However, "investing in assessment is investing in teaching and learning, as assessment drives learning" and perceived resource-intensive assessment methods may turn out to be rewarding in terms of return on cost in practice.[1] Therefore, the cost-effectiveness of assessment, evaluating the benefits of a particular assessment against its cost, seems more important than the cost alone. For example, a one-from-five MCQ test may be the cheapest mode of valid and reliable assessment of 'knows' and 'knows how' levels of Millers pyramid.[17] However, it is not suitable for assessing competence or performance. A portfolio assessment, which is costly compared to a MCQ test, may be the cost-effective method of assessing performance credibly.

ACCEPTABILITY

A test may be acceptable to some of those dealing with it and not to others.[9] The beliefs and attitudes of both examiners and examinees towards assessment may not always be in line with the research and empirical evidence. Therefore, certain assessments may not be acceptable to all.[1] Provision of necessary information and willingness to compromise may increase the commitment of both examiners and examinees.[27] However, if the beliefs, opinions and attitudes of both examiners and examinees are not considered in choosing and designing assessments, the survival of assessment procedure is threatened.[1] For example, strongly structured assessments may not be acceptable to examiners, as the examiners have little opportunity to exploit their expertise to vary questioning from candidate to candidate.[1] Therefore, a compromise between an acceptable degree of freedom for such issues and the exam structure enhances the sustainability of the assessment system.[1] For example, in an OSCE, a checklist can be used together with a global rating where the examiners can express their overall judgment on candidates, enhancing both psychometric properties and acceptability.

THE UTILITY FORMULA

Combining the utility elements: validity; reliability; educational impact; cost-effectiveness; and acceptability, van der Vleuten[1] introduced a utility formula.

Utility = R x V x E x A x C

(R= Reliability, V= Validity, EI= Educational impact, A= Acceptability, C= Cost)

However, feasibility has also been shown to be important for the utility of an assessment.[27] On the other hand, in practice, cost-effectiveness of assessment may be a better determinant of its utility than the cost alone. Therefore, we have found it

helpful to modify this formula to include practicability and cost-effectiveness.

Utility = R x V x EI x P x A x CE

(R= Reliability, V= Validity, EI= Educational impact, P = Practicability, A= Acceptability, CI= Cost-effectiveness)

According to this utility formula, the utility value of the assessment becomes null and void, if any of the utility factors becomes zero.

CONCLUSION

Good assessment practices in medical training, at all levels, enhance both quality and accountability of medical education. The utility of assessments depends on reliability, validity, educational impact, acceptability, cost-effectiveness and practicability. Although the rigor of assessments is determined by validity, reliability and the educational impact, measures employed in achieving rigor should be balanced against the practicability and cost-effectiveness of using an assessment system in a particular setting, and the acceptability of assessments to their stake holders.

References

1. van der Vleuten C: The assessment of professional competence:
developments, research and practical implications. Advances in Health Science Education; 1996;1: 41 – 67.
2. Fowell S, Southgate L, Bligh J: Evaluating assessment: the missing link? Medical Education; 1999; 33: 276 – 28.
3. Harden R. AMEE Guide 21: curriculum mapping: a tool for transparent and authentic teaching and learning. Medical Teacher; 2001; 23: 123 – 137.
4. Eraut M: A wider perspective on assessment. Medical Education; 2004; 38: 800 – 804.
5. Boud D: Assessment and learning: contradictory or complementary? In: Knight P. eds. Assessment for Learning in Higher Education; London; Kogan; 1995: 25 – 48.
6. Postgraduate Medical Education and Training Board: Standards for Curricula and Assessment Systems; London; PMETB; 2008: 6.
7. Committee of Vice-chancellors and Directors: Quality Assurance Handbook for Sri Lankan Universities. Colombo: University Grants Commission; 2002: 105.
8. Case S, Swanson D: Constructing written test questions for basic and clinical sciences. 3rd ed. Philadelphia; National Board of Medical Examiners; 2002: 26.
9. Lowry S: Medical Education; London; BMJ books; 1993;46 – 47.
10. Hager P, Gonczi A. What is competence? Medical Teacher; 1996; 18: 15 – 18.
11. Schuwrith L, van der Vleuten C: Different written assessment methods: what can be said about their strengths and weaknesses. Medical Education; 2004; 38: 974 — 979.
12. Spencer J: Learner-centered approaches in medical education British Medical Journal; 1999; 318: 1280 – 1283.
13. Wood D: Evaluating the outcomes of undergraduate medical education. Medical Education; 2003; 37: 580 – 581.
14. Bloom S, Hastings T, Modays J: Handbook of Formative and Summative Evaluation of Students Learning; New York; McGraw Hill; 1971: 103.
15. Scale F, Chapman J, Davey C: The influence of assessments on the students motivation to learn in a therapy degree course. Medical Education; 2000; 34: 614 - 621.
16. Des Marchas J, Vu V: Developing and evaluating the student assessment system in the preclinical problem-based curriculum at Sherbrook. Academic Medicine; 1996; 71: 274 – 281.
17. Miller J: The assessment of clinical skills, competence, performance. Academic Medicine; 1990; 65: s63 — s67.
18. Bridge D, Musial J, Frank R, Roe T, Sawilowsky S: Measurement practices: methods tom developing content valid student examinations. Medical Teacher; 2003; 25: 414 – 421
19. van der Vlouten C, Schuwirth L: Assessing professional competence. Medical Education; 2005; 39: 309 - 317.
20. Wass V, van der Vleuten C, Shatzer J, Jones R: Assessment of clinical competence. Lancet; 2001; 357: 945 – 949
21. Haynes S, Richard D, Kubany E: Content validity in psychological assessment: a functional approach to concepts and methods. Psychological Assessment; 1995; 7: 238 – 247.
22. Southgate L, Cox J, David I, Hatch D, Howes A, et al: The assessment of poorly performing doctors: the development of the
assessment programmes for the General Medical Councils Performance Procedures. Medical Education; 2001; 35 (sl): 28.
23. Gronlund N: Reliability and other desired characteristics. Measuring and evaluating in teaching. 3rd ed; London; Collier Macmillan; 1976: 105 – 135.
24. Boulet J: Generalizability theory: Basis. In: Everitt S, Howell C. eds. Encyclopedia of Statistics in Behavioural Science; 2nd ed; Chichester; John Wiley & Sons; 2005: 704 – 711.
25. Malik L: The attitudes of medical students to the objective structured practical examination. Medical Education; 1988; 22: 40 – 46.
26. Ram P, Grol R, Rethans J, Schouten B, van der Vleuten C, Kester A: Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. Medical education; 1999; 33: 447 – 454.
27. Crossley J, Humphris G, Jolly B: Assessing health professionals. Medical Education; 2002;36: 800–804.

Author Information

M. N. Chandratilake
Research Officer, Centre for Medical Education, University of Dundee

M. H. Davis
Director, Centre for Medical Education, University of Dundee

G. Ponnamperuma
Senior Lecturer, Development and Research Centre, University of Colombo