# Principles of Sound Assessment Practice in Health Professions Education

**Craig Peck***

*Clinical Supervisor/Lecturer, Department of Conservative/Restorative Dentistry, UWC Dental Faculty, Tygerberg Oral Health Centre, University of the Western Cape, Bellville, Cape Town, South Africa*

**\*Corresponding Author:** Craig Peck, Clinical Supervisor/Lecturer, Department of Conservative/Restorative Dentistry, UWC Dental Faculty, Tygerberg Oral Health Centre, University of the Western Cape, Bellville, Cape Town, South Africa.

## Introduction

In the broad sense, assessment is a systematic method of obtaining information or sampling data about teaching and learning in order to make specific inferences about characteristics which reflect student learning and skills acquisition. Assessment in Health Professions Education is vital as it impacts on healthcare delivery outcomes for patients and must be based on sound research evidence [1,2]. For any assessment method to be successfully used as an instrument of competency measurement, standard setting and blue printing must be done in advance [3].

The use of assessment as an instrument for educational measurement is a complex process which ultimately searches for consensus between application and practice of assessment methods. Such methods should always be characterised as sound, defensible, accurate and meaningful in their intent of what and how they are measuring student learning outcomes. Assessment is intrinsically linked to the curriculum itself and should be an explicit reflection of what the pre-determined course objectives are [1,4,5]. It is well accepted that assessment drives learning and that students instinctively learn what is being assessed. This fact provides teachers with a very powerful tool which could drive student learning in the direction of the course objectives [1,3,4].

As no one method of assessment is able to test all competencies equally well, it is imperative that a variety of the most appropriate assessment instruments be selected which are aligned with the competency being tested [1,2,4]. Educators must therefore know on what level they are assessing students – whether it is a more verbal-type of knowledge as corresponding to the two lower tiers of Miller's Pyramid, or a doing-type of knowledge as indicated by the two upper tiers of Miller's Pyramid [3,6].

Whether assessment is directed at a cognitive level (such as with written exams, multiple choice questions and orals), observational methods (such as clinical assessment) or performance exams (such as standardised patient exams), all aspects of sound assessment practice must be given full consideration. The chief principles which underpin such practices include what the purpose of assessment is, validity, reliability, educational impact, acceptability, fairness and feasibility. The degree to which these principles and indicators of sound assessment are met by various assessment methods, vary greatly and hence the assessment instruments used must be specified by their intended purpose [4].

This guideline serves to elucidate the meaning, application and use of these key determinants of sound assessment practice as a way of clarifying their relevance and importance for assessment in general. The intent is to provide colleagues with a better understanding of the complexity and value of sound assessment as a useful instrument for and of learning, in order for them to be able to apply this to both theoretical and practical/clinical assessments. It hopefully illuminates what measures can be implemented by educators to ensure that sound assessment practices are adhered to.

## Principles of Sound Assessment

### Purpose of Assessment

When assessment is used as an educational tool, it can drive student learning [1,3]. Assessment is a way to monitor student progress, with the intent of determining competency by evaluating its success against pre-determined criteria [7]. Assessment is crucial for student feedback – to students as well as to the educator – which ushers reflective thought and practice from both students and teachers [8]. There is no perfect assessment and various assessment techniques should be applied to determine which tools are more optimal for the specific objectives and which facilitate greater learning outcomes [3,4]. This means there should be constructive alignment between teaching and learning activities (TLA's), assessments and learning outcomes [5,9]. Any successful assessment plan should be done on the principle of "fit-for purpose" [2].

Assessment can also be defined as the process of gathering data about teaching and learning and once data is collected, one can more accurately evaluate student performance [1,2]. Hence, assessment is not a once-off event but is crucial for quality improvement and for greater assurance of educational efforts, which represents a continuous cycle of assessment [2].

