CHAPTER 1

Review of the assessment literature

This literature review provides a broad overview of some of the most important advances in medical education assessment practices over the past three decades.[1-7] For the purpose of this thesis these advances can be broadly clustered into two overarching themes: (1) the purpose of assessment and (2) the utility of assessment. The purposes of assessment, as addressed in this thesis, are: (1) measurement of student achievement in order to make judgement decisions, e.g. selection, placement, promotion to next year of study, graduation or certification, (2) facilitation of student learning, and (3) improvement of the quality of training programmes by initiating and sustaining curriculum change.[2-4,8] The utility of assessment, as referred to in this thesis, is understood to mean the perceived usefulness or fitness for purpose of an assessment process. The key parameters that determine assessment utility include: (1) parameters indicating the rigour of assessment – validity and reliability, and (2) parameters indicating the practicality of assessment – feasibility, cost and resource requirements.[4,9]

Most of these advances have been implemented in, and have impacted upon, medical training programmes in developed countries, and, much less is known about assessment practice advances in resource-constrained settings typical of developing world countries. This was highlighted by a recent publication in which the paucity of data from the developing world, particularly sub-Saharan Africa, was apparent. Tutarel reviewed the geographical distribution of all papers published in two international peer-reviewed prestigious medical education journals, Academic Medicine (USA-based) and Medical Education (UK-based) between 1995 and 2000.[10] He found that only 15 of 2 953 published articles were from sub-Saharan Africa. At least 80% of these papers, accounting for only 0.5% of all publications reviewed over the five-year period, came from South Africa. The reasons for this phenomenon are beyond the focus of this dissertation, but the findings highlight the lack of published data regarding medical education practices in the developing world. This thesis, therefore, specifically focuses on advances in medical education practices, assessment in particular, implemented in medical training programmes in South Africa. A key purpose of the publications in this thesis is to inform the broader medical education community about the implementation of widely endorsed assessment practice advances in a developing country challenged by significant resource constraints.

In the first section of the literature review I explore issues dealing with the measurement of clinical competence. I initiate the discussion by providing a working definition of

professional competence and clarifying the relationship between observed performance and implied competence. Thereafter, a user-friendly taxonomy of the assessment of competence is graphically depicted and illustrated using a simple clinical example of hypertension. This is followed by a brief discussion of four major factors that have led to the plethora of assessment methods currently in use: (1) the dual purpose of student assessment, (2) the hierarchical nature of competence, (3) the psychometric adequacy of assessment instruments, and (4) the educational and vocational alignment of assessment practices. Using the taxonomy referred to earlier, some examples of written tests, in vitro (simulated clinical environment) and in vivo (authentic clinical workplace) performance assessment tools are listed. Finally, this section highlights the need for multi-component (composite) assessment strategies. Since no one assessment tool comprehensively assesses the many outcome competencies of medical training programmes, it has become obvious that it is necessary to craft assessment packages that broadly address programme outcome competencies using a variety of testing instruments.

The second major topic addressed in the literature review is the use of assessment to facilitate student learning. Three specific strategies are targeted for discussion. Firstly, the importance of educational concordance is highlighted again because of its profound influence on student learning, Secondly, feedback, the key component of formative assessment used to guide and direct student learning is described. A few studies evaluating the educational impact of formative assessment are briefly reviewed. Secondly, summative assessment practices, for judgement purposes, are known to powerfully influence student learning behaviour. Two well known examples from the medical literature are described. The literature is increasingly advocating the strategic use of assessment methods to steer student learning towards a more desirable approach. Some recent attempts to do this are briefly discussed.

Thirdly, I review the use of assessment results (student performance data) to initiate and sustain curriculum change. Two published examples, relevant to the papers contained in this thesis, are used to demonstrate the principle. Firstly, I return to the limited procedural skills competence of new graduates and briefly outline the long overdue changes that have been made, or are being implemented, in response to well demonstrated curriculum deficiencies regarding procedural skills training and assessment in undergraduate medical programmes internationally. Secondly, the method of problem-based learning (PBL) is used as an example to demonstrate the impact of student performance on sustaining curriculum innovation. Despite little evidence in the literature that PBL benefits the academic performance of students, this method of instruction provides numerous other educational benefits that endorse its use worldwide. In addition, new data is providing a better understanding of strategies that can be used to refine and improve PBL.

Finally, the parameters that determine the fitness for purpose, or utility, of assessment practices are discussed. They include reliability, validity, feasibility, acceptability and resource requirements. Educational impact, another fundamental determinant of assessment utility, is not discussed in this section since it is separately addressed in the preceding section. After providing a brief description of each parameter, I highlight the need for a simple, robust way of using these parameters to make rational, i.e. educationally sound, resource-based, decisions when selecting assessment tools. The potential value of such a strategy in the developing world, given the resource constraints (described in Chapter 2) and limited medical education expertise among most developing world clinician-educators is highlighted. The section closes with a brief description of a model of rational drug prescribing developed by the World Health Organization.[11] The potential application of this model to the outlined problem is mentioned.

Having provided a broad outline of the structure and purpose of this literature review, I now deal with each section in detail. The first three sections describe the multiple purposes of assessment: measure professional competence, facilitate student learning and drive and sustain curriculum development and change. The final section of the literature review addresses the issue of selecting assessment tools on the basis of their fitness for purpose (utility).

Assessment to measure professional competence

Competence, in any profession, forms the cornerstone of professional practice.[12] However, before embarking on a discussion of the assessment of competence, a critical function of all professional training programmes, it is necessary to define the term "competence" and delineate the relationship between competence and performance. In terms of professional practice, competence is best defined as "the degree to which an individual can use the knowledge, skills and judgement associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice".[12] This definition highlights four key aspects of professional competence: (1) it is a composite construct comprising profession-specific cognitive, psychomotor and affective skills, often referred to as knowledge, skills and attitudes, (2) it requires the integrated use of these profession-specific skills, (3) it is demarcated by the expected scope of practice of the specific profession or vocation, and (4) it is best determined by observing performance. Professional competence is thus constituted by a relationship between an individual and his or her work, i.e. "it is not something that is directly observed, rather, competence is inferred from performance".[13] In the context of medical education this means that decisions regarding professional competence are best made by observing the proficiency of trainees performing tasks, including cognitive, psychomotor and affective elements, authentic to the practice of medicine. This understanding

of the relationship between competence and performance considerably simplifies the interchangeable and often conflicting use of both these terms in the medical education literature. For the purpose of consistency, I will use the terms as defined here. It is important to note that I have elected to use the term professional competence rather than clinical competence, the term more commonly encountered in the medical education literature. The preceding discussion clearly indicates that competence is not restricted to one specific domain, for example clinical proficiency; it also requires cognitive and affective proficiency appropriate to the profession, i.e. competence appropriate to all the demands of the profession, or professional competence.

Since professional competence is the cornerstone of good clinical practice, it is not surprising that performance-based assessment has demonstrated enormous growth, in medical education terms, over the past 30 years. A number of key issues pertaining to the assessment of professional competence merit brief discussion since they are central to an understanding of the work described in this thesis. They include: (1) a user-friendly taxonomy to facilitate an understanding of the assessment of competence, (2) major factors responsible for driving the development of the vast array of assessment strategies currently in use, (3) an overview of the range of assessment strategies used to assess professional competence, and (4) the need for composite assessment strategies to comprehensively assess the multiple outcome competencies of medical training programmes. Each issue is briefly outlined.

Taxonomy for the assessment of professional competence

Criteria for judging quality are embedded in the scoring processes of all assessment events.[12] These assessment criteria have evolved from the simple concept of "right or wrong" to a considerably more sophisticated understanding of the hierarchical nature of human cognition and behaviour. Such advances in assessment practices have largely been driven by the development of taxonomies describing increasingly complex levels of human cognition. For example, Bloom described six levels of human cognition: (1) simple recall of knowledge or information, (2) comprehension – the ability to explain the meaning of knowledge or information, (3) application – the ability to apply knowledge or information to a specific circumstance or situation (in theory or in practice), (4) analysis – the ability to deconstruct knowledge or information into its constituent components in order to determine their interrelatedness, (5) synthesis – the ability to construct a hypothesis to explain a novel circumstance or situation based on an understanding of knowledge or information previously acquired, and (6) evaluation – the ability to make a judgement decision, i.e. determine the value or worth of something, in a specific circumstance or situation, based on an understanding of knowledge or information previously acquired.[14] Many others have gone on to refine this taxonomy and develop numerous other taxonomies in the cognitive domain. A number of examples are contained in a recent publication by Nitko.[15]

More recently, however, a better understanding of the intimate relationship between cognition and behaviour has lead to the concept of "performances of understanding", i.e. "if you understand something properly you act differently in the contexts understood".[16] This significant advance in our appreciation of the relationship between understanding and behaviour was paralleled by the development of a taxonomy describing growth of competence in terms of increasingly complex task performances.[17] This taxonomy, referred to as the Structure of the Observed Learning Outcomes (SOLO), describes five advancing levels of performance as a function of greater understanding: (1) prestructural – the task has not been approached appropriately (understanding = none), (2) unistructural – one or a few aspects of the task have been appropriately performed (understanding = nominal), (3) multistructural – several aspects of the task have been performed, but each component of the task has been treated as a separate entity (understanding = knowing about), (4) relational – all components of the task have been performed as a coherent whole, with each part contributing to the overall meaning (understanding = appreciating relationships), and (5) extended abstract – the task has been understood and performed as a coherent whole, and has been reconceptualised at a higher level of abstraction enabling generalisation to a new topic or area (understanding = far transfer, involving metacognition). The overlap between the cognitive constructs of this taxonomy, and that described by Bloom, are apparent. Unfortunately this taxonomy, widely known in the general education literature, has not significantly impacted upon medical education assessment practices.

In 1990 George Miller, a medical practitioner, provided a simple description of the hierarchies of human performance as a function of growth in understanding.[18] He depicted the hierarchical nature of professional competence as a pyramid of increasing performance proficiency ultimately culminating in the delivery of good-quality health care (Figure 1).

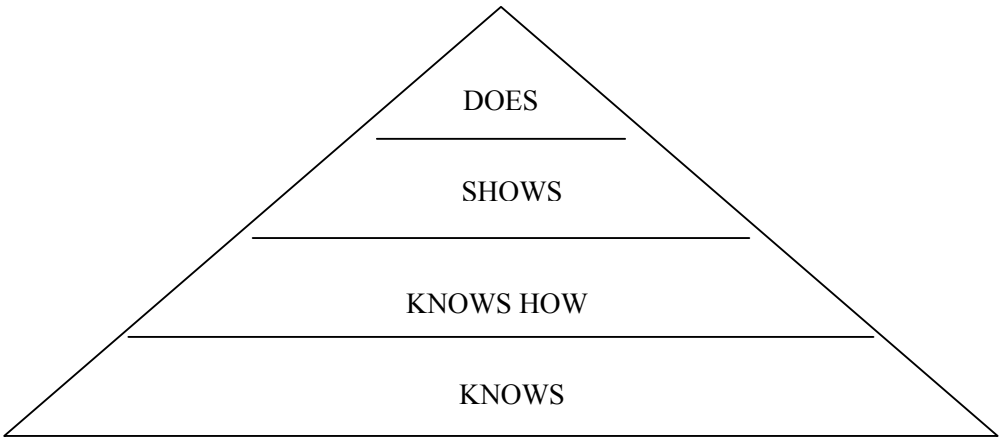


Figure taken from Miller, 1990[18]

Figure 1. *Miller's pyramid of competence*

Figure 1, often referred to as "Miller's pyramid of competence" describes each level of performance using a simple verb which clearly defines the advancing level of proficiency that must be achieved by trainees as they increasingly take on the role and responsibility of providing appropriate health care. Although Miller's pyramid is not traditionally described as a taxonomy of performance proficiency, from which professional competence may be inferred, it certainly functions as such, and almost parallels the performance levels described in the SOLO taxonomy. The simplicity of Miller's pyramid has, however, had an enduring appeal in the medical education literature, and since it continues to form the framework of many discussions regarding the assessment of professional competence,[1,2,19] I use it as the basis for all further discussion regarding the assessment of professional competence in this thesis. A brief outline of the use of Miller's pyramid to stratify the assessment of professional competence is provided in order to orientate the reader to further discussions contained in the literature overview.

While knowledge is embedded in each level of Miller's pyramid, the first two levels specifically focus on assessing the knowledge and theoretical constructs that underpin professional tasks; demonstration of the psychomotor and affective skills required to perform the tasks is not required. Levels three and four of the pyramid, however, require trainees to demonstrate proficiency at performing professional tasks. To achieve this outcome, trainees are required to use, in an integrated manner, the specific cognitive, psychomotor and affective skills appropriate to the task. The concordance between this outcome, and the definition of professional competence provided earlier, is apparent.

The key difference between the upper two levels of Miller's pyramid is the physical location (environment) in which the task is performed – level three tasks take place in a simulated (in vitro) clinical environment (such as a clinical examination setting), while level four tasks take place in the clinical workplace (in vivo). An example serves to illustrate the hierarchy of performance assessment using Miller's pyramid of professional competence. The care of patients with hypertension (a professional task) requires that medical practitioners: (1) are knowledgeable about the causes, consequences, clinical features, investigation and treatment of high blood pressure; (2) understand the techniques for measuring blood pressure, examining the heart, eyes and kidneys for signs of hypertension-related target organ damage and the principles of prescribing treatment for hypertension; (3) are able to correctly measure the blood pressure and examine the eyes, heart and urine of a patient with hypertension as well as appropriately manage hypertension; and (4) appropriately manage hypertensive patients encountered in daily clinical practice, including regular examination of their blood pressure, eyes, heart and urine, as well as adjusting their treatment as required. Levels one and two of the pyramid require candidates to demonstrate cognitive proficiency – knowledge and understanding of the causes, consequences, clinical features and management of hypertension,

while levels three and four require candidates to demonstrate cognitive, psychomotor and affective proficiency relevant to the care of hypertensive patients.

Upon reflection it becomes apparent that Miller's pyramid, formulated some 15 years ago, implicitly advances three concepts critical to our current understanding of the assessment of professional competence: (1) knowledge and skills, including cognitive, psychomotor and affective, should be assessed in an integrated manner rather than attempting to partition the components of professional competence into measurable subcomponents that are separately assessed as knowledge, skills and attitude; (2) the assessment of competence should be hierarchically arranged consistent with the growth of professional competence that occurs over time; (3) assessment processes should be role-based rather than trait-based, i.e. competence represents the increasing ability of trainees to perform the professional functions of a doctor (e.g. manage a patient with hypertension) requiring the integrated use of a diverse range of knowledge and skills rather than the decontextualized performance of specific subcomponents (e.g. measure a patient's blood pressure). This stepwise progression towards more professionally authentic assessment tasks is discussed again later.[1-3,7]

Factors driving the development of assessment strategies

Another major advance in our understanding of the assessment of professional competence has been the realisation that no one assessment tool is able to adequately address all the assessment needs of medical training programmes. Factors responsible for driving the development of multiple assessment methods over the past 30 years can be loosely clustered into four categories: (1) the intended purpose of assessment, (2) the level of competence being assessed, (3) the psychometric adequacy of the assessment process, and (4) the educational and vocational concordance of the assessment process. Each of these factors is briefly outlined in this section, followed by a more detailed in the next section where specific assessment methods are used to illustrate the points highlighted here.

As was mentioned in the preface to this chapter, student assessment serves three fundamental purposes. In this section I explore two of those purposes: (1) the measurement of student achievement for judgement (summative) purposes, i.e. the use of assessment to make decisions regarding selection, placement, promotion to the next year of study, graduation and certification, (2) the measurement of student achievement for student learning (formative) purposes, i.e. the use of assessment strategies to provide trainees with feedback[20] regarding cognitive, psychomotor and affective performance in order to identify their learning needs, guide their learning and motivate them to learn.[8,21,22] Most assessment methods developed over the past three decades can be successfully used for either purpose. By tradition, however, most assessment tools have been developed for summative purposes and continue to be used for this

purpose. The educational (formative) role of assessment, a more recent focus of attention, is separately addressed later.

Secondly, assessment strategies differ in their ability to address the hierarchy of clinical competence depicted in Miller's pyramid. The need for multiple assessment methods to make decisions regarding the various levels of competence of medical trainees, as identified by the tiers of Miller's pyramid, was initially highlighted by Miller himself;[18] many others have done so since.[1,2,4,6-8] Examples of the range of assessment tools that have been developed to measure the different levels of professional competence depicted in Miller's pyramid, are outlined in the next section of this chapter.

The psychometric adequacy of individual assessment tools basically refers to the consistency of the test results obtained using specific assessment methods. The importance of this characteristic of any assessment method is apparent. A more detailed discussion outlining the mathematical expression of assessment result consistency, the reliability coefficient of assessment tools, is provided later. At this point it is sufficient to understand that psychometric adequacy or assessment result consistency can be mathematically expressed as a reliability coefficient, and that this characteristic of all assessment processes has, and continues to be, a major force driving the development of a host of assessment methods over the past 30 or more years. Examples are provided in the next section of this discussion.

Finally, the educational and vocational concordance of assessment processes has becomes a focus of attention more recently. Educational concordance refers to the measure of alignment between learning programme outcomes, the learning methods used in the programme and the assessment methods used to determine that the learning outcomes have been achieved, i.e. the relevance of assessment practices to programme demands. Vocational concordance refers to the measure of alignment between programme learning outcomes and clinical practice demands, i.e. the relevance of the training programme outcomes to professional practice demands. Educational alignment ensures that learning programmes are not driven by the "backwash" effect of assessment, while vocational alignment ensures that graduates are able to meet professional practice demands within the context in which they practice. The critical role of concordance, in both the educational and vocational sense, is discussed in more detail later in this chapter. At this point it suffices to recognise concordance as one of the major factors driving the development of multiple assessment processes, particularly more recently.

Strategies to assess professional competence

Using Miller's pyramid it is possible to cluster the array of assessment tools, currently in use in medical education practice, into four categories (Table 1).[2,19] The list in Table 1 is not exhaustive; only a few examples of assessment tools, representative of each tier of the pyramid,

are listed. A detailed description of each assessment tool is beyond the scope of this thesis. A broad understanding of the essential elements of each category of assessment strategies is all that is required. Performance assessment measures, specifically in vivo (located in the authentic workplace environment) assessment tools, are discussed in more detail because they are increasingly being advocated as the preferred way to assess professional competence – tasks that a qualified medical practitioner should be able to handle successfully.[23] This represents an important shift in the emphasis of programme outcomes from trait-based roles to competency-based roles, a construct compatible with professional workplace demands.[3] Additionally, this theme forms the focus of attention of two papers presented in this thesis. The reader should thus be familiar with a slightly more elaborate understanding of these assessment tools.

Table 1. *Assessment tools used to measure performance*

Levels of competence	Format of assessment	Assessment tools used
Knows	Written assessment	Multiple choice questions, essay questions, short answer questions
Knows how	Written assessment	Multiple choice questions, essay questions, short answer questions, patient management problems
Shows	In vitro performance assessment	Bedside oral examination, objective structured clinical examination, practical assessment of clinical examination skills, directly observed clinical encounter examination
Does	In vivo performance assessment	Clinical work sampling, mini-clinical evaluation exercise , portfolios

Written assessment methods. Assessment strategies focusing on the lower two tiers of the pyramid require candidates to demonstrate an understanding of vocation-specific knowledge, including knowledge relevant to the basic and clinical sciences that underpin medical practice, e.g. Physiology, Anatomy, Biochemistry, Microbiology, Pathology, etc.[2,24] The second tier of the pyramid additionally requires proficiency in the theoretical application of this knowledge to specific clinical contexts. These two tiers are assessed using written tests classified according to the format of the stimulus (indicates what the question wants the candidate to answer) and the format of the response (indicates how the response of the candidate is captured).[25] The format of the stimulus is the most important determinant of the competency being tested, i.e. it determines what is being tested: (1) is the question testing factual knowledge, or (2) does the question require the application of factual knowledge to a specific context as part of a problem-solving process? Understanding, analysis and application of knowledge, encapsulated by the term clinical reasoning, is best tested using contextualised questions i.e. case-based questions.[25,26]

The response format – how the answer is recorded – takes one of two forms: (1) candidates select the correct answer from a number of provided options – multiple choice questions (MCQ), or (2) candidates are required to provide a written response of variable length. MCQs either require selection of an answer from a choice of options provided with each question or selection of an answer from an extended list of options. The latter, a more recent development in question response formats, is referred to as extended matching item (EMI) questions.[27] Open ended responses vary both in length and the cognitive complexity of the task required. Good essay questions require candidates to process information or knowledge rather than just reproduce it, while short answer questions (SAQ) require limited responses where spontaneous generation of the answer is an essential aspect of the stimulus.[25]

While these methods are able test the application of knowledge to clinical contexts, patient management problems (PMP) were specifically developed for this purpose.[28,29] This assessment method was designed to "walk" students through clinical case scenarios by providing information in a stepwise fashion. Students were required to provide written responses to specific questions as the case unfolded. Unfortunately this strategy fell into disuse owing to its psychometric inadequacy.[30] In summary, the bottom two tiers of Miller's pyramid are currently predominantly assessed using MCQs and SAQs because they are more efficient to mark, specifically MCQs, and they demonstrate better reliability for equivalent test times than the other test forms listed.[2,7,19]

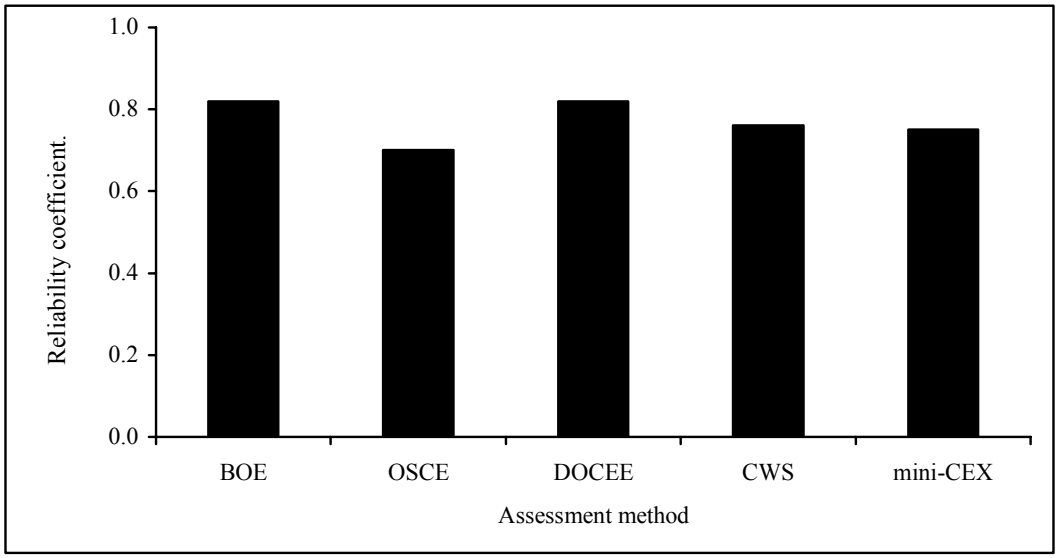
In vitro performance assessment methods. Assessment strategies focusing on the practical demonstration of professional competence, the upper two tiers of the pyramid, "have long served as a rite of passage from training to medical practice".[6] The intrinsic appeal of all performance tests is that they better approximate the context and proficiencies required in authentic clinical practice. Level three tests are usually located within a simulated clinical environment. The best known example is the traditional bedside oral examination (BOE), an assessment method that has been used for decades.[31] Typical BOEs consist of an oral examination based on an unobserved patient encounter (interview and physical examination of a patient) of variable duration – short cases are usually based on a 30-minute patient encounter, while long cases are based on a 60-minute patient encounter. In the mid 1970s the psychometric inadequacy of BOEs, based on an evaluation of the American Board of Internal Medicine's cardiovascular disease oral examination comprising just two BOEs (reliability coefficient of 0.46),[32] resulted in the declining use of this assessment method over the ensuing 30 years.

Attempts to provide more reliable performance-based measures of professional competence led to the development of the objective structured clinical examination (OSCE) in the late 1970s.[33] This assessment method specifically addressed the problem of context-specificity and the resultant poor reliability of assessment events based on limited sampling of

trainee performance.[4,6,7,19] OSCEs were designed to sample a greater number, and wider range, of performance assessment tasks. Typically candidates rotate through 10-20 stations, 10 minutes or less per station, at which examiners directly observe candidates performing one or more clinical tasks.[34] The dramatic increase in the number of events (10-20) scored per candidate in an OSCE, as compared to two or three patient encounters in a traditional BOE, is the single most important factor accounting for the vastly superior reliability of OSCE assessment results.[2,6,7,19] Psychometric adequacy of the OSCE has resulted in widespread implementation of this strategy in many undergraduate and postgraduate medical training programmes worldwide. More recently, the practical assessment of clinical examination skills (PACES), a variant of the OSCE strategy, replaced the BOE (one long case and two short cases) component of the specialist certification examination of the Federation of Royal Colleges of Physicians of the UK.[35]

While the superior reliability of the OSCE strategy continues to make it a very popular performance assessment tool, the time constraints of such short patient encounters – 10 minutes or less – have raised concerns about the "atomisation" and subsequent trivialisation of the complex, integrated (cognitive, psychomotor and affective components) clinical tasks routinely required in daily patient care.[2,3,7,36] An attempt to address this limitation of OSCE assessment, by increasing the duration of each patient encounter and reducing the total number of patient encounters per assessment event, has recently been shown to yield a psychometrically acceptable performance assessment tool referred to as the directly observed clinical encounter examination (DOCEE).[37,38] Indeed, the DOCEE represents a move back towards the traditional BOE previously discarded because of its alleged poor reliability.

Re-evaluation of the traditional BOE has, in recent times, demonstrated that this assessment strategy's previously documented poor reliability was primarily a function of the limited number of items tested (three or fewer clinical cases). If a sufficient number of clinical cases are assessed (3-hour testing time = 9 x 20-minute cases), observed BOEs demonstrate the same psychometric adequacy as that demonstrated by OSCEs of a similar time duration.[39-44] Figure 2 demonstrates that all three in vitro performance assessment tools (OSCE, DOCEE and BOE) perform similarly, reliability coefficients of approximately 0.8, given an examination time of three hours – BOE = 9 x 20-minute cases, DOCEE = 4 x 45-minute cases, OSCE = 27 x 6-minute stations. Indeed, BOEs perform better than OSCEs. Thus, while OSCEs currently dominate in vitro performance-based assessment practice, DOCEEs and observed BOEs – given sufficient sampling – are likely to become popular performance-based assessment strategies in the future.



Data derived from Hamdy et al, 2003;[38] Wass et al, 2001c;[40] Brennan and Norman, 1997;[45] Norcini et al, 1995;[46] Hatala and Norman 1999[47]

BOE = bedside oral examination, OSCE = objective structured clinical examination, DOCEE = directly observed clinical encounter examination, CWS = clinical work sampling, mini-CEX = mini clinical evaluation exercise

Figure 2. *Reported reliability coefficients for different performance tests when 3-hour testing times are used.*

In vivo performance assessment methods. Increasingly the literature is emphasising the need to assess professional competence in professionally authentic environments using tasks that closely approximate or even engage actual clinical practice.[1,3,7] The principal reasons for advocating in vivo performance assessment include: (1) trainees become proficient at performing professionally authentic tasks that constitute part of routine clinical practice (professional authenticity), and (2) the assessment process requires use of the appropriate cognitive, psychomotor and affective skills, essential to routine clinical practice, in an integrated manner (integration). The importance of professional authenticity and integration are discussed in more detail later. At this point it suffices to recognise that the drive promoting the in vivo assessment of professional competence initiated the development of the most recent battery of performance assessment strategies. Both clinical work sampling (CWS), developed in Canada[45,47,48] and the mini clinical evaluation exercise (mini-CEX), developed in the USA,[46,49-51] focus on assessing trainee performance in the workplace using authentic patient encounters. Essentially these are "blinded" patient encounters since the patient is not known to the candidate prior to commencing the interview and/or examination process.[52] Provided sufficient events (approximately 10 patient encounters of 20-25 minutes each) are sampled, both strategies achieve reliability parameters comparable to those demonstrated for the in vitro performance assessment events previously described (Figure 2). This represents a major advance in work-

based assessment strategies and ushers in a new era of performance assessment. To date, these real-time bedside assessment strategies have largely been used for formative assessment purposes. Published reliability parameters,[45-47] however, suggest that these test instruments could be used as summative assessment tools. This has recently been recommended to postgraduate certification bodies in both the USA and the UK.[49-51]

Portfolios of learning. While the workplace-based assessment strategies just described represent significant advances in assessment practices, and clearly show part of the way forward, the use of portfolios, in undergraduate clinical clerkships, represents an additional exciting development. Before discussing the educational merits of portfolios, it is essential to understand the meaning of the term "clinical clerkship" and the word "portfolio", as used in the medical education context.

According to a recent survey of more than 800 medical schools worldwide,[53] medical students are primarily expected to develop clinical competence by undertaking periods of "apprenticeship" attachment to clinical units representing the various disciplines relevant to the practice of medicine, e.g. Internal Medicine, General Surgery, Orthopaedic Surgery, Obstetrics, Gynaecology, Paediatrics, etc. During these clerkship attachments (apprenticeships) trainees are expected to acquire, by observation and supervised practise, the cognitive, psychomotor and affective skills appropriate to the specific clinical discipline represented by the clerkship attachment. Thus, during clinical clerkships trainees actively participate in the delivery of clinical service under the supervision of more senior qualified staff. Clinical clerkships constitute most or all of the training time in the final two or three years of study of most basic undergraduate medical degree programmes. Clerkship attachments vary in duration from 4-8 weeks, and most programmes offer at least two clerkship attachments in the major disciplines, including Internal Medicine, General Surgery, Obstetrics, Gynaecology, and Paediatrics, prior to graduation. The educational value of portfolios, in the context of clinical clerkships, is addressed in more detail later in this chapter.

Derived from the graphic arts, the term portfolio refers to a collection of work for the purpose of demonstrating the development of specific expertise, e.g. fine arts, graphic art design or architecture. Simply put, a portfolio is a "collection of evidence that learning has taken place".[54,54] In the medical education context, therefore, the content of a learning portfolio is largely dictated by the educational intent of the various health professional training programmes in which this learning tool has been implemented over the past decade, e.g. nursing programmes,[56] the training of medical students,[57-59] medical trainees in postgraduate training programmes[60] and medical practitioners participating in continuing medical education programmes.[61] By way of illustration, a student portfolio in an undergraduate medical training programme may include: (1) records of patient encounters; (2) reports reflecting on critical

incidents involving patient care, such as counselling an AIDS patient, informing a mother of the death of her newborn, reflecting on the unexpected death of a patient presenting with acute pulmonary embolism; (3) critical reviews of journal articles relevant to patients in the care of trainees; (4) journal entries reflecting on emotionally or physically demanding experiences such as the first delivery of an infant in an Obstetric clerkship attachment, or a journal entry reflecting on the experience of counselling a rape victim.[57] Regardless of the specific purpose, portfolio tasks usually have two features in common: (1) they consist of a collection of evidence (paper-based, video material, electronic on-line entries, or another mode of data capture) reflecting learning by participation in, or completion of, a number of professionally authentic tasks,[54,55,57,62] and (2) patient encounters form the basis of most learning tasks.[54,57,62] That patient-centred professionally authentic tasks form the basis of portfolio learning activities is entirely appropriate; competent patient care is the most important training outcome of any medical training programme worldwide (IIME,2002).[63]

There are three main reasons why portfolios are thought to be particularly useful learning tools in the training of health professionals: (1) students are given the opportunity to engage in authentic clinical encounters and learn by experience, (2) portfolio tasks provide students with structured learning activities that are suitable for improving the quality of learning in the clinical workplace setting, and (3) students are given the opportunity to demonstrate growth of competence over a period of time. I deal with each aspect in turn. Firstly, professional authenticity, both in terms of the tasks undertaken and the physical location, is probably the major reason for adopting this learning strategy.[54,55,57,62] Trainees are given the opportunity to engage in vocationally relevant learning experiences in the professional workplace. The educational value of experiential learning, first proposed by Kolb,[64] is a widely accepted construct of the theory of human learning. While a detailed discussion of this topic is well beyond the focus of this thesis, a brief explanation of experiential learning is required in order to better appreciate the relevance of portfolio-based learning to medical education, in particular.