Factors to consider when assessment tools are devised and implemented include what their purpose is - that is formative/developmental or 'for learning', and/or summative/judgemental or 'of learning' [1,2,4,7]. The main purpose of formative assessment is to provide useful feedback to students regarding their strengths and weaknesses in terms of their performance and with respect to meeting the learning objectives successfully [2,4,10]. As this takes places during and throughout the course of study, students have the opportunity of evaluating and considering what they understand well and those aspects which require more attention. It also gives teachers a better indication of which aspects of the content need more attention in terms of TLA's, which need more explicit direction and possibly more time made available to areas which are felt to be less understood by students [4,10]. Examples include short weekly quizzes, feedback and reflection from teachers after any assignment or test and is usually associated with lower-stakes assessments [1].

Summative assessment implies the summation of achievement within a course, with greater emphasis on the final measurement of achievement (such as a final exam). Its primary purpose being to evaluate and measure what students know based on the specific assessment employed [2,4,10]. Although feedback can form part of summative assessment tools, it is not its main purpose and summative assessment is usually associated with higher-stakes assessments [1,11].

With the acquisition of skills, it is reported that students learn best when feedback is given timeously and effectively incorporated as part of the academic program. This improves learning outcomes greatly [12]. The value of formative as well as summative assessment of dental students' clinical skills is undisputed in dental education and is essential for stimulating and promoting student interest through active involvement [12,13]. Formative assessment is reported to encourage a deeper approach to learning and reflective practice from students [14].

It is important that educators must be aware of on what level assessment is being done according to Blooms Taxonomy or Miller's Pyramid [1,3]. Assessment must be specific to each level or tier educators are evaluating, which represent various levels of skills development (remember, understand, apply, analyse, evaluate and create for Bloom's Taxonomy or knowledge, competence, performance and action for Miller's Pyramid [6,15].

### Reliability

Reliability is analogous with consistency of assessment and pertains to the degree of reproducibility of results [2,3,16,17]. There is an inherent relationship between reliability and objectivity, whereby consistency of assessment outcomes determine the degree of reliability of assessment methods used [16]. A general rule applicable to reliability, is "more is better". This principle underpins ways in

which reliability can be improved. These include using varied assessment [18], employing more examiners and increasing consistency between examiners with standardised questions, check-lists and rubrics [7]. Improving consistencies between examiners and students and increasing the level of objectivity of assessments would similarly result in increased reliability [2].

Other factors which influence reliability and should be given explicit attention to, include ensuring that test environments are safe, consistent and educationally conducive spaces [9,19]. Sufficient time should be allocated for testing [20] and educators must strive for broad sampling of cases/content, like in multi-case testing [2,3,17].

Reliability can be evaluated in two main ways - test-re-test reliability or parallel-test reliability. The former evaluates reliability after re-testing the same candidate with the same test at a different occasion, while the parallel-test compares assessment outcomes or scores in the same student using a different set of questions of equivalent nature and value. This is particularly useful in identifying if the specific assessment method is actually testing the intended content or domain appropriately and accurately [2]. Large variations in assessment scores after re-testing would reflect low reliability and would render the interpretation of assessment results less meaningful [18].

Two important aspects of reliability include inter-rater and inter-case reliability. Inter-rater reliability measures the consistency of assessment between examiners, where more examiners would result is a higher level of reliability. Inter-case reliability remains one of the most important aspects in clinical competency testing, as not all students perform equally well in all assessment tasks [3]. This explains why broad sampling with multi-case testing is essential for sound assessment practices as it greatly improves the degree of reliability of testing method [2].

Reliability is especially important in high-stakes assessments (such as exiting and licencing examinations) and one should always ask oneself whether or not the assessment produces stable and consistent results, which are comparable to each other and are a true reflection of what is being tested [1,11].

## Validity

Validity is the extent to which a specific evaluation or assessment instrument is valid. That is, does it measure what it is intending to? [2,3,17] Validity is not always easy to measure and therefore one should collect evidence from different sources and perspectives of assessment to be able to demonstrate evidence thereof within assessment methods [2]. To be able to identify validity, one must first specify what is being assessed and ensure that there is alignment between assessments and learning objectives [3,5]. It is important to remember that no single assessment instrument is capable of measuring all levels of competencies and learning objectives equally and with the same level of specificity [3]. Hence, assessment methods should be tailor-made and targeted at the desired level of competency being assessed, as specified by Bloom's Taxonomy and/or Miller's Pyramid [1,3,6,15].