In the 1970s, Knowles proposed that adult learners: (1) shift away from dependence towards self-directedness; (2) use accumulated experience as a learning resource; (3) are increasingly orientated towards developmental tasks of social roles, and (4) shift from a subject-centred approach to a problem-centred approach.[65] He also noted that adults enter novel situations with a background of experience, and learning from that experience, thereby highlighting the role of experience in learning. Riegel further advanced this model by suggesting that adult learners are also influenced by their ability to: (1) use logic to identify problems or pose questions, and (2) unite concrete and abstract ideas, thus, facilitating the exploration of complex problems.[66] In the 1980s Kolb expanded our understanding of experiential learning by describing adult learning as a cycle that explicitly incorporates and

builds on the experiences from which learning are derived (Figure 3).[64] Closely allied to both the work of Knowles and Kolb is the principle of "andragogy", described by Mezirow as the facilitation of learning in a manner that enhances the ability of adults to function as self-directed learners.[67]

Integral to all this work is an understanding of the ability of adults to reflect on, and learn from experience. Our understanding of reflection, a critical stage of Kolb's learning cycle, has been further advanced by Schon who highlighted the difference between "reflection in action" and "reflection on action".[68] The former, likened to intuition, is best described as an immediate, appropriate response to an apparently new situation that is sufficiently similar to previous experiences to permit such a rapid response. "Reflection on action" involves revisiting the experience after the event in order to build a memory for future "reflection in action". I return to a more detailed discussion of the process of clinical reflection later in this chapter. At this point it is sufficient to recognise the central role that reflection plays in learning.

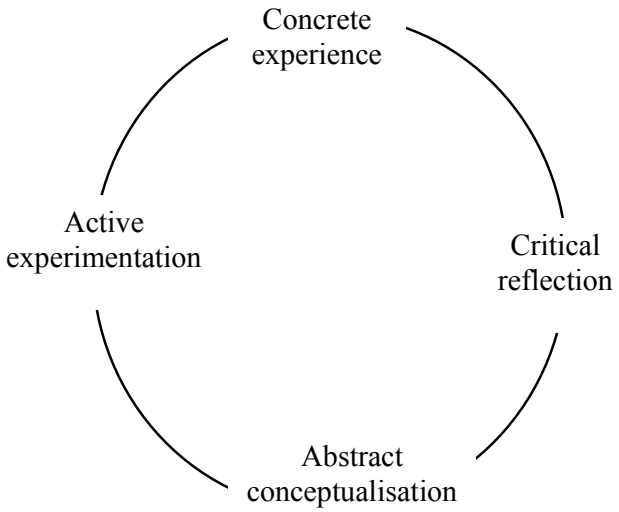


Figure taken from Kolb,1984[64]

Figure 3. *Kolb's learning cycle*

Secondly, portfolios provide students with structured learning activities which can be completed in busy clinical environments under limited supervision. This is a critical issue in clinical clerkships located in the professional workplace, a poorly structured learning environment in which the quality and efficacy of learning is highly variable.[69-74] Two important reasons why the workplace environment is not an ideal learning environment are: (1) learning programmes in clinical practice settings are often poorly structured and learning is frequently opportunistic; (2) senior staff have busy clinical schedules and only a limited amount of time is available to directly observe trainee performance when engaging in daily clinical practice

activities.[72-77] I address the issue of poorly structured programmes in this section and return to the issue of observed, supervised clinical practice later.

The need to structure on-the-job learning in clinical training programmes is well recognised. A study conducted several years ago clearly defined the extent of the problem. Using a recognised classification of on-the-job learning,[78] Friedman Ben-David surveyed faculty opinion and determined the proportion of time students spent learning on-the-job in one of three ways: (1) formal teaching or training (30%) – a formal structured and planned learning activity, e.g. bedside tutorial sessions facilitated by senior clinician-educators, (2) informal (30%) – an intentional approach by the learner who selects a topic and through self-directed, experiential learning gathers the information, e.g. students read up on the electrocardiographic diagnosis of acute myocardial infarction after encountering such a patient on a ward round or (3) incidental (40%) – unintentional or opportunistic workplace experiences in which learners engage and interpret clinical information during routine clinical practice activities, e.g. students are asked to interpret a chest radiograph of a patient seen during a ward round.[1] This important observation, i.e. up to 70% of workplace learning may be unplanned, unstructured and often opportunistic, highlights the urgent need for improving the quality of clerkship learning by structuring learning activities within the clinical environment.

The literature suggests that portfolios may be particularly suited to improving the structure of clinical clerkship programmes if (1) the portfolio tasks are well structured and clear guidelines are provided for both students and supervising clinicians, (2) there is a sufficient range of different experiences in which students can engage, (3) students are given sufficient feedback and coaching to reflect on their performance and identify personal learning needs and (4) the portfolios are formally assessed.[79] These four programme design aspects are thought to promote reflection, defined as "thought processes that help students improve their professional performance".[79] The importance of reflection, a critical component of experiential learning[80,81] and the basis of much portfolio learning, has already been highlighted. Indeed, "reflection is a prerequisite for learning in the context of real practice",[79] as much as it is a desirable feature of working in daily clinical practice.[82,83] Since portfolio work does not automatically result in reflection,[84] identifying practical aspects of portfolio programme design that promote reflection, central to achieving learning, is critical to the successful use of the tool. Programme design elements, similar to those identified by Driessen and colleagues, have also been suggested by Wade and Yarbrough.[85]

The importance of an appropriate portfolio assessment process is based on the universally recognized observation that students value what is assessed and preferentially engage in learning activities that are assessed, i.e. assessment drives learning. This has been

shown to be applicable to portfolio work too.[60.79] Suitable methods for assessing portfolio work are discussed later in this chapter.

Finally, since learning portfolios are compiled over a period of time; in some cases a year of more, trainees are provided with an opportunity to demonstrate growth of competence over a period of time.[57] Review of work collated over a longer period of time provides a more accurate reflection of the trainee's true ability as compared to a single "snapshot" assessment at the end of a clerkship rotation. Thus, from an educational perspective, there are several sound reasons to pursue the use of portfolio learning methods in medical training programmes, particularly clerkship attachments located within the professional workplace. As van der Vleuten and colleagues recently commented: "From an educational perspective, clerkship training is for the most part a black box.".[72] Well-structured portfolio learning programmes, as described here, should go a long way towards eliminating the traditional "black box" approach to clerkship learning.

While enthusiasm for the use of portfolios as a learning tool continues to grow, concerns about the resource-intensive nature of portfolio assessment have been expressed.[54,56,57] Two examples illustrate the problem. At the University of Dundee, Scotland, portfolio examination requires a total of 170 minutes per candidate.[57] Maastricht University, in the Netherlands, recently reported an assessment strategy requiring an examination time of only 11 minutes per candidate.[59] The Dutch authors, however, excluded the time spent on biannual in-course progress meetings between students and their mentors. Since the outcome of these meetings critically informed the final decision made by the assessment committee, they should have been included the interview time in the total assessment time. If conservatively estimated at 30 minutes per session, not stated in the paper, this translates into an estimated total examination time of at least 70 minutes per candidate.

In contrast to the developed world where the human resource implications of such examination strategies are a source of some concern, they are completely prohibitive in world regions like sub-Saharan Africa where up to 50% of medical trainee teaching, supervision and assessment is the responsibility of clinicians not employed as full-time university staff.[53] Furthermore, African clinicians (less than 5 doctors per 10 000 patients) as compared to doctors in Western Europe (up to 30 doctors per 10 000 patients) (WHO,2005),[86] are barely able to cope with clinical service demands, aside from the teaching and assessment needs of the local medical schools that so heavily rely upon them for training-related activities.[53] The extent of the human resource crisis is further aggravated by the massive burden of disease present in sub-Saharan Africa, recently identified as the world region worst affected by poor health and illness (WHO,2005).[86] The extent of the human resource crisis and the burden of disease in sub-Saharan Africa are described in more detail later in this chapter. At this point it suffices to

broadly appreciate the extent of the resource constraints present in this world region. Given these findings, it should, therefore, not be surprising that resource-intensive assessment strategies, such as portfolio assessment, are not widely used in African medical schools.[87] The use of portfolio-based learning, including suitable assessment processes, can only become a feasible option in developing world regions if resource-efficient, reliable assessment methods are developed.

The second major concern regarding portfolio assessment relates to the limited reliability of current assessment methods.[88-91] In a recent study, trained examiners only achieved an overall pass /fail inter-rater reliability kappa score of 0.26 which improved to 0.5 when discussion between examiners was permitted.[91] Improving the psychometric rigour of portfolio assessment is clearly required. Suggestions include the standardisation of portfolio entries, examiner training, structured assessment criteria and a clear idea of the competencies being assessed.[55,56,88]

Thirdly, concerns regarding the suitability of current portfolio assessment methods have been raised. At most institutions examiners read student portfolios in order to provide a final score indicating their satisfaction that the submitted work adequately demonstrates achievement of the specified learning outcomes.[57-59,91,92] The use of student interviews to supplement the portfolio reading process,[57-59] and obtain critical information regarding the development of expertise during the process of compiling a learning portfolio is not universal practice. Where interviews form part of the assessment process, published work does not indicate the contribution these interviews make to the final score awarded or provide detailed descriptions of the interview process. The question has, thus, been asked: "Do portfolios provide educators with real insight into practitioners' clinical ability or simply show that they are good at writing about what they do?"[56] In other words, "Are we assessing what we want to assess, which is the capacity of the professional to integrate knowledge, values, attitudes and skills in the world of clinical practice?"[13] These observations suggest that current portfolio assessment processes may not be the most appropriate way of determining a student's ability to deal with complex professional tasks requiring integration of the "relevant cognitive, psychomotor and affective skills"[7] A number of years ago Friedman Ben-David made a powerful argument for an assessment process in which "the interplay between the contextual evidence and the cognitive processes involved in presenting the evidence becomes a major focus in portfolio assessment".[1] Challis shares the opinion that portfolios should offer students a unique opportunity to participate in a "professional conversation between learner and assessor"[93] A firm rationale does, therefore, exist for assessing portfolios using an interview-based strategy. Aside from the two quoted examples, this has not been further explored in the literature.

In 2002, the University of Cape Town (UCT), South Africa, launched an extensively revised MBChB programme. A more detailed discussion of the programme revision undertaken at UCT, within the current South African socio-political and economic context, is provided in Chapter 2. At this point it is only necessary to appreciate that, despite the difficulties highlighted in the literature, portfolio learning was introduced into the new undergraduate medical degree programme launched at UCT in 2002. In Chapter 4 of this thesis, I describe the implementation of a structured interview technique as a primary portfolio assessment strategy for summative (judgement) purposes. The paper specifically addresses the issues of examination time per candidate, the psychometric adequacy (internal consistency) of the assessment method, and the impact of this assessment strategy on student learning behaviour in the clinical clerkship context.

Educational and vocational alignment of assessment strategies

As mentioned earlier, the drive for professional authenticity and the integrated use of cognitive, psychomotor and affective skills critical to authentic clinical practice, are not performance assessment objectives in their own right.[1-3,7] Rather, they represent important steps aimed at addressing two critical assessment issues: (1) educational alignment – ensuring concordance between training programme learning outcomes and assessment processes (content and method) used to determine achievement of these learning outcomes; (2) vocational alignment – ensuring alignment between medical training programme outcomes and the context-specific vocational demands of professional practice, including the use of assessment strategies that test the appropriate cognitive, psychomotor and affective skills required in routine clinical practice in an integrated, authentic manner,[3,7] In essence both these issues refer to relevance – the relevance of assessment practices to teaching (content and method) practices, and the relevance of educational (teaching and assessment) practices to professional practice. Each is briefly explained so as to provide a framework within which to locate some of the work described in this thesis.

Educational alignment of assessment practices. The success of any learning programme is critically dependent upon an assessment strategy which seeks to determine achievement of the competencies the learning activities were designed to engage, i.e. assessment strategies must focus on determining whether students have achieved the stated outcomes of the learning programme.[16,94] Based on this observation it is self-evident that the learning outcomes of training programmes should clearly articulate the cognitive, psychomotor and affective skills trainees are required to demonstrate upon completion of the programme. Essentially these outcomes should dictate the content and format of all learning and assessment activities. This concept, termed outcomes-based education,[95] was adopted by a number of medical schools in the 1990s, e.g. Dundee University in Scotland.[96,97] In the Netherlands a single outcomes-based

programme, applicable to all medical training institutions, was developed more than 10 years ago.[98,99] More recently it has been recognised that programme outcomes may be better expressed as competencies, i.e. what the doctor is able to do in clinical practice.[3,23] A few medical schools have produced excellent examples of clearly articulated, vocationally relevant, competency-based learning programmes, e.g. Brown University in the USA[100] and Calgary University in Canada.[101] Similar work has also been initiated at the postgraduate level, e.g. the Can MEDS project of the Royal College of Physicians and Surgeons of Canada (CANMeds)[102] and the Outcome project of the Accreditation Council for Graduate Medical Education in the USA (ACGME)[103].

Educational alignment of assessment practices is demonstrated in two papers contained in this thesis. The paper in Chapter 4 highlights the articulation between the interview strategy adopted as the primary portfolio assessment tool (previously described), the learning activities included in the portfolio and national guidelines outlining the training requirements of medical practitioners in South Africa. The latter are described in Chapter 2. The educational impact of alignment in the context of this summative (judgement) assessment process is discussed in some detail in the paper in Chapter 4. The use of a workplace-based formative (learning) assessment strategy, implemented in the 4th year Internal Medicine clerkship of the MBChB programme at the University of Cape Town, is described in Chapter 6. This assessment strategy was specifically designed to provide structured feedback based on directly observed student performance. The congruence between this formative assessment activity and the outcome competencies of the programme, as articulated by the university and the national training guidelines document, as well as the summative clinical examination conducted at the end of the clerkship is highlighted. Once again the educational impact of alignment on student learning behaviour is discussed in some detail. While this seems a logical course to follow, the international emergence of the term "hidden curriculum" more than 30 years ago[104] points to the long history of educational discordance to which trainees have been subjected. The extent to which student learning may simply be driven by the "backwash" effect of discordant assessment processes[16] is again referred to later in this chapter.

Vocational alignment of assessment practices. One of the issues raised earlier was the need for assessment strategies to mimic the complex process of simultaneously engaging the appropriate cognitive, psychomotor and affective skills required to deal with everyday clinical practice. While in vitro performance assessment strategies (BOE, DOCEE, OSCE) do approximate the workplace setting, to a greater or lesser extent, the problem with artificial "assessment" environments is that they passively permit "atomisation" and trivialisation of the complex, integrated clinical tasks routinely required in daily patient care.[2,3,6,7,36] From earlier

discussions it is clear that this issue is probably best addressed by adopting in vivo assessment processes based on authentic patient encounters.

A second critical issue regarding vocational alignment is that assessment processes should aim to determine the competence of skills (cognitive, psychomotor and affective) that are closely aligned with the demands of clinical practice conditions.[96] While clinical skills have been a major focus of attention of undergraduate programmes for a long time, and graduates are generally accepted to be clinically competent in terms of their ability to interview and examine patients, procedural skills (diagnostic and therapeutic) competence has become a more recent focus of attention. The ability of recent medical graduates to safely perform basic therapeutic and diagnostic procedures has been an assumed outcome of medical training since the beginning of the history of clinical practice. More recently, however, it has been recognised that the procedural skills competence of new or recently qualified graduates may not be adequate.[105-112] For example, in a recent survey of medical graduates in Ireland, up to 84% indicated that they had received insufficient undergraduate procedural skills training to function competently during their internship.[109] This has become a matter of concern internationally,[113] and was recently included as a basic minimum requirement of all medical education programmes worldwide (IIME, 2002).[63] This issue is addressed in more detail later, and has been raised here simply to highlight one of the major forces driving the move towards workplace-based assessment processes. If all authentic professional tasks, not just patient interviewing and physical examination skills, were the focus of performance assessment processes, critical issues like procedural skills competence would not have been overlooked to the extent that it has been until recently.

The importance of vocational alignment of undergraduate assessment practices explains the motive for adopting authentic patient encounters as the basis of all portfolio learning activities, as outlined in Chapter 4 of this thesis, as well as the use of authentic patient encounters as the basis of the bedside formative assessment strategy described in Chapter 6. Authentic, workplace-based integration of the appropriate cognitive, psychomotor and affective skills relevant to each clinical encounter is a prerequisite for both assessment activities conducted during the clerkship attachment. This use of clinical records, a documented series of workplace-based patient encounters, to conduct authentic, vocationally relevant assessment processes, such as described in the paper in Chapter 4, is increasingly advocated in the literature.[114,115]

Vocational alignment of undergraduate assessment practices is further examined in Chapter 7 of this thesis. In this paper I describe the outcome of an OSCE evaluating basic procedural skills proficiency in a cohort of South African medical graduates at the start of their first year of clinical service (internship). The results described in the paper highlight the critical

importance of vocational alignment and the need to implement remedial steps to address the "skills gap" identified in this paper. The use of this kind of information to drive curriculum change is referred to later in this chapter.