Miller's Pyramid of competencies provides an effective and usable model on which valid assessment methods can be based as they relate directly to the level being tested – be it "knows", "knows how", "shows how" or "does" [1,3,6]. Bloom's Taxonomy is valuable in identifying what the pre-defined learning objectives are and on what level the assessment is designed and specified for – be it for lower-order cognitive function (such as listing, stating or identifying) or higher-order functioning (such as comparing, analysing, evaluating, critiquing and hypothesising) [15].

There are four main categories of validity, namely content (direct) validity, construct (indirect) validity, predictive validity and consequential validity (educational impact). Content validity defines validity in terms of the homogenous spread of assessment over the entire testable material/content [2].

An important aspect of content validity is blue printing, which ensures that the test content is planned against, and aligned with, the learning objectives. This ultimately ensures that assessment validates the explicitly set objectives of the curriculum by being embedded

therein. Blue printing ensures that assessments are aligned with competencies which students need to learn and with TLA's being employed for effective transfer of knowledge and skills. In this way, blue printing forms a matrix for assessment, by deciding the details of the spread and grading of questions from various categories and subsections within the content tested, as well as determining the relevance of the items being tested [3,20]. Blue printing is as much about informing students how question papers and assessments are constructed as they are moderation tools for education.

Construct validity makes reference to the ability and success of assessment instruments to be able to distinguish between student competencies using a single assessment method. That is, can the specific assessment differentiate between students with higher and lower competencies successfully? [2] Predictive validity is simply the degree to which test scores obtained from assessment are able to predict the future performance or competency of candidates [3].

Evidence for validity can be found by looking at factors such as the usability of tests and the relationship between the test and the candidate being assessment. Furthermore, a high validity would be represented by a weak correlation between assessments which measure different constructs and strong correlations between assessments which are intended to measure the same constructs [3]. The accuracy of assessment scores determine the degree to which one can make inferences about the appropriateness and adequacy of assessment and how well this correlates to pre-determined criteria of learning outcomes and objectives [2].

There are several ways to ensure greater validity in assessment. By properly defining the purpose of the evaluation tools and selecting them appropriately based on what is being measured (as specified in Miller's Pyramid and Bloom's Taxonomy), would increase the validity of the assessment [1,2,3,6]. Ensuring that assessment is done comprehensively in terms of the content covered and always assessing according to set standards and pre-defined criteria, would similarly improve the level of validity [3].

In this regard, it is pertinent to mention the importance of using criterion-referencing in standard setting and not norm-referencing. Criterion-referencing allows for a clearly defined standard to be set for each assessment, which accounts for the variances in student abilities while greatly being influenced by the level of difficulty of the assessment. Norm-referencing is not an acceptable method of standard setting to assess clinical competency, as it compares and relates candidate test scores to each other and not to a pre-determined minimum criteria of clinical competence [1]. Naturally, norm-referencing would have a negative impact on patient care, as the standard and assessment outcomes would be greatly dependant on the performance of candidate's peers.

Validity is linked to reliability, in as much as any assessment which is not reliable cannot be valid.[3] Below is a graphic representation of the relationship between reliability and validity:
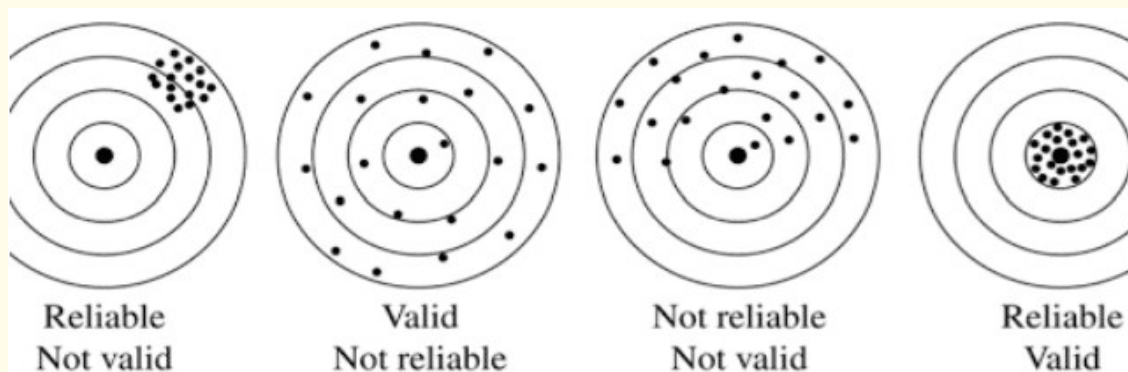
## Educational Impact

Educational impact can be defined as "assessment for learning" and it also known as consequential validity. It implies using appropriately selected assessment instruments for learning as assessment itself influences student learning behaviour [2,17]. The type and nature of assessment will define student attitudes and behavioural consequences to learning due to the external motivating effect that assessment has on driving and specifying student learning. Students regard assessment as a defining aspect of the curriculum and objectives, as it specifies to students what is to be learnt [2,4,5]. Any form of assessment, be it formative or summative, will directly influence learning [4,21].

Assessment affects student learning in several ways, the most obvious of which is the influence that the content itself has. Students perceive that the most important aspects of the course content will be assessed which implicitly directs their studying efforts to those aspects which are given more attention during assessments. By teachers repeating and emphasising the most important issues of the content, improves the level of educational impact of assessment [4,19,22].

It is also important to pay attention to scheduling of assessment tasks in a bid to avoid clustering exams/tests close together. This would result in strategic selection of content studied by students and inadequate preparation allowed for assessments [2]. Poor scheduling would similarly increase student stress, anxiety and cramming of knowledge which would foster a more surface approach to learning as represented by poorer assessment outcomes [5,23].

The format of assessment is instrumental in facilitating greater educational impact. A variety of assessment formats should be used to assess different skills so that students do not intentionally prepare for one type of assessment method [2,24].

A further factor which impacts on the degree of educational impact of assessment, is the academic program itself [17]. The structure of the program defines academic success and therefore credit scores, weighting of scores and minimum criteria set by the initial standard setting process are all important influences of and for student learning. In this way, it is accepted that continuous assessment will drive continuous learning with a deeper approach being employed by students, whereas pin-point assessments would result in the phenomenon of cramming and a more surface approach to learning [2].

Again, the importance of ensuring constructive alignment between course objectives and assessment methods is vital for improving educational impact [4,5]. Educators must be cognisant of the various psychological, social, emotional and affective needs of student and fully aware of how these influence student learning and, by extrapolation, also assessment outcomes [23,25].

## Acceptability

Any method or instrument of assessment can only be regarded as valid if it is able to be used in the proper, intended manner. This implies that assessment must be acceptable for use – by educators and stakeholders, but for students and patients as well. Often a compromise has to be made between what is acceptable assessment and what is deemed to be educationally and diagnostically superior assessment [2].

## Fairness

This makes reference to the degree of transparency of assessment based on known outcomes and refers to the degree of objectivity of assessment tools [7]. Fairness ensures that assessments do not disadvantage students by possibly making demands which are deemed inappropriate or unsuitable for candidates. It also represents the right of students to be able to view and appeal their assessment results [8].

Fairness in assessment reflects the degree of multiply inclusiveness within education and facilitates greater alliance between students, teachers and the subject matter by echoing sound and defensible assessment practices [9,23]. There must be a rational system or plan to connect the content tested to the knowledge, skills and abilities which are deemed important learning outcomes [1]. Using varied as-

sessment methods addresses the great diversity of students' preferences for different styles of assessment, which improves fairness and maximises on the value of assessment as an educational measuring instrument [1,18].

Fairness implies assessment being free from pre-determined judgments, prejudices and biases (such as race, religion, culture and gender) and ensures that all students have the necessary freedom in which to demonstrate the degree to which they were able to meet the learning objectives [26]. Aspects such as workload, scheduling of assessments, the complexity/difficulty of assessment tasks and providing accurate and adequate feedback on assessments affect the degree to which an assessment is considered to fair or unfair [20]. Unfair assessment practices would impact negatively on the validity of assessment - such as biasness, in which subjective opinion is given precedence over objectivity [27].