Composite assessment strategies

As previously discussed, no single assessment event can adequately assess all the competencies expected of medical trainees. This has, over time, given rise to the concept of composite examinations, sometimes referred to as "assessment packages" or "assessment programmes".[7] These terms simply refer to a comprehensive assessment process that collectively addresses all the assessment needs of a training programme using a variety of test instruments. Van der Vleuten and Schuwirth recently suggested that assessment should be viewed as a matter of instructional design rather than the measurement of student achievement.[7] This emphasises the need for assessment strategies to form an integral part of training programme design and development, rather than an odd assortment of tests arbitrarily selected by tradition, habit or ignorance. The critical importance of this fundamental concept is highlighted by briefly considering the context of high stakes assessment processes. These examinations have significant long term implications for candidates, i.e. graduation or specialist certification,[116] and examining bodies have both a social and professional responsibility to ensure that these examinations are credible, fair and defensible.[5,7,117] The literature contains only a few examples of comprehensive (written and clinical components), psychometrically robust composite assessment packages.[118-122] Two examples illustrate the point. Wass and colleagues recently described the composition of a final year undergraduate medical programme examination comprising an MCQ paper, an SAQ paper, an essay paper, a 20-station OSCE and two BOE cases.[122] The second example comes from the Royal Australian College of General Practitioners.[118] Their postgraduate specialist certification examination combines the use of seven different assessment tools including an MCQ paper, an 80-item data interpretation test, two written case commentaries (1 500-2 000 words), two computerised diagnostic problems, five role-play performance assessments, four BOE cases and a 30-minute structured oral examination of a logbook of 100 cases seen. Both examples illustrate the increased scope of assessment that can be achieved using composite assessment systems, a practice widely advocated in the literature.[3,4,6,7] The psychometric adequacy of these assessment packages is discussed later in this chapter.

An example of the use of a composite assessment package in a South African postgraduate specialist training programme is described in Chapter 5 of this thesis. The paper outlines the structure of a composite high stakes postgraduate specialist certification examination conducted by the College of Physicians of South Africa, a member of the Colleges of Medicine of South Africa. In this paper, multivariate generalizability theory, discussed later

in this chapter, is used to determine the overall reliability of the composite examination. In addition, the use of multivariate generalizability theory, to objectively explore options for improving composite examination reliability, is specifically highlighted in this paper.

Assessment to facilitate student learning

For just over two decades leading educationalists, including a number of medical education experts,[1,3,4,8,123,124] have urged the medical education community to recognise the critical role assessment plays in the learning process. Indeed, they argue that all assessment processes should facilitate learning. There are three key mechanisms by which this may be achieved: (1) ensuring educational alignment between programme content, competency outcomes and assessment practices, both in terms of content and method, (2) providing student feedback during or after assessment events, and (3) strategically using assessment events to steer student learning towards a more desirable approach.[6,123-127] The importance of educational alignment has already been emphasised in an earlier section in this chapter, and will not be discussed in any further detail. The other two strategies referred to are addressed in this section.

Feedback to motivate and guide student learning

Feedback, described as "the heart of medical education",[128] is central to the process of learning and constitutes the core purpose of formative assessment.[20,129] The process of providing feedback is thought to promote student learning by informing trainees of their progress, advising them regarding observed learning needs and resources available to enrich their learning, and motivating them to engage in appropriate learning activities.[21,22] Formative assessment strategies are thought to best prompt learning when they are integral to the learning process, performance assessment criteria are clearly articulated for students, feedback is provided immediately after the assessment event, and students engage in multiple assessment opportunities.[125,129] More recently it has been suggested that the efficacy of feedback may be improved by promoting trainee "ownership" of feedback by: (1) encouraging trainees to engage in a process of self-assessment prior to being given feedback, (2) permitting trainees to respond to feedback given and (3) ensuring that feedback translates into a feasible plan of action for the trainee.[130] Failure to formulate an action plan addressing the deficiencies noted in the trainee's performance results in failure to close the "learning loop" and correct the identified performance deficiency. Unfortunately there appears to be a significant gap between medical education advice and "on the ground" practice. Holmboe and colleagues[130] evaluated the type of feedback provided after mini-CEX encounters and found that while 61% of feedback sessions included a response from the trainee to the feedback given, only 34% included any form of self-evaluation by the trainee.

Of greatest concern, however, was the finding that only 11% of mini-CEX encounters translated into a plan of remedial action.

Not only do feedback practices not mirror educational advice, but literature suggests five reasons why so little has been published regarding the use of feedback, given the recognition of its importance in the learning process: (1) current in vivo assessment methods, e.g. the mini-CEX, may be focusing on assessing performance at the expense of providing feedback,[130] (2) the score sheets currently used for in-vivo assessment processes are not designed to provide feedback, and may in fact limit feedback,[130] (3) clinician-educators may not fully appreciate the role of feedback as a fundamental clinical teaching tool,[128] (4) clinician-educators do not regularly observe trainees engaging in routine clinical practice activities,[74,77,130-137] and (5) clinician-educators may not be skilled in the process of providing high quality feedback.[77,128,136,137] Of all the problems listed, the most significant problem regarding feedback is that it can only take pace if trainee performance is directly observed or supervised. The lack of faculty observation of trainee performance, often exceeding 70% of the time is, thus, the most significant problem limiting effective feedback in most medical training programmes.[74,77,131-135] Limited time for teaching and giving feedback, on the part of faculty, is the most frequent explanation offered for this fundamental failure of current assessment practice.[75,76] Daelmans, however, recently succinctly summarised the core issue of the problem: "Supervision is not a structured educational event".[74]

Based on the observation made by Daelmans, it may be reasonable to suggest that the educational value of feedback will only be truly harnessed if direct observation of trainee performance and feedback become a structured activity embedded within clinical training programmes. This is indeed the attempt of the mini-CEX or CWS strategy. To date limited success has been achieved. Only 28% of 114 undergraduate medical programmes in the USA recently surveyed, have successfully employed the mini-CEX as an undergraduate formative assessment strategy.[138] The frequency of assessment events, however, remains largely opportunistic because trainees are usually required to self-initiate feedback on their performance; hence the failure of this strategy in some settings.[74] The statement made by Daelmans is really hinting at the need to make trainee observation and structured feedback a formal component of learning programmes rather than an assessment event. In this way the opportunistic element, i.e. self-initiated student requests for observation and feedback, would be eliminated. Furthermore, it has been suggested that structured feedback forms improve the quality and frequency of feedback,[139,140] as does the training of clinician-educators in the observation and rating of trainees performance.[136]

In Chapter 6 of this thesis, I describe the implementation of a bedside-based formative assessment strategy to provide 4th year medical students with structured feedback during a 14-

week medical clerkship at the University of Cape Town. In contrast to other formative assessment strategies, this process was embedded in the weekly bedside teaching sessions and thus formed an integral part of the training programme. This obviated the need for student-initiated requests for feedback. In addition, I designed a structured feedback form based on examples contained in the literature.[47,51] The frequency of feedback obtained during the clerkship attachment as well as student and faculty perceptions of the educational value of feedback are discussed in this paper.

Impact of assessment practices on student learning behaviour

The impact of summative assessment practices on student learning behaviour is well documented.[6,124,125,127,141] Crooks provides a comprehensive review of all the literature relevant to classroom-based assessment practices and concludes that test format, content and frequency significantly impact upon student learning behaviour.[125] These are similar to the observations made earlier by Frederiksen.[124] These early papers urged educators to recognise the educational value of assessment events and focus attention on making learning, rather than measurement, the primary outcome of assessment activities. Frederiksen considerably broadened the concept of assessment by stating that a "test may be thought of as any standardised procedure for eliciting the kind of behaviour we want to observe and measure".[124] This recognition of the potential to strategically use assessment processes to manipulate student behaviour and reinforce desirable learning behaviour has again been recently emphasised.[127,142] They also emphasise the critical importance of concordance between programme learning outcomes and the format and content of assessment processes used to determine achievement of these outcomes. This point was highlighted earlier in this chapter.

The literature contains two well known examples of the impact of summative assessment practices on the behaviour of medical trainees. In the paper by Newble and Jaeger[123] they describe the impact of changing a final year assessment process from a performance-based bedside oral examination (BOE) and a written examination (MCQs) to a written examination only. Not surprisingly, the students spent less time in the wards and almost all their time studying in the library. This unanticipated negative impact of the change in assessment practice was rapidly remedied when performance-based BOEs were re-instated. In the second paper, Stillman and colleagues report an increase in the number of observed student clinical encounters undertaken by faculty after a performance-based clinical examination was instituted in the final year of the medical degree programme.[126] This paper suggests that the increase in observed clinical encounters was largely driven by a conscious effort on the part of faculty to improve the clinical skills of students during their clinical clerkship attachments. Thus, it may be plausible to suggest that changed assessment practices may also impact upon staff teaching behaviour, to a greater of lesser extent.

More recently, attempts have been made to strategically direct student learning by selecting assessment methods that reinforce desirable learning behaviour.[127,142] A good example in the medical education literature is lacking, but Driessen and van der Vleuten have provided a useful example from a Law faculty in the Netherlands.[127] They introduced a portfolio of assignments as an educational tool in a legal skills training programme comprising tutorials which were poorly attended and for which students did not adequately prepare. The portfolio assignments (e.g. writing a legal contract, drafting a legislative document), reviewed by peers and the tutors, were used as the basis for subsequent skills training sessions. Assessment feedback given by peers and tutors was kept by students in a file with the assignments. This portfolio learning and assessment process resulted in a twofold increase in the time spent preparing for the skills training sessions, 2.9 as compared to 7.4 hours per week, and both faculty and students were in favour of the strategy. Students, in particular, recognised the learning value of the portfolio assignments.

In Chapter 4 of this thesis I describe the use of a structured interview to determine the learning achieved by students compiling a portfolio containing a prescribed number of authentic clinical encounters, including the provision of supervised care for patients admitted during their clinical clerkship attachment. The impact of this summative assessment strategy on learning behaviour is discussed in the paper.

More recently, the use of workplace-based formative assessment strategies such as the mini-CEX, or variants thereof, has led to an interest in the potential impact of feedback, provided during formative assessment events, on trainee learning behaviour in the workplace environment. A recent paper from Argentina evaluated the learning strategies adopted by postgraduate cardiology specialist trainees in response to the use of the mini-CEX strategy.[143] The authors suggest that formative assessment, using the mini-CEX strategy promoted a desirable approach to learning, i.e. an attempt to (1) understand the meaning of the subject matter rather than learn by rote, (2) adapt study strategies according to personal interest, knowledge and needs and (3) construct a relationship between personal experience and topics studied. These findings are concordant with a "deep" approach to learning, first described by Marton and Säljö some 30 years ago.[144,145] Unfortunately the study by de Lima and colleagues included only 16 candidates exposed to one CEX event each. Furthermore, the candidates were postgraduate students enrolled in a highly competitive academic training setting. Thus, the generalizability of these results, although encouraging, may be limited.

A lengthy discussion of learning approaches is clearly beyond the scope of this thesis, but a brief explanation is provided so as to enable the reader to appreciate the significance of the study finding. Essentially Marton and Säljö identified two key approaches to learning: a "surface" approach characterised by rote learning and (2) a "deep" approach whereby learners

attempted to understand underlying principles, concepts and ideas and interpret them in a personally meaningful way. Subsequently Entwistle and colleagues added a third approach based on their observation of the impact of assessment on learning strategies.[146,147] They called this approach "strategic" learning – "the conscientious, well-organised learner whose study methods are closely linked to achievement motivation and the desire to excel in an upcoming assessment event".

In contrast, two Dutch studies failed to show dramatic changes in undergraduate student learning behaviour in response to the implementation of workplace-based formative assessment strategies with structured feedback in a surgical and medical clerkship, respectively.[134,135] In the surgical clerkship it is encouraging to note that students spent more time performing clinical procedures and less time engaged in "waste-of-time" activities, e.g. collecting blood samples and finding X-rays. However, of concern is the observation that they spent less time on ward rounds or engaged in authentic patient encounters. This seems contrary to the intended outcome of clerkship learning, especially since it is increasingly recognised that most clinical competencies required in professional medical practice can really only be attained by spending many hours working in authentic clinical practice settings.[23] Many of the tasks doctors have to learn to perform cannot be taught on manikins or with simulators, and simulated patients provide a restricted range of authentic clinical training.[148] The authors of both papers express concern that part of the failure of their educational initiatives may have been due to limited participation on the part of the supervising clinical staff. This is supported by the finding that the majority of students in both studies indicated that supervision, observation of competencies performed and feedback regarding performance were largely obtained from trainees rather than senior clinicians.

In Chapter 6 of this thesis I present a paper describing the implementation of a bedside formative assessment strategy. This paper, already described, reports on the implementation of a bedside formative assessment process based on the use of "blinded" patient encounters.[52] Essentially, this technique requires that students conduct a directly observed interview and /or examination of a real patient, without access to the patient's clinical record, as part of a bedside teaching session. The case is then presented by the student and the discussion forms the basis of the bedside tutorial session. As indicated earlier, students received structured feedback after theses tutorial sessions. The impact of this assessment strategy on learning behaviour is discussed in the paper.

Assessment to initiate and sustain curriculum change

The impact of student performance data on curriculum development and change is an not increasingly recognised function of assessment i.e. the programme evaluation role of assessment.[149-151] In this section I focus on only two specific issues relevant to the work presented in this thesis: (1) the use of student assessment results to initiate curriculum change, and (2) the use of student assessment results to endorse and sustain curriculum change. Both these topics are worthy of lengthy discussions in their own right, but I wish to focus only on the role assessment results may play in bringing about curriculum change and then sustaining change achieved. In order to remain within the context of this thesis, I focus on only two examples in the literature. The first example illustrates the use of assessment data to identify curriculum changes needed, and the second example addresses the issue of sustaining curriculum change using assessment data.

Initiate curriculum change

While trainee performance has not greatly influenced curriculum design and content in the past, this trend is changing. The one domain, in particular, in which trainee performance has demonstrated an incremental influence on curriculum design and content is in the area of procedural skills proficiency. This is largely related to two recent global trends: increasing societal concerns about the quality of practising doctors[152] and an international recognition that the ability to competently perform a wide range of diagnostic and therapeutic procedures should be a core learning outcome of modern undergraduate medical training programmes.[63,97] While basic diagnostic and therapeutic procedures performed in routine medical practice have not changed dramatically over recent years, competence performance of these procedures, particularly at the level of junior doctors, has assumed greater importance than previously.[113] Traditionally, it has always been assumed that the skills needed to practice medicine, particularly diagnostic and therapeutic procedural skills – for example, the ability to aspirate a sample of fluid from a collection of fluid on the lungs (diagnostic procedure) or the ability to insert a catheter into the bladder to drain urine (therapeutic procedure) – are acquired during the clinical clerkship years of medical training programmes. As already alluded to earlier, the clinical workplace environment is not an ideal training environment,[69,70,72,74] and unstructured clerkships should no longer be relied upon to ensure adequate skills training.[72] A number of publications spanning two decades suggest that the procedural skills proficiency of new medical graduates has failed, and continues to fail, to meet the expectations of senior clinicians for whom new graduates work in their first or early years of clinical service.[105,107-109,111,153]

The reason for the limited alignment between undergraduate programme outcome competencies, and the demands of professional practice, regarding this particular learning

domain, may relate to the fact that national training guidelines are often produced by statutory government or parastatal organizations responsible for registering and /or accrediting medical training programmes, e.g. the General Medical Council (GMC) of the UK, the Health Professions Council of South Africa (HPCSA), the Association of American Medical Colleges (AAMC). These regulatory bodies generally provide only broad educational outcomes, including skills proficiency, without specifically articulating a list of core skills needed at the time of commencing clinical practice.[154-156] Some countries, e.g. the Netherlands[98,99] have addressed this serious shortcoming by developing comprehensive lists of learning outcomes, including lists of basic procedural skills that graduates are expected to be proficient at performing when commencing clinical practice. There is thus an international need to address this matter urgently. The lack of adequate procedural skills proficiency at the commencement of clinical service can no longer be ignored.