Various means can be employed by educators to foster greater fairness. These include creating a learning environment of trust [5,9,19,28,29], ensuring equitable levels of student comprehension, language and understanding [8] and to ensure that all students are fully aware of the learning outcomes [2,5]. By employing varied methods of assessment [18] and being aware of on what level one is executing assessment on (according to Bloom's Taxonomy and/or Miller's Pyramid) would similarly increase the fairness of the assessment method used [1,3,5,6].

## Feasibility

Feasibility refers to the ease at which an assessment is executable or doable in reality. For the most part, it is a reflection of available resources, infra-structure, staff and budget required to design, implement and monitor assessment effectively in order to determine the best fit for the circumstances and educational needs. A large part of feasibility is determined by the cost-effectiveness of the assessment method [2]. In order for an assessment to be regarded as cost-effective, the goals of the specific assessment must be explicitly described in terms of what is assessed and how it is assessed [3].

Several factors impinge on cost-effectiveness, which include previously held misconceptions of traditional assessments and a lack of research-based evidence on which assessment should be grounded. Issues such as poor infra-structure, lack of adequate resources (financial and human), misutilisation of appropriate staff during assessment and poor collaborative efforts all detract from a cost-effective assessment process [2].

For optimal collaboration in assessment, one should give priority to proper planning of the assessment tasks and ensure that there is sufficient buy-in and commitment from all parties involved in the process. It may even be necessary to secure a degree of pre-investment for the assessment process to reach fruition. It is always advisable to first complete a needs analysis and then select the most appropriate assessment which best suits the circumstances and specific requirements. This reduces the possibility of the assessment not achieving its goals in terms of reliability, validity, fairness and educational impact [2].

There tends to be an indirectly proportional relationship between feasibility and reliability, as factors which improve reliability often result in a decrease in feasibility of assessment by negatively impacting on cost-effectiveness [2].

## Conclusion

Although assessment offers information regarding student progress and performance and sets acceptable standards of practice, assessment instruments themselves can be considered as "memory boxes" of how successful they were in assessing student learning outcomes – be it factual and/or clinical. Hence, assessment must be constantly reviewed and adapted accordingly in order to meet the ever-changing demands of the curriculum [4]. This enables teachers to get a better indication of which aspects of the curriculum are taught well and which parts may perhaps require more explicit attention by possibly adjusting TLA's or investing more time on certain aspects of the content [1,2].

Assessment is needed to provide meaningful feedback (to students as well as to teachers) with the intention of monitoring the success of the competency of the assessment method used, as only that which is measured can be properly evaluated, compared and improved upon. In this way, assessment becomes important for quality improvement and assurance of education and should never be consider a once-off event. The assessment process should always be considered as a continuous process of collecting, analysing and reflecting on evidence in order to make conversant and consistent judgements which improve student learning, without the limitation or capacity of disadvantaging any student.

Employing tiered assessment of both knowledge and skills helps to identify weaknesses within the curriculum and provides educational institutions with a unified assessment of the overall level of student competencies [30]. By giving explicit attention to all aspects of sound assessment principles would ensure that the assessment methods are fit-for-purpose in terms of their appropriateness and effectiveness in not only facilitating student learning, but also their dependability as instruments of educational measurement [2]. For medical education, this has tremendous implications by ensuring and fossilising safe and effective patient healthcare delivery.

## Bibliography

1. Downing SM and Yudkowsky R. "Introduction to assessment in Health Professions. Assessment in Health Professions Education". Routledge, New York (2009): 1-20.

2. Schuwirth LWT and van der Vleuten CPM. "How to design a useful test: the principles of assessment". In: Stanwick T. Understanding Medical Education: Evidence, Theory and Practice, Second Edition. Chichester, Wiley-Blackwell, John Wiley & Sons Ltd (2010): 243-254.

3. Wass V., *et al*. "Assessment of clinical competence". The Lancet 357.9260 (2001): 945-949.

4. Boud D. "Assessment and the Promotion of Academic Values". *Studies in Higher Education* 15.1 (1990): 101-111.

5. Biggs J. "What the Student Does: teaching and enhanced learning". *Higher Education Research and Development* 18.1 (1999): 57-75.