While the need for structured procedural skills training programmes is clear,[113] the vocational relevance of the procedures taught is also of critical importance. This is particularly true in developing countries where junior doctors often commence clinical practice in relatively poorly supervised settings immediately after graduation. It is, therefore, highly likely that the skills proficiency of graduates from different world regions may differ greatly, depending upon the vocational demands made of them in their initial year(s) of service delivery. While an international set of minimum medical training outcomes (IIME, 2002)[63] have recently been published, the literature emphasises the need to tailor training needs to the specific contexts in which training programmes are located.[157,158]

Based on the preceding discussion, it is clear that specific procedural skills need to be identified on a national or regional basis and undergraduate training programmes need to be revised so as to ensure that the identified procedural skills are adequately taught and assessed prior to graduation. This may require significant curriculum revision in many circumstances.

In Chapter 7 of this thesis I describe evaluation of the basic procedural skills proficiency of newly-qualified South African medical graduates using a seven-station OSCE. The findings of the study are discussed and the need for formal skills training and assessment prior to graduation is highlighted.

Endorse and sustain curriculum change

Achieving and sustaining curriculum change are challenging tasks.[159-161] The use of student performance data to endorse and sustain curriculum change, the programme evaluation role of assessment, has become a topic of growing interest over the past 30 years. Perhaps the most well known programme innovation that has been the subject of intense debate for the past 15 years is problem-based learning (PBL). Before embarking on a discussion of the evaluation

of PBL programmes, a working definition of PBL and a brief description of the process are required so as to highlight the major departure of this educational strategy from traditional large-class lecture-based teaching programmes. PBL may be defined in many ways, but I prefer to think of it as an instructional approach which attempts to apply our growing understanding of human cognition and learning to educational practice.[162] Four modern insights regarding human learning form the foundation for PBL.[163] They are: (1) learning is a constructive process – students must actively construct knowledge networks by engaging in a process of creating meaning and building interpretations of the world based on personal experience and interaction; (2) learning should be a self-directed process – students need to play an active role in planning, monitoring and evaluating their own learning; (3) learning should be a collaborative process – students need to engage in a collective learning process in which learning tasks are shared and mutual interaction results in a shared understanding of the problem; (4) learning should be contextualized – all learning should take place in a context, i.e. all learning should be situated because the situation in which the knowledge is acquired determines the use of the knowledge. Each of these points could be considerably elaborated upon, but this basic description suffices for the purposes of this thesis.

Although PBL is conducted in a variety of ways in different medical schools across the world, the process has three essential characteristics:[163] (1) clinical problems serve as the stimulus for learning – students are given a clinical problem for which they are required to identify and find, by personal study, new knowledge needed to understand the biomedical and psychosocial concepts illustrated by the problem, i.e. they actively construct new knowledge which is linked to their prior knowledge; (2) learning takes place in small groups – students, in groups of 10 or less, work together on identifying the new knowledge needed to understand the clinical problem being discussed, and then learn by interacting with each other when they discuss the knowledge acquired during the period of personal study that takes place between the small group sessions; (3) tutors act as facilitators of the learning process – the tutor's task is to keep the group learning process going by probing students' level of understanding of the information being discussed, ensuring participation by all group members, modulating the direction of the group discussion in order to stay focused on the relevant learning issues and, finally, the tutor also monitors the educational progress of each student in the group so as to initiate remediation if and when necessary. The striking differences between this educational strategy and a traditional large-class lecture-based programme are apparent. In some medical schools, PBL is supported by supplementary large-class lectures, laboratory or skills centre practical sessions and small group tutorials.[164] These activities serve as additional learning resources rather than the primary mode of instruction.

This educational learning strategy, formally described by Schmidt at Maastricht University,[165] was implemented in a handful of medical schools in the late 1960s and early 1970s, including McMaster University in Canada, Newcastle University in Australia, the University of New Mexico and Michigan State University, both in the USA. Since then PBL has become a worldwide phenomenon. A recent survey of more than 800 medical schools around the world found that some element of the PBL approach was being used in more than half of schools surveyed.[53] Mamede and colleagues recently made the interesting observation that this learning method has been widely implemented despite limited empirical evidence emerging from numerous studies comparing traditional curricula with PBL programmes.[166] One of the earliest papers reviewing the evidence in support of PBL was published in 1987.[167] Schmidt and colleagues found data to support the idea that PBL encouraged an inquisitive learning style and appeared to influence the career choice of graduates; a career in primary care practice was favoured. Evidence that PBL students performed better in conventional knowledge tests was, however, not forthcoming. In 1992, Norman and Schmidt[168] reviewed the evidence substantiating the theoretical advantages of PBL, and concluded that there was a reasonable theoretical basis for the idea that PBL promotes better transfer of concepts to novel situations, evidence that PBL group discussions stimulate the activation and elaboration of prior knowledge which facilitates retention of new knowledge, and evidence that PBL enhances self-directed learning. Since these two early papers, a multitude of studies and a number of reviews have been published. Most of the large review papers have not consistently shown significant differences in favour of PBL when students in PBL and traditional programmes are compared on conventional measurements of knowledge.[169-174] A number of other outcome measures have, however, been shown to be better in PBL programmes. These include: student and staff satisfaction,[170,171] problem-solving ability, clinical reasoning and diagnostic accuracy,[170,171,175-177] the use of a "deep" approach to learning,[170,171,178] the integration of biomedical and clinical knowledge,[179] self-directed learning skills[169-171,177] and communication skills.[177,180,181] These comparative studies have sparked an ongoing debate in the literature about the relative merits of using different measures of student performance as indicators of curriculum innovation success.[182-185]

A recent paper by Mamede, Schmidt and Norman[166] has added a further dimension to the ongoing debate regarding the evaluation of PBL. They review a number of recent papers focusing on specific aspects of the PBL process rather than the product, e.g. the impact of the group process on student learning,[186] the impact of clinical experiences on learning in PBL sessions[187] and the impact of learning resource availability on student performance in PBL programmes.[188] They conclude that future research should pay more attention to factors influencing the educational impact of PBL, i.e. the process, rather than just evaluating the

graduate, i.e. the product. This conclusion endorses the suggestion recently made by Dolmans and colleagues,[163] and furthermore supports the suggestion that future research needs to focus on identifying ways of closing the gap between the theory of PBL and the actual findings in practice, i.e. a better understanding of how PBL does or does not work. This builds on earlier suggestions by Norman and colleagues[184,185] that PBL research should not attempt to identify the educational impact of PBL at curriculum level since such effects are unlikely to be detected owing to the presence of unidentified confounding variables and the unreasonably large effect size that would need to be present in order to generate a measurable impact.[189] Rather, they make a plea for theory-driven research and the use of research tools, such as structural equations modelling,[184,190] that are able to dissect out the various contributing factors in order to improve our overall understanding of how PBL actually works.[184,185] A detailed explanation of the principle of structural equations modelling is not relevant to the work done in this thesis, but the elegance of this statistical method is well demonstrated in a recent paper by de Bruin and colleagues.[191] The interested reader is referred to this paper for further details.

While future research regarding the PBL process is clearly a matter of great importance, the considerable financial and human resource costs involved in the use of this educational method,[169,171] make measurable graduate outcome parameters an ongoing source of concern, particularly in resource-limited settings. Although traditional measures of student academic performance have not yielded useful data, an alternative avenue of research that deserves further exploration is the use of student retention rates as an indicator of programme efficiency. Clearly retention rates cannot serve as a direct marker of academic performance, since students may drop out of programmes for many reasons other than academic performance, but they do provide an early and ongoing indication of the capacity of a programme to produce medical graduates. In resource-limited settings, where traditional programmes are considered significantly more cost-effective, the need to demonstrate curriculum innovation efficiency early on in the change process is of considerable importance.[192] In the review by Vernon and Blake[170] and a paper by Mennin and colleagues[193] brief mention is made of student attrition rates in PBL programmes as compared to conventional programmes. Both papers, published in 1993, reported similar dropout rates. The issue does not seem to have attracted any further research interest.

Recently, however, Iputo and Kwizera from the Walter Sisulu University (WSU) in South Africa reported significantly lower attrition rates in their PBL programme as compared to a traditional programme.[194] The origin of this historically Black university, and the significantly academically disadvantaged profile of their student enrolments, as compared to historically White universities in South Africa, is discussed in some detail in Chapter 2. At this point the reader is only required to appreciate the fact that this medical school predominantly admits

students from severely academically disadvantaged backgrounds.[195] The findings of this study are, thus, of importance in developing countries where academically disadvantaged students may make up a considerable proportion of medical school entrants, e.g. South Africa.[196] This issue is also addressed in more detail in Chapter 2. Furthermore, the novelty of the recent finding at the WSU is even better appreciated if it is recognised that the University of New Mexico School of Medicine elected to keep academically-at-risk students in their conventional programme when they first implemented PBL.[193] They made this decision based on their concern that academically weaker students would be at greater risk in a less structured, more self-directed learning environment. Thus, the findings documented by Iputo and Kwizera deserve further evaluation. Is it possible that student retention or attrition rates may be a useful early indicator of successful curriculum innovation, in particular the implementation of PBL, in resource-constrained environments where the enrolment and throughput of academically-at-risk students is a priority? Given that resource constraints dictate medical education practices to a considerable extent, a point elaborated upon in Chapter 2, it would be very useful to find an early indicator of curriculum change success so as to sustain change in the face of ongoing resource limitations. The data discussed, suggest that student retention or attrition rates may be a useful measure, specifically in circumstances where student enrolment (admission) criteria are not uniform. The potential impact of this factor on the evaluation of programme innovations is mentioned in the literature, but not explored to any great extent.[197]

The impact of the political legacy of Apartheid in South Africa, discussed in Chapter 2, makes the enrolment of educationally disadvantaged students a priority. Determining early success of programme innovations benefiting these academically-at-risk students is thus of critacal importance. The situation may be similar in other developing countries, e.g. India, and even developed countries attempting to improve the ethnic representation of minority groups in medical schools enrolments, e.g. the Aboriginal people in Australia or Native Indians in Canada and the USA.

In Chapter 8 of this thesis I evaluate the retention rates and academic performance of academically-at-risk students admitted to the PBL programme recently initiated at the University of Cape Town (UCT) in South Africa. The retention rates and performance of these at-risk students is compared to that of similarly at-risk students previously admitted to an extended traditional programme operational at UCT between 1991 and 2000. The details of this programme, including its political origin, are explained later in Chapter 2. The findings of this study are presented in the paper in Chapter 8 and the implications thereof are highlighted.

Utility of assessment practices

The utility of an assessment procedure, as defined in this thesis, refers to the overall usefulness or fitness for purpose of a specific test instrument.[4,9] Factors determining the utility, or fitness for purpose, of assessment practices are well defined in the literature.[4,9] Methods by which these utility parameters may be used to make rational decisions regarding the selection of appropriate assessment tools are, however, limited. These two issues are briefly outlined and the relevance thereof, to the work presented in the thesis, is highlighted.

Parameters determining the utility of assessment practices

The selection of a particular assessment method inevitably involves compromises and trade-offs.[4,6,7] The critical issue is whether these compromises and trade-offs are made at random, i.e. out of ignorance, or whether they represent a conscious decision on the part of the clinician-educator choosing the assessment tool(s). In order to facilitate the decision-making process, van der Vleuten has provided a very useful conceptual framework outlining the key parameters that determine test utility: (1) reliability, (2) validity, (3) educational impact, (4) acceptability, and (5) resources required.[4] Crossley recently rephrased these five determinants of test utility into two categories: (1) parameters indicating the rigour of a test, i.e. reliability and validity; and (2) parameters determining the practicality of a test, i.e. feasibility, cost (resources required) and acceptability.[6] The classification suggested by Crossley and colleagues is useful because it permits separate evaluation of the educational impact of assessment, a critical function of assessment discussed earlier in this chapter.[4,7,124,125] For the purpose of this thesis, I focus on the assessment utility parameters defined by Crossley and colleagues.[6]

Parameters indicating test rigour. Validity, the appropriateness and meaningfulness of the inferences made from test results, and reliability, the consistency of test results, are two basic concepts that have acted as the principal determinants of test utility since their description more than 50 years ago.[8,198] Neither of these assessment utility parameters is an inherent property of the test itself. Rather, they refer to the results obtained from a test process, and the manner in which these results are interpreted or used.[8,199] While each is important, they are inextricably linked – "reliability provides the consistency of results that makes valid inferences possible."[8] Although these two assessment utility parameters are not the major focus of attention of the work described in this thesis, a few basic concepts relevant to each are referred to in four of the six papers included in the dissertation. For this reason it is imperative to provide a brief overview of key reliability and validity issues relevant to this thesis.

Reliability is most simply defined as the consistency or repeatability of test results, i.e. the amount of error, random and systematic, inherent in any measurement.[8,198] The concept is mathematically expressed as the ratio of true variance, i.e. variance between candidates, to total

variance, which includes variance due to measurement "error". This calculated ratio, known as the reliability coefficient, thus expresses the relationship between true variability and measurement error over time (test-retest reliability), test items (inter-item reliability) or examiners (inter-examiner reliability). The reliability coefficient of test scores can be determined in a number of ways, including calculation of Cronbach's alpha coefficient, a measure of internal test consistency.[200,201] A detailed discussion of the various methods used to calculate test score reliability is beyond the scope of this thesis. Detailed explanations of the various methods used are contained in standard reference texts.[8,198]

The minimum acceptable reliability coefficient for any given test is critically dependent upon the intended purpose of the test, i.e. what decisions are going to be made on the basis of the test results?[3] There is, therefore, no standard minimum reliability coefficient for any given assessment method. Rather, the importance of the relationship between the accuracy of the scoring and the importance of the consequences of the decision to be made should be appreciated. For example, test results having significant consequences, e.g. graduation from a degree programme or postgraduate specialist certification, should demonstrate good reliability, i.e. a high reliability coefficient is desirable. The Royal College of Physicians and Surgeons of Canada have suggested a minimum reliability coefficient of 0.75 to 0.85 for high stakes assessment processes (Royal College of Physicians and Surgeons of Canada, 2000).[202] This convention, widely followed in the literature, is based largely on consensus opinion rather than specific criteria or evidence.

The reliability of a test score can also be expressed as the standard error of measurement (SEM), which is derived from the reliability coefficient using the following formula:

$$\text{Standard error of measurement} = s\sqrt{1-r_n}$$

where s = standard deviation and r_n = the reliability coefficient.[8] The advantage of expressing test score reliability as the SEM is that it provides an indication of the amount of error to allow for when interpreting individual test score results. For example, if a candidate achieves a test score of 65% for a test having a standard deviation of 4.5 and a reliability coefficient of 0.6, then the SEM would be 2.8. The 95% confidence interval (CI) of the candidate's test score, calculated by multiplying the SEM by 1.96 and adding /subtracting it from the measured test score, would be 59.5% to 70.5%. This provides a more meaningful interpretation of test score reliability for individual candidates, specifically academically weak (borderline) candidates where a score of 45% for the same test would have yielded a 95% CI of 39.5% to 50.5%. Given this wide confidence interval it is clear that borderline candidates would be at a considerable risk of failing this hypothetical test because of the poor consistency of the test results. The example clearly demonstrates the limitations of test score results for borderline candidates, and

the value of having a numerical expression of the reliability of such scores in order to make informed decisions about the appropriate use of test score results in such circumstances.

When classical test theory first permitted calculation of test reliability,[198] the drive to "objectification" of all assessment tools became an obsession that did not necessarily improve the quality of tests.[203] Reliability, or the consistency of test results, while clearly a desirable feature of good assessment practice, should not be pursued at the expense of other important features of good assessment practice.[51] Indeed, it is critical that the educational context, or basic purpose, of an assessment event always remains the primary determinant of the test content and format selected.[203]

Three key factors influence the reliability of test scores: (1) the number of test items included, (2) trained examiners using clearly defined scoring methods that limit inter-examiner variability, and (3) carefully selected test items that demonstrate limited inter-item variation.[8] More than two decades of research has shown that the number of test items sampled is the most important determinant of test reliability.[6,8,19,36,43,203] The main reason for this consistent observation is that candidate variability across test items is far greater than variability between test items (inter-item reliability) or examiners (inter-examiner reliability).[6,36,203]

When assessing clinical competence (performance assessment), the marked variation in candidate performance across test items, for example bedside oral examination cases or OSCE stations, is best explained in terms of context-specificity and the dependent relationship between performance and domain-specific knowledge.[5,6,36,203] The phenomenon of context-specificity, also referred to as case-specificity, was first described by Elstein and colleagues in the late 1970s.[202] This concept is best understood by referring to a clinical example previously used. In a performance test situation, the ability of a candidate to perform a competent physical examination of a patient with hypertension is largely dependent upon knowledge of the pathological processes involved in the development of hypertension, the nature of the major target organ damage caused by hypertension and the relevant accompanying clinical signs. Not only are performance tests dependent upon such context-specific knowledge, but more importantly, proficiency in one performance test item does not predict proficiency in another performance test item. For example, candidates proficient at examining a patient with hypertension may not demonstrate the same proficiency when examining a patient with pneumonia. The knowledge required to competently assess a patient with pneumonia (a disorder of the respiratory system) is unrelated to the knowledge required to competently assess a patient suffering from hypertension (a disorder of the cardiovascular system). Indeed, even within the same discipline, performance of one task does not predict performance of another task – for example, the competent assessment of a patient with hypertension does not predict the ability of a candidate to recognise the electrocardiographic signs of acute myocardial infarction.