6. Miller G. "Educational theories you must know". Miller's Pyramid. St. Emlyn's (1990).

7. Churches A. "A guide to formative and summative assessment and rubric development. 21st century fluency project (n.d.).

8. Hays RB., *et al*. "Twelve tips for increasing the defensibility of assessment decisions". *Medical Teacher* 37.5 (2015): 433-436.

9. Entwistle NJ and Petersen ER. "Conceptions of learning and knowledge in higher education: Relationships with study behaviour and influences of learning environments". *International Journal of Education Research* 41.6 (2004): 407-428.

10. Black P and Wiliam D. "Assessment and Classroom Learning". *Assessment in Education: Principles, Policy and Practice* 5.1 (1998): 7-74.

11. Jonsson A and Svingby G. "The use of scoring rubrics: Reliability, validity and educational consequences". *Educational Research Review* 2.2 (2007): 130-144.

12. Rolfe IE and Sanson-Fisher RW. "Translating learning principles into practice: a new strategy of learning clinical skills". *Medical Education* 36.4 (2002): 345-352.

13. Taylor CL., *et al*. "Assessing the Clinical Skills of Dental Students: A Review of the Literature". *Journal of Education and Learning* 2.1 (2013): 20-31.

14. Rushton A. "Formative assessment: a key to deep learning?" *Medical Teacher* 27.6 (2005): 509-513.

15. Bloom BS. "Bloom's Taxonomy of educational objectives: The classification of educational goals" (1956).

16. van der Vleuten CPM. "The assessment of professional competence: developments, research and practical implications". *Advances in Health Sciences Education* 1.1 (1996): 41-67.

17. van der Vleuten CPM and Schuwirth LWT. "Assessing professional competence: from methods to programs". *Medical Education* 39.3 (2005): 309-317.

18. Epstein RM. "Assessment in Medical Education". *New England Journal of Medicine* 326.4 (2007): 387-396.

19. Gravett S. "Adult Learning. Designing and implementing learning events – A dialogic approach". Second Edition, (2004): 31-39.

20. Patil SY., *et al*. "Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology". *International Journal of Applied Basic Medicine Research* 5.1 (2015): S76-S79.

21. Cilliers FJ. "Is assessment good for learning or learning good for assessment? A. Both? B. Neither? C. It depends?" *Perspectives on Medical Education* 4.6 (2015): 280-281.

22. Newble DI and Jaeger K. "The effect of assessments and examinations on the learning of medical students". *Medical Education* 17.3 (1983): 165-171.

23. Haggis T. "Constructing Images of Ourselves? A Critical Investigation into 'Approaches to Learning' Research in Higher Education". *British Educational Research Journal* 29.1 (2003): 89-104.

24. Stalenhoef-Halling BF., *et al*. "A new approach to assessing clinical problem-solving skills by written examination: conceptual basis and initial pilot test results, 1990 as cited by Schuwirth LWT, van der Vleuten CPM. How to design a useful test: the principles of assessment". In: Stanwick T. Understanding Medical Education: Evidence, Theory and Practice, Second Edition. Chichester, Wiley-Blackwell, John Wiley & Sons, Ltd, (2010): 243-254.

25. Rowe M. "Beyond the lecture: Teaching for professional development". *African Journal of Health Professions Education* 8.2 (2016): 208-210.

26. Stupart D., *et al*. "Does examiner bias in undergraduate oral and clinical surgery examinations occur?" *South African Medical Journal* 98.10 (2008): 805-807.

27. McManus IC., *et al*. "Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP (UK) PACES and nPACES examinations". *BMC Medical Education* 13 (2013): 103.

28. Vermunt JD and Verloop N. "Congruence and friction between teaching and learning". *Learning and Instruction* 9.3 (1999): 257-280.

29. Donald J. "Student Learning and Intellectual Development. Tomorrow's Professor Postings". Stanford University (2008): 1-4.

30. Doroshenko OM., *et al*. "Competency Assessment of Clinical Approach During Primary Specialization in Dentistry: Integrated Licensing Exam". *Intermedical Journal* 1.9 (2017): 10-15.