Highly variable performance across test items, therefore, requires a sufficient number of test items to minimise the error incurred by inadequate sampling due to case-specificity.[5,6,8,203] The importance of comparing the reliability of different test formats using a standard duration of testing time, a function of the number of test items included, is, therefore, critically important. This issue has already been highlighted. The reader is referred back to Figure 2 where the reliability coefficients of a number of performance assessment methods are shown using equivalent examination times.

More recently the issue of case-specificity has been re-examined. Current data suggest that a general ability to deal with problems is also relevant to the problem-solving ability of students.[205] This observation does not negate the earlier findings, but serves to add to our understanding of clinical reasoning. Case-specificity is, therefore, more recently understood to be one of two important contributing variables. Further discussion of this concept is not required for the purposes of this thesis. The most recent findings have only been mentioned so as to provide an accurate reflection of the current understanding of factors accounting for case-specificity in the context of performance assessment.

While the reliability of individual test scores can be calculated using the classical statistical methods referred to earlier, e.g. test-retest method or internal consistency methods, the reliability of composite assessment packages containing a number of different tests cannot be calculated in the same manner.[8,198] This stumbling block in the evaluation of composite test reliability has been elegantly addressed by the development of multivariate generalizability theory,[206,207] first advanced by Brennan in the early 1970s.[208] This theory, based on the analysis of variance – a basic statistical concept,[209] enables estimation of the reliability of composite assessment processes containing a number of different tests. In addition to determining the reliability coefficient of a composite examination, prediction studies can also be done. Such studies, called decision studies, use existing data to predict the optimal composition of assessment packages containing a number of different tests.[198] A detailed description of the technical issues relevant to the use of multivariate generalizability theory is not required for the purpose of this thesis. A basic appreciation of the concept, a major statistical advance that has the capacity to greatly improve the quality of multi-component assessment packages, is all that is required.[206]

The medical education literature contains two good examples of the use of multivariate generalizability theory to determine the reliability and optimal composition of high stakes multi-component examinations. Both examples have already been referred to earlier in this chapter. The composite undergraduate qualifying examination of a UK medical programme was recently evaluated by Wass and colleagues and found to have a reliability coefficient of 0.76.[122] Based on current practice this is reasonable, but further improvement was desirable. To this end, a

series of D-studies were conducted which showed that the examination reliability could be further improved by altering the weighting of the examination subcomponents, i.e. reducing the weighting of the clinical component of the examination. In another study by Hays and colleagues,[118] the technical details of how to determine a composite examination reliability coefficient are described. Their multi-part examination had an overall reliability of 0.8 and was considered adequate. A number of D-studies identified changes that could have been implemented to further improve the quality of the examination. Given the elegance and educational insight this statistical technique provides, in terms of improving the quality of assessment packages using objective data, this is indeed surprising to find such a limited number of publications.[206] Limited access to user-friendly software[206] and a possible reluctance to publish negative findings or subject examination processes to external scrutiny[116] have been suggested as possible reasons for this gap in the assessment literature. It also suggests an ongoing emphasis on the psychometric evaluation of individual assessment instruments rather than evaluation of the psychometric rigour of composite assessment packages, previously identified as the preferred way of assessing the multiple, complex dimensions of professional competence.[4,6,7]

In Chapter 5 of this thesis, as already mentioned, I describe the composition of a multi-component postgraduate specialist certification examination conducted in South Africa. A major part of the paper focuses on the use of multivariate generalizability theory to determine the component and composite reliability of the examination and objectively identify resource-appropriate ways of improving the examination reliability by altering the examination composition using the results of data obtained from a series of decision studies.

As has already been mentioned, validity is not a property of any test procedure. Rather, it refers to the inferences made from the test results. Essentially this means that the test itself cannot be described as valid; the inferences made from the test result need to be appropriate and meaningful.[8] Thus, when considering test validity one should carefully consider whether the inferences being made, on the basis of test results, are indeed valid. Kane has proposed a model in which the inferences made are likened to the links in a chain.[12] The chain is only as strong as its weakest link. According to Kane, the interpretation of a performance test score, as a reflection of an individual's competence, requires at least three inferences: (1) evaluation – the criteria embedded in the scoring system used to judge the candidate's performance need to have a clear and credible basis for differentiating good from bad performance; (2) generalisation – the generalisations made from the observed sample of performance to "conclusions about a larger universe of similar observations"; and (3) extrapolation – extrapolation from the "behaviour actually observed to the behaviour of ultimate interest".[12] Kane has emphasised the importance of recognising that all assessment processes make inferences at these three levels. Validity at

each level needs to be carefully examined. Kane has also used the analogy of linking bridges to explain his model.[210]

Having outlined two key concepts regarding validity, I return now to the expanded concept of validity. This is a lengthy issue that has been extensively reviewed. I draw most of this brief discussion from a particularly useful recent review written by Roy Killen from the University of Newcastle in Australia.[199] Until the 1980s validity was viewed as a multi-faceted concept comprising five separate components: (1) content validity – the extent to which test measures what has been learnt or taught; (2) construct validity – the extent to which a test score reflects human behaviour or ability, i.e. the meaning of a test score; (3) consequential validity – the consequences of a test score, e.g. whether a candidate is awarded a degree or not, (4) concurrent validity – the extent to which one test score agrees with some other test score measuring the same attributes or abilities; and (5) predictive validity – the extent to which a test score is able to predict future performance. This cumbersome view of validity has been considerably simplified – "validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment".[211] The basic question being asked is: "Am I making justifiable inferences and decisions on the basis of the evidence I have gathered?"[199] The evidence that needs to be gathered to defend the decisions being made is still clustered into the same five categories referred to earlier.[8] However, these categories now represent the types of evidence (e.g. construct-related evidence or content-related evidence) needed to validate inferences made, rather than the types of validity sought. Further discussion of issues relevant to validity is not required for the purpose of this thesis. The key issues, as they relate to the papers discussed later, have been addressed.

Parameters determining test practicality. The practicality (feasibility) of assessment processes is primarily determined by the resources, human and infrastructure (including equipment), required to perform the test procedure and the acceptability of the test procedure to the primary stakeholders, i.e. the examiners and the examinees. These two factors are briefly outlined and the relevance of each, to the thesis, is highlighted.

The resources required to conduct an assessment process are entirely dependent upon the format of the test process. For example written tests usually require a venue, adequate seating and desk space for all candidates, basic stationery and supervision by one of more members of staff. By contrast, an OSCE examination comprising 15 stations requires (1) one of more examiners per station (15-30 examiners), (2) a selection of patients, other clinical material (e.g. electrocardiograph tracings or chest radiographs), and /or equipment required to perform specific tasks – for example, life-size human manikins and equipment for performing procedures such as endotracheal intubation, cardiac defibrillation, etc., and (3) a venue large

enough to accommodate 15 separate spaces which vary in size according to the nature of the task to be performed. Often the venue needs to accommodate a number of beds, chairs and desks and other surfaces suitable for performing procedures on in a safe manner. The striking difference between the resource requirements of written assessment tools and performance-based assessment processes is apparent. The cost of acquiring, storing and maintaining all the relevant equipment needed for performance-based examinations further amplifies the vastly different resource requirements of these two fundamentally different assessment processes.

While the resource requirements of performance-based assessment processes are well recognised in the developed world, they seldom dictate the use of a limited number of resource-efficient or "cheap" assessment processes.[34] Widespread implementation of the OSCE approach to the assessment of clinical competence in the developed world provides good evidence in this regard. It is, therefore, not surprising that very little is written in the literature about the practicability of assessment processes in well-resourced settings.

In severely resource-constrained environments typical of developing countries, however, performance-based assessment processes are difficult to initiate and sustain. Although hospitalised patients or clinic attendees often serve as examination cases in bedside oral examinations, the need for a cohort of examiners and a suitable venue in which to conduct the examination remain significant limiting factors. While infrastructural constraints, such as a venue in which to conduct a multi-station OSCE, can be accommodated to a greater or lesser extent, the major limitation of performance-based assessment in the developing world is usually the lack of sufficient clinician-educators to serve as examiners. The reasons for this are discussed later in this chapter.

The challenges in resource-constrained settings are twofold: (1) adapt resource-intensive assessment processes in such a manner that they can be sustained given the relevant resource constraints,[3] and (2) use published psychometric data to determine the most efficient use of examiners, often the critical limiting parameter. For example, it is recognised that the reliability of a performance assessment test is only marginally improved by using pairs of examiners as compared to single examiners.[38,40,203,207] Hamdy and colleagues demonstrated that the reliability coefficient of a 4-case DOCEE (180 minute test) was improved from 0.82 to 0.84 by using two examiners per case rather than one examiner.[38] Similarly, Wass and colleagues showed that two examiners improved the reliability coefficient of a 9-case BOE (180-minute test) from 0.82 to 0.83.[40] These two examples clearly demonstrate the inefficiency of using pairs of examiners rather than single examiners, particularly in resource-constrained settings. Unfortunately, however, the use of inefficient assessment methods frequently persist without giving due consideration to these fundamental issues.

The final determinant of assessment utility, the acceptability of assessment processes to the relevant stakeholders, is not extensively discussed in the literature. Norman and colleagues addressed the question some 15 years ago, and concluded that student perceptions of tests were influenced by their beliefs about the fairness of the test, the perceived educational value of the test, and the intended use of the test results.[36] In the same paper they also discuss the factors impacting upon examiner satisfaction – the time spent on preparing, conducting and marking the test, the time required to train examiners and a belief in the intrinsic value of the test. More recently, Norcini and colleagues evaluated examiner and examinee satisfaction with the use of the mini-CEX assessment process.[46] Overall, both examiners and examinees were satisfied. They did not explore factors contributing to user satisfaction. Thus, although the acceptability of a test process to both examiner and examinee is clearly important, it has not been a major focus of attention the medical education literature.

This aspect of assessment utility is addressed in Chapter 6 of this thesis where the acceptability of a bedside formative assessment process, implemented in an undergraduate 4th year clinical clerkship programme at the University of Cape Town, is examined from the perspective of both students and clinician-educators. Specific issues evaluated, relevant to user acceptability, include the perceived educational value, fairness, validity and feasibility (time required to complete the task) of the assessment process. The results are discussed in the paper.

Rational selection of assessment methods using utility parameters

The key parameters that determine the overall utility of assessment processes have already been outlined. While we understand that these factors significantly influence our choice of test selection, no work has been done on developing a rational method for assessment tool selection based on these key utility parameters. A superb model facilitating the rational selection, in the context of multiple options does, however, exist in the medical literature. Polypharmacy, a term used to refer to the unnecessary and /or inappropriate use of multiple drugs in the management of patients, is a common occurrence in daily clinical practice. Unfortunately this practice often leads to adverse drug side effects and drug interactions, the outcome of which may range from harmless to fatal. The principal reason for the persistence of this undesirable clinical practice is that prescribing clinicians often fail to objectively consider a few key issues before selecting drugs for the pharmacotherapeutic treatment of medical problems. The World Health Organization (WHO) recognised this problem more than a decade ago and developed a simple model of rational drug prescribing that requires consideration of four key issues when selecting a drug for therapeutic use: (1) suitability of the drug for the intended purpose, (2) efficacy of the drug, (3) cost of the drug, and (4) the safety profile of the drug.[11] In the WHO model each factor is assigned a numerical score according to the prescriber's perception of the importance of the factor to a specific patient context. After a final

score for each of a range of potential drugs has been calculated, the drug with the highest score is selected as the favoured therapeutic option. This model of drug prescribing is widely advocated in many developing countries. It should be apparent that this tool is not directed at expert clinicians who have a vast experience of drug prescribing. Rather, it focuses on providing less experienced clinicians with an objective way of selecting appropriate drugs after due consideration of all the critical issues that usually bedevil appropriate drug selection.

This model of rational drug selection, based on identified utility parameters, is ideally suited for the purpose of rational assessment process selection, based on the utility parameters previously identified. Such a tool, designed to assist clinicians with limited medical education expertise in rationally selecting resource-appropriate assessment methods for use in medical training programme assessment processes, would be of great value in the developing world. From a developed world perspective it may seem improbable that clinicians without formal medical education expertise may be required to design and implement medical training programme assessment packages. However, in the resource-constrained environments typical of developing countries this is, unfortunately, common practice. The truth of this statement is supported by the observation that South Africa, the most affluent country in sub-Saharan Africa, does not have an Office of Medical Education in each of its eight medical schools. Certainly a measure of medical education expertise exists in all these eight schools, but educational expertise in African medical schools is generally a scarce resource. Given the relative affluence of South Africa, as compared to other African countries, it is unlikely that the situation in poor developing countries is better than in South Africa. Published data to support or refute this statement are not available.

Faced with this reality I, and a number of colleagues from other African medical training facilities, developed an assessment selection tool based on the WHO drug prescribing model. A description of the model and evaluation of some assessment tools commonly used in African medical schools is provided in Chapter 9 of this thesis.

Summary

In this literature review I have provided a broad overview of some of the most important advances in medical education assessment practice achieved over the past three decades. In terms of the work described in this thesis, four specific themes have been highlighted: (1) the use of assessment to measure clinical competence, (2) the use of assessment to facilitate student learning, (3) the use of assessment to initiate and sustain curriculum change, and (4) the selection of assessment tools on the basis of their utility. Each theme is briefly summarised.

Clinical competence, the cornerstone of professional practice, is best thought of as the extent to which an individual can use the relevant knowledge, skills and judgement required to perform effectively within the scope of practice defined by the profession. It is, therefore, constituted by a relationship between an individual and his or her work and cannot be directly observed. Hence, competence is inferred from performance. A popular taxonomy of professional competence, measured by observing performance, is provided by "Miller's pyramid of competence". This simple taxonomy elegantly stratifies the hierarchical nature of professional competence using four verbs: knows, knows how, shows how and does. Furthermore, it distinguishes between the knowledge required to complete a clinical task (knows, knows how) and the ability to proficiently perform the task (shows how, does). The latter two levels only differ with regard to the physical location in which the tasks are performed, i.e. "shows how" tasks take place in a simulated environment (in vitro), e.g. a clinical skills laboratory, while "does" tasks are performed in the clinical workplace (in vivo). This taxonomy readily classifies the plethora of assessment strategies that have been developed over the past 30 years e.g. written tests most appropriately test the lower levels of competence, while patient-based clinical tests examine the upper levels of the taxonomy. Examples of written tests include multiple choice questions (MCQ) and short-answer questions (SAQ), while in vitro clinical examinations include objective structured clinical examinations (OSCE), traditional bedside oral examinations (BOE) and directly observed clinical encounter examinations (DOCEE). Currently used examples of in vivo clinical assessment methods include clinical work sampling (CWS), mini clinical encounter examinations (mini-CEX) and portfolios of learning.

Four major factors have led to the plethora of assessment methods currently in use. They are: (1) the dual purpose of student assessment, i.e. to make judgement decisions (summative) and to provide feedback so as to facilitate learning (formative); (2) the hierarchical nature of competence, already articulated by the levels indicated in Miller's pyramid of competence; (3) the variable psychometric adequacy of individual assessment instruments; (4) the educational and vocational alignment of assessment processes. The variable psychometric adequacy of performance tests is largely a function of the number of test items included in an assessment event. Given three hours of testing time, the OSCE, DOCEE and BOE all achieve a reliability coefficient of 0.8 or more. Similarly 10 or more patient encounters of approximately 25 minutes each, using the CWS or mini-CEX strategy, achieves a similar reliability coefficient. In the context of assessment, educational alignment refers to the concordance between learning programme outcomes and assessment processes (method and content) used to measure achievement of these learning outcomes. Vocational alignment refers to the concordance between programme outcomes and the demands of professional clinical practice. Educational

concordance is essential to prevent learning programmes being driven entirely by the "backwash" effects of assessment, while vocational alignment ensures that graduates are adequately equipped to deal with clinical service delivery. The need to use assessment methods that demonstrate relevance in both regards is apparent.

Key reasons for the growing popularity of portfolios include: (1) the capacity to structure portfolio tasks in such a manner that educational and vocational concordance are ensured, (2) the professional authenticity of portfolio learning, both in terms of task(s) and location, (3) the opportunity for students to demonstrate growth of competence over a period of time, i.e. the assessment process is not a single "snapshot" of the trainee's competence, and (4) the ability of portfolio learning programmes to provide a structured educational basis for clinical clerkships in which learning activities are often poorly structured and most learning is opportunistic. The challenges of portfolio assessment – excessive examination time per candidate, limited psychometric adequacy and questions regarding the suitability of current assessment methods – are the key reasons why this innovative learning tool has not yet found universal application. Finally, it is well recognised that no one assessment tool can comprehensively assess the many outcome competencies of medical training programmes. For this reason, multi-component assessment packages, using a variety of testing instruments, are widely advocated in the literature and increasingly being used in practice.

The literature outlines three strategies that should be used to ensure that assessment facilitates student learning. Firstly, educational concordance is of paramount importance because of the profound influence assessment practices have on student learning. Secondly, feedback, the key component of formative assessment, should be used to guide and direct student learning. The lack of observed trainee performance is the key reason why most formative assessment strategies fail to significantly impact on student learning, particularly in clinical clerkships. The need to change this situation is a current focus of attention in the literature. Thirdly, the strategic use of summative assessment, for judgement purposes, to steer student learning towards a more desirable approach is critically important. Although it is well documented that "assessment drives learning", it is a fairly recent development in our thinking that assessment should be purposely used to manipulate student learning behaviour. There is a need for more published data on methods exploiting this strategy.

The use of assessment results (student performance data) to initiate and sustain curriculum change, the programme evaluation function of assessment is an emerging role. Two examples from the literature demonstrate the principle. Firstly, the limited procedural skills competence of new graduates, recognised for more than two decades, has recently started impacting on curriculum design. Long overdue changes, to remediate this demonstrated curriculum deficiency in undergraduate medical programmes, are being effected internationally.

Secondly, despite little evidence in the literature that problem-based learning (PBL) benefits the academic performance of students, as measured by traditional knowledge tests, this method of instruction provides numerous other educational benefits which endorse it use worldwide. The use of student retention rates, as an indicator of the success of PBL, highlighted by a recent paper from South Africa indicating better retention rates in PBL programmes, deserves further work. More data is needed to support this finding. Finally, new data, emerging in the literature, is providing a better understanding of strategies that can be used to refine and improve the PBL process.

Parameters that determine the fitness for purpose, or utility, of assessment practices include: reliability, validity, feasibility, acceptability and resource requirements. Educational impact, another fundamental determinant of assessment utility, was not discussed in this section since it was previously addressed. While a clear understanding of each of these parameters exists in the literature, there is need for a simple, robust way of using these parameters to make rational, i.e. educationally sound, resource-based, decisions when selecting assessment tools. The potential value of such a strategy in the developing world, given the resource constraints and limited formal medical education training of most developing world clinician-educators, is apparent. A model of rational drug prescribing, developed by the World Health Organization, serves as a useful example of a model that could be adapted to develop a tool for the evaluation of assessment utility.

Concluding remarks

This review of the medical education literature, restricted to four key themes, has highlighted the most significant advances over the past 35 years. Most of these assessment advances have been implemented in, and have impacted upon, medical training programmes in developed regions of the world. Much less is known about their use in medical training programmes in resource-constrained settings typical of developing world regions. In Chapter 2 of this thesis I provide an outline of the current state of medical education in a sub-Saharan African country, South Africa, before moving on to address six questions which explore specific aspects of the challenges faced by medical educators attempting to implement these major assessment practice advances in resource-constrained settings.

References

1. Friedman Ben-David M. The role of assessment in expanding professional horizons. Medical Teacher 2000; 22: 472-477.
2. Shumway JM, Harden RM. AMEE medical education guide no. 25. The assessment of learning outcomes for the competent and reflective practitioner. Medical Teacher 2003; 6: 569-584.
3. Schuwirth LWT, van der Vleuten CPM. Changing education, changing assessment, changing research? Medical Education 2004; 38: 805-812.
4. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education 1996; 1: 41-67.
5. Friedman M, Mennin SP. Rethinking critical issues in performance assessment. Academic Medicine 1991; 66: 390-395.
6. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. Educational Researcher 1995; 24: 5-11.
7. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Medical Education 2005; 39: 309-317.
8. Gronlund NE. Assessment of student achievement. 6th ed. Needham Heights, MA: Allyn and Bacon; 1998.
9. Crossley J, Humphris G, Jolly B. Assessing health professionals. Medical Education 2002; 36: 800-804.
10. Tutarel O. Geographical distribution of publications in the field of medical education. BMC Medical Education 2006; 2:3. Accessed on 14 October 2006. URL: http://www.biomedcentral.com/1472-6920/2/3
11. De Vries TPGM, Henning RH, Hogerzeil HV, Fresle DA. Guide to good prescribing. A practical manual. Geneva: World Health Organization – Action programme on essential drugs; 1994. Accessed on 16 October 2006. URL:http://www.dundee.ac.uk/facmedden/APT/downloads/Resource%20Materials/WHO%20prescribing%20guide.pdf
12. Kane MT. The assessment of professional competence. Evaluation & the Health Professions 1992; 15: 163-182.
13. Hager P, Gonczi A, Athanasou J. General issues about assessment of competence. Assessment & Evaluation in Higher Education 1994; 19: 3-16.
14. Bloom B. Taxonomy of educational objectives. The classification of educational goals. Handbook 1: cognitive domain. New York: David Mackay; 1971.
15. Nitko A. Educational assessment of students. 3rd ed. Merril: Prentice-Hall; 2001.

16. Biggs J. Enhancing teaching through constructive alignment. Higher Education 1996; 32: 347-364.
17. Biggs JB, Collis KF. Evaluating the quality of learning: the SOLO taxonomy. New York: Academic Press; 1982.
18. Miller GE. The assessment of clinical skills /competence /performance. Academic Medicine 1990; 65 (Suppl.): S63-S67.
19. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001; 357: 945-949.
20. Sadler R. Formative assessment and the design of instructional systems. Instructional Science 1989; 18: 119-144.
21. Gipps C. Socio-cultural aspects of assessment. Review of Educational Research 1999; 24: 355-392.
22. Shepard L. The role of assessment in a learning culture. Educational Researcher 2000; 29: 4-14.
23. Schuwirth LWT, van der Vleuten CPM. Challenges for educationalists. British Medical Journal 2006; 333: 544-546.
24. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine. Written assessment. British Medical Journal 2003; 326: 643-645.
25. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education 2004; 38: 974-979.
26. Schuwirth LWT, Verheggen MM, van der Vleuten CPM, Boshuizen HPA, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? Medical Education 2001; 35: 348-356.
27. Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. Teaching and Learning in Medicine 1993; 5: 107-115.
28. Rimoldi HJA. The test of diagnostic skills. Journal of Medical Education 1961; 36: 73-79.
29. McGuire CH, Babbott D. Simulation technique in the measurement of problem-solving skills. Journal of Educational Measurement 1967; 4: 1-10.
30. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Medical Education 1985; 19: 238-247.
31. Muzzin LJ. Oral examinations. In: Neuveldt VR, Norman GR, editors. Assessing clinical competence. New York: Springer; 1985. p. 71-93.
32. Meskauskas JA. Studies of the oral examination: the examinations of the subspeciality Board of Cardiovascular Disease of the American Board of Internal Medicine. In: Lloyd

JS, Langsley DG, editors. Evaluating the skills of medical specialists. Chicago: American Board of Medical Specialties; 1983.

33. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Medical Education 1979; 13: 41-51.
34. Smee S. Skill based assessment. British Medical Journal 2003; 326: 703-706.
35. Federation of Royal Colleges of Physicians of the United Kingdom. PACES: Practical assessment of clinical examination skills. Journal of the Royal College of Physicians of London 2000; 34: 57-60.
36. Norman GR, van der Vleuten CPM, de Graaf E. Pitfalls in the pursuit of objectivity: issues of validity/ efficiency and acceptability. Medical Education 1991; 25:119-126.
37. Abouna GM, Hamdy H. The integrated direct observation clinical encounter examination (IDOCEE) – an objective assessment of students' clinical competence in a problem-based learning programme. Medical Teacher 1999; 21:67-72.
38. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). Medical Education 2003; 37:205-212.
39. Daelmans HEM, Scherpbier AJJA, van der Vleuten CPM, Donker AJM. Reliability of clinical oral examination re-examined. Medical Teacher 2001; 23: 422-424.
40. Wass V, Jones R, van der Vleuten C. Standardised or real patients to test clinical competence? The long case revisited. Medical Education 2001; 35: 321-325.
41. Norcini J. The validity of long cases. Medical Education 2001; 35: 720-721.
42. Norman G. The long case versus objective structured clinical examinations. British Medical Journal 2002: 324: 748-749.
43. Norcini JJ. The death of the long case? British Medical Journal 2002; 324: 408-409.
44. Wass V, van der Vleuten C. The long case. Medical Education 2004; 38; 1176-1180.
45. Brennan BG, Norman GR. Use of encounter cards for evaluation of residents in obstetrics. Academic Medicine 1997; 72 (Suppl.1): S43-S44.
46. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Annals of Internal Medicine 1995; 123: 795-799.
47. Hatala R, Norman GR. In-training evaluation during an internal medicine clerkship. Academic Medicine 1999; 74 (Suppl.): S118-S120.
48. Turnbull J, MacFayden J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. Journal of General Internal Medicine 2000; 15: 556-561.
49. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. Academic Medicine 2002; 77: 900-904.

50. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: A method for assessing clinical skills. Annals of Internal Medicine 2003; 138: 476-481.
51. Norcini JJ. The mini clinical evaluation exercise (mini-CEX). Clinical Teacher 2005; 2: 25-30.
52. Mcleod PJ, Meagher TW. Educational benefits of blinding students to information acquired and management plans generated by other physicians. Medical Teacher 2001; 23:83-85
53. Boelen C, Boyer MH. A view of the world's medical schools. Defining new roles. 2001. Accessed on 8 June 2006.
URL: http://www.the-networktufh.org/download.asp?file=med_schools.pdf
54. Snadden D, Thomas M. The use of portfolio learning in medical education. Medical Teacher 1998; 192-199.
55. Challis M. AMEE. Medical education guide no. 11 (revised). Portfolio-based learning and assessment in medical education. Medical Teacher 1999; 21: 370-386.
56. Webb C, Endacott r, Grey MA, Jasper MA, McMullam M, Scholes J. Evaluating portfolio assessment systems: what are the appropriate criteria? Nurse Education Today 2003; 23:600-609.
57. Davis MH, Friedman Ben-David M, Harden RM, Howie P, Ker J, McGhee C, et al. Portfolio assessment in medical students' final examination. Medical Teacher 2001; 23: 357-366.
58. Driessen EW, van Tartwijk J, Vermint JD, van der Vleuten. Use of portfolios in early undergraduate medical training. Medical Teacher 2003; 25: 18-23.
59. Driessen EW, van der Vleuten C, Schuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. Medical Education 2005; 39: 214-220.
60. Snadden D, Thomas ML. Portfolio learning in general practice vocational training – does it work? Medical Education 1998; 32: 401-406.
61. Challis M, Mathers NJ, Howe AC, Field NJ. Portfolio-based learning: continuing medical education for general practitioners – a mid-point evaluation. Medical Education 1997; 31:22-26.
62. Friedman Ben David M, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. AMEE medical education guide no. 24. Portfolios as a method of medical assessment. Medical Teacher 2001; 23:535-551.
63. Institute for International Medical Education (IIME). Global minimum requirements in medical education. Medical Teacher 2002; 24: 130-135.
64. Kolb DA. Experiential learning. Chicago: Prentice Hall; 1984.

65. Knowles M. Andragogy: an emerging technology for adult learning. In: Tight M, editor. Education for Adults: adult learning and education. London: Croom Helm; 1970.
66. Riegel KF. Dialectic operations: the final period of cognitive development. Human Development 1973; 16:346-370.
67. Mezirow J. A critical theory of adult learning and education. Adult Education 1981; 32: 3-24.
68. Schon D. The reflective practitioner: how professionals think in action. London: Basic Books; 1983.
69. Irby DM. Teaching and learning in the ambulatory setting, a thematic review of the literature. Academic Medicine 1995; 70: 898-931.
70. Jolly BC, Macdonald MM. Education for practice: the role of practical experience in undergraduate and general clinical training. Medical Education 1989; 23: 189-195.
71. Remmen R, Denekens J, Scherpbier A, Hermann I, van der Vleuten CPM. An evaluation of the study of the didactic quality of clerkships. Medical Education 2000; 34: 460-464.
72. Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LTW, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. Medical Teacher 2000; 23:600-609.
73. Van den Hem-Stokroos HH, Scherpbier AJJA, van der Vleuten CPM, de Vries H, Haarman HJTM. How effective is a clerkship as a learning environment? Medical Teacher 2001; 23: 608-613.
74. Daelmans HEM, Hoogenboom RJI, Donker AJM, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Effectiveness of clinical rotations as a learning environment for achieving competence. Medical Teacher 2004; 26: 305-312.
75. Samuel S, Shaffer K. Profile of medical student teaching in radiology: methods, staff perceptions and rewards. Academic Radiology 2000; 7: 868-874.
76. Regan-Smith M, Young WW, Keller AM. An efficient and effective teaching model for ambulatory medicine. Academic Medicine 2002; 77: 593-599.
77. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangora LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? Annals of Internal Medicine 1992; 117: 757-765.
78. Marsick VJ, Watkins KE. Informal and incidental learning in the workplace. London: Routeledge; 1990.
79. Driessen EW, van Tartwijk J, Overeem K, Vermunt JD, van der Vleuten CPM. Conditions for successful reflective use of portfolios in undergraduate medical education. Medical Education 2005; 39: 1230-1235.

80. Boud D, Keogh R, Walker D, editors. Reflection: turning experience into learning. London: Kogan Page; 1985.
81. Boud D, Cohen R, Walker D, editors. Using experience for learning. Buckingham: Open University Press; 1993.
82. Mamede S, Schmidt HG. The structure of reflective practice in medicine. Medical Education 2004; 38: 1302-1308.
83. Epstein RM. Mindful practice. Journal of the American Medical Association 1999; 282:833-839.
84. Pearson DJ, Heywood P. Portfolio use in general practice vocational training: a survey of GP registrars. Medical Education 2004; 38:87-95.
85. Wade RC, Yarbrough DB. Portfolios: a tool for reflective thinking in teacher education? Teaching and Teacher Education 1996; 12: 63-79
86. World Health Organization (WHO). Health and the millennium development goals. Geneva: WHO; 2005. Accessed on 04 September 2006.
URL: http://www.who.int/mdg/publications/mdg_report/en/index.html
87. Walubo A, Burch V, Parmar P, Raidoo D, Cassimjee M, Onia R, et al. A model for selecting assessment methods for evaluating medical students in African medical schools. Academic Medicine 2003; 78:899-906.
88. Roberts C, Newble DI, O'Rourke AJ. Portfolio-based assessments in medical education: are they valid and reliable for summative purposes? Medical Education 2002; 36:899-900.
89. Pitts J, Coles C, Thomas P. Educational portfolios in the assessment of general practice trainers: reliability of assessors. Medical Education 1999; 33:515-520.
90. Pitts J, Coles C, Thomas P. Enhancing the reliability in portfolio assessment: "shaping" the portfolio. Medical Teacher 2001; 23:351-355.
91. Pitts J, Coles C, Thomas P, Smith F. Enhancing reliability in portfolio assessment: discussions between assessors. Medical Teacher 2002; 24:197-201.
92. Karlowicz KA. The value of student portfolios to evaluate undergraduate nursing programmes. Nurse Educator 2000; 25: 82-87.
93. Challis M. Portfolios and assessment: meeting the challenge. Medical Teacher 2001; 23: 437-440.
94. Cohen SA. Instructional alignment: Searching for a magic bullet. Educational Researcher 1987; 16: 16-20.
95. Spady WG. Organising for results: the basis of authentic restructuring and reform. Educational Leadership 1988; October: 4-8.
96. Harden RM, Crosby JR, Davis MH. AMEE guide no. 14. Outcome-based education: part 1. An introduction to outcome-based education. Medical Teacher 1999; 21: 7-14.

97. Harden RM, Crosby JR, Davis MH, Friedman M. AMEE guide no.14. Outcome-based education: part 5. From competency to meta-competency: a model for the specification of learning outcomes. Medical Education 1999; 21: 546-552.
98. Metz JCM, Stoelinga GBA, Pels R, van Erp T, Kip EH, van den Brand-Valkenburg BWM. Blueprint 1994: training of doctors in the Netherlands. Nijmegen: University of Nijmegen Publications Office; 1994.
99. Metz JCM, Verbeek-Weel AMM, Huisjes HJ. Blueprint 2001: training of doctors in the Netherlands. Adjusted objectives of undergraduate medical education in the Netherlands; 2001. Accessed on 29 October 2006.
URL:http://www.lumc.nl/5030/rapportages/documenten/KRUL%20Voortgangsrapportage%20200310%20_%20200312.pdf
100. Smith SR, Dollase RH. Planning, implementing and evaluating a competency-based curriculum. Medical Teacher 1999; 21: 5-22.
101. Mandin H, Harasym P, Eagle P, Watanabe M. Developing a "clinical presentation" curriculum at the University of Calgary. Academic Medicine; 1995; 70: 186-193.
102. Royal College of Physicians and Surgeons of Canada. CanMEDS 2005 Framework. Accessed on 29 October 2006.10.29
URL: http://www.healthcare.ubc.ca/residency/CanMEDS_2005_framework.pdf
103. Accreditation Council for Graduate Medical Education (ACGME). Outcome project: the general competencies. Accessed on 29 October 2006.
URL: http//www.acgme.org
104. Snyders BR. The hidden curriculum. Cambridge, MA: MIT Press; 1971.
105. Wakeford R, Roberts S. An evaluation of medical students' practical experience upon qualification. Medical Teacher 1982; 4: 140-143.
106. Martin YM, Harris DL, Karg MB. Clinical competencies of graduating medical students. Journal of Medical Education 1985; 60: 919-925.
107. Kowlowitz V, Curtis P, Sloane PD. The procedural skills of medical students: expectations and experiences. Academic Medicine 1990; 65: 656-658.
108. Board P, Mercer M. A survey of the basic practical skills of final-year medical students in one UK medical school. Medical Teacher 1998; 20: 104-108.
109. Hannon FB. A national medical education needs' assessment of interns and the development of an intern education and training programme. Medical Education 2000; 34: 275-284.
110. Ringstedt C, Schroeder TV, Henriksen J, Ramsing B, Lyngdorf P, Jønsson V, et al. Medical students' experience in practical skills is far from stakeholders' expectations. Medical Teacher 2001; 23: 412-416.

111. Barnsley L, Lyon PM, Ralston SJ, Hibbert EJ, Cunningham I, Gordon FC, et al. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. Medical Education 2004; 38: 358-367.
112. Fox RA, Ingham Clark CL, Scotland AD, Dacre JE. A study of pre-registration house officers' clinical skills. Medical Education 2000; 34: 1007-1012.
113. Gastel B. Towards global consensus on quality medical education: serving the needs of populations and individuals. Summary of the consultation. Academic Medicine 1995; 70 (Suppl.):S3-S7.
114. Norcini JJ. Work based assessment. British Medical Journal 2003; 326: 753-755.
115. Davies H. Work based assessment. British Medical Journal 2005; 331: 88-89.
116. Hutchinson L, Aitken P, Hayes P. Are medical postgraduate certification processes valid? A systematic review of the evidence. Medical Education 2002; 36: 73-91.
117. Lew SR, Page CG, Schuwirth LW, Baron-Malondolo M, Lescop JM, Paget NS, et al. Procedures for establishing defensible programmes for assessing practice performance. Medical Education 2002; 36:936-941
118. Hays RB, Fabb WE, van der Vleuten CPM. Reliability of the Fellowship examination of the Royal Australian College of General Practitioners. Teaching and Learning in Medicine 1995; 1: 43-50.
119. Thompson AN. An assessment of a postgraduate examination of competence in general practice: part I- reliability. New Zealand Medical Journal 1990; 103:182-184.
120. Thompson AN. An assessment of a postgraduate examination of competence in general practice: part II- validity. New Zealand Medical Journal 1990; 103:1217-219.
121. Handfield-Jones R, Brown JB, Rainsberry P, Brailovsky CA. Certification examination for the College of Family Physicians of Canada. Part II: conduct and general performance. Canadian Family Physician 1996; 42:1188-1195.
122. Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? Medical Education 2001; 35: 326-330.
123. Newble DI, Jaeger K. The effect of assessments and examination on the learning of medical students. Medical Education 1983; 17: 165-171.
124. Frederiksen N. The real test bias. Influences of testing on teaching and learning. American Psychologist 1984; 39: 193-202.
125. Crooks TJ. The impact of classroom evaluation practices on students. Review of Educational Research 1988; 58:438-481.
126. Stillman PL, Haley H-L, Regan MB, Philbin MM. Positive effects of a clinical performance assessment programme. Academic Medicine 1991; 66: 481-483.

127. Driessen E, van der Vleuten C. Matching student assessment to problem-based learning: lessons from experience in a law faculty. Studies in Continuing Education; 2000; 22: 235-248.
128. Branch WT, Paranjape A. Feedback and reflection: teaching methods for clinical settings. Academic Medicine 2002; 77: 1185-1188.
129. Gibbs G, Simpson C. Conditions under which assessment supports student learning. Learning and Teaching in Higher Education 2004-05; 1: 3-31.
130. Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the mini clinical evaluation exercise. Journal of General Internal Medicine 2004; 19: 558-561.
131. Blank LL, Grosso LJ, Benson JA. A survey of clinical skills evaluation practices in internal medicine residency programmes. Journal of Medical Education 1984; 59:401-406.
132. Szenas P. The role of faculty observation in assessing students' clinical skills. Contemporary Issues in Medical Education 1997; 1:1-2.
133. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behaviours in medical school. Academic Medicine 1999; 74: 842-849.
134. Daelmans HEM, Hoogenboom RJI, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Effects of an in-training assessment programme on supervision of and feedback on competencies in an undergraduate Internal Medicine clerkship. Medical Teacher 2005; 27: 158-163.
135. Van den Hem-Stokroos HH, Daelmans HEM, van der Vleuten CPM, Harrman HJThM, Scherpbier AJJA. The impact of multifaceted educational structuring on learning effectiveness in a surgical clerkship. Medical Education 2004; 38: 879-886.
136. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Annals of Internal Medicine 2004; 140: 874-881.
137. Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. Academic Medicine 2004; 79: 16-22.
138. Kogan JR, Hauer KE. Brief report: use of the mini-clinical evaluation exercise in internal medicine core clerkships. Journal of General Internal Medicine 2006; 21: 501-502.
139. Lane JL, Gottlieb RP. Structured clinical observations: a method to teach clinical skills with limited time and financial resources. Paediatrics 2000; 105: 973-977.
140. Paukert JL, Richards ML, Olney C. An encounter card system for increasing feedback to students. American Journal of Surgery 2002; 183: 300-304.
141. Messick S. The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 1994; 23:13-23.

142. Gibbs G. Using assessment strategically to change the way students learn. In: Brown S, editor. Assessment matters in higher education. Choosing and using diverse approaches. Buckingham: Society for Research into higher Education and Open University Press; 1999.
143. De Lima AA, Henquin R, Thierer J, Paulin J, Lamari S, Belcastro F, et al. A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. Medical Teacher 2005; 27: 46-52.
144. Marton F, Säljö R. On qualitative differences in learning: 1 – Outcome and process. British Journal of Educational Psychology 1976; 46: 4-11.
145. Marton F, Säljö R. Approaches to learning. In: Marton F, Hounsell DJ, Entwistle NJ, editors. The experience of learning. Edinburgh: Scottish Academic Press; 1984.
146. Entwistle NJ, Hanley M, Hounsell DJ. Identifying distinctive approaches to studying. Higher Education 1979; 8: 365-380.
147. Entwistle NJ, Ramsden P. Understanding student learning. London: Croom Helm; 1983.
148. Issenberg SB, Mc Gaghie WC, Petrusa ER, Lee-Gordon D, Sacalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BME systematic review. Medical Teacher 2005; 27:10-28
149. Wilkes M, Bligh J. Evaluating educational interventions. British Medical Journal 1999; 318:1269-1272.
150. Morrison J. ABC of learning and teaching in medicine. Evaluation. British Medical Journal 2003; 326:385-387
151. Hutchinson L. Evaluating and researching the effectiveness of educational interventions. British Medical Journal. 1999; 318:1267-1269.
152. Schuwirth LWT, Southgate L, Page GC, Paget NS, Lescop JMJ, Lew SR, et al. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. Medical Education 2002; 36: 925-930.
153. Langdale LA, Schaad D, Wipf J, Marshall S, Vontver L, Scott CS. Preparing graduates for the first year of residency: Are medical schools meeting the need? Academic Medicine 2003; 78: 39-44.
154. Association of American Medical Colleges (AAMC). Report 1. Learning objectives for medical student education. Guidelines for medical schools. Washington, DC: Medical School Objectives Project, AAMC; 1998.
155. Association of American Medical Colleges (AAMC) and American Medical Association (AMA). Standards for accreditation of medical education programs leading to M.D. degree. Washington, DC, and Chicago: Liaison Committee on Medical Education, AAMC and AMA; 2003. Accessed on 20 April 2004
URL: http://www.lcme.org

156. General Medical Council (GMC). Tomorrow's doctors. Recommendations on undergraduate medical education. London: GMC; 2003. Accessed on 29 October 2006. URL: http://www.gmc-uk.org/education/undergraduate/tomorrows_doctors. asp
157. Scott CS, Barrows HS, Brock DM, Hunt DD. Clinical behaviours and skills that faculty from 12 institutions judged were essential for medical students to acquire. Academic Medicine 1991; 66: 106-111.
158. Bass EB, Fortin AH, Morrison G, Wills S, Mumford LM, Goroll AH. National survey of clerkship directors in internal medicine on the competencies that should be addressed in the medicine core clerkship. American Journal of Medicine 1997; 102: 564-571.
159. Mennin SP, Kaufman A. The change process and medical education. Medical Teacher 1989; 11:9-16
160. Mennin SP, Krackov SK. Reflections on relevance, resistance and reform in medical education. Academic Medicine 1998; 73(Suppl.):S60-S64
161. Bland CJ, Starnaman S, Wersal L, Moorhead-Rosenberg L, Zonia S, Henry R. Curricular change in medical schools: how to succeed. Academic Medicine 2000; 75:575-594.
162. Dolmans D, Schmidt H. The advantages of problem-based curricula. Postgraduate Journal of Medicine 1996; 72: 535-538.
163. Dolmans DHJM, de Grave W, Wolfhagen IHAP, van der Vleuten CPM. Problem-based learning: future challenges for educational practice and research. Medical Education 2005; 39: 732-741.
164. Burch VC, Sikakana CNT, Yeld N, Seggie JL, Schmidt HG. Performance of academically-at-risk students in a problem-based learning programme. A preliminary report. Advances in Health Sciences Education 2006. In press.
165. Schmidt HG. Problem-based learning: rationale and description. Medical Education 1983; 17: 11-16.
166. Mamede S, Schmidt HG, Norman GR. Innovations in problem-based learning: what can we learn from recent studies? Advances in Health Sciences Education 2006. In press.
167. Schmidt HG, Dauphinee, WD, Patel VL. Comparing effects of problem-based and conventional curricula in an international sample. Journal of Medical Education 1987; 62: 305-315.
168. Norman GR, Schmidt HG. The psychological basis of PBL. A review of the evidence. Academic Medicine 1992; 67: 557-565.
169. Albanese MA, Mitchell S. Problem-based learning: A review of literature on its outcomes and implementation issues. Academic Medicine 1993; 68: 52-81.
170. Vernon DTA, Blake RL. Does problem-based learning work? A meta-analysis of evaluative research. Academic Medicine 1993; 68: 550-563.

171. Berkson L. Problem based learning: have the expectations been met? Academic Medicine 1993; 68(Suppl.) S79-S88.
172. Colliver J. Effectiveness of problem-based learning curricula. Academic Medicine 2000; 75: 259-266.
173. Newman M. A pilot systematic review and meta-analysis of the effectiveness of problem-based learning. Newcastle, UK: Campbell Collaboration Systematic Review Group on the effectiveness of problem-based learning, University of Newcastle, Learning and Teaching Support Network; 2003.
174. Dochy F, Segers M, van den Bossche P, Gijbels D. Effects of PBL: a meta-analysis. Learning and Instruction 2003; 13: 533-568.
175. Hmelo CE, Gotterer GS, Bransford JD. A theory-driven approach to assessing the cognitive effects of PBL. Instructional Science 1997; 25: 387-408.
176. Schmidt HG, Machiels-Bongaerts, Hermans H, ten Cate TJ, Venekamp R, Boshuizen HPA. The development of diagnostic competence: comparison of a problem-based, an integrated, and a conventional medical curriculum. Academic Medicine 1996; 71: 658-664.
177. Schmidt HG, Vermeulen L, van der Molen HT. Longterm effects of problem-based learning: a comparison of competencies acquired by graduates of a problem-based and conventional school. Medical Education 2006; 40: 562-567.
178. Newble DI, Clarke RM. The approaches to learning of students in a traditional and in an innovative problem-based medical school. Medical Education 1986; 20: 267-273.
179. Boshuizen HPA, Schmidt HG, Wassmer L. Curriculum style and the integration of biomedical and clinical knowledge. In: Bouhuijs PAJ, Schmidt HG, van Berkel HJM, editors. Problem-based learning as an educational strategy. Maastricht, the Netherlands: Network Publications; 1994.
180. Van Dalen J, Kerkhofs E, van Knippenberg-van den Berg BW, van den Hout HA, Scherpbier AJ, van der Vleuten CP. Longitudinal and concentrated communication skills programmes: two Dutch medical schools compared. Advances in Health Sciences Education 2002; 7: 29-40.
181. Prince KJAH, van Eijs PWLJ, Boshuizen HPA, van der Vleuten CPM, Scherpbier AJJA. General competencies of problem-based learning (PBL) and non-PBL graduates. Medical Education 2005; 39: 394-401.
182. Colliver JA. Educational theory and medical education practice: a cautionary note for medical school faculty. Academic Medicine 2002; 77: 1217-1220.
183. Colliver J. Full-curriculum interventions and small-scale studies of transfer: implications for psychology-type theory. Medical Education 2004; 38: 1212-1213.

184. Norman GR, Schmidt HG. Effectiveness of problem-based learning curricula: theory, practice and paper darts. Medical Education 2000; 34: 721-728.
185. Norman GR, Eva KW, Schmidt HG. Implications of psychology-type theories for full curriculum interventions. Medical Education 2005; 39: 243-249.
186. Dolmans DHJM, Schmidt HG. What do we know about small group tutorials in problem-based learning? Advances in Health Sciences Education 2006. In press.
187. O'Neill P, Duplock A, Willis S. Using clinical experience in discussion within problem-based learning groups. Advances in Health Sciences Education 2006. In press.
188. Te Winkel WWR, Rikers RMJP, Loyens SMM, Schmidt HG. Influence of number of learning resources on self-directed learning in a problem-based curriculum. Advances in Health Sciences Education 2006. In press.
189. Albanese M. Problem-based learning: Why curricula are likely to show little effect on knowledge and clinical skills. Medical Education 2000; 34: 729-738.
190. Arbuckle JL, Wothke W. AMOS 4.0 user's guide. Chicago: SmallWaters Corporation; 1996.
191. De Bruin AHB, Schmidt HG, Rikers RMJP. The role of basic sciences knowledge and clinical knowledge in diagnostic reasoning: A structural equation modelling approach. Academic Medicine 2005; 80: 765-773.
192. Mennin SP, Friedman M. Evaluating innovative medical education programmes: common questions and problems. Annals of Community-Orientated Medical Education 1992; 5:123-133
193. Mennin SF, Friedman M, Skipper B, Kalishman S, Snyder J. Performances on the NBME I, II and III by medical students in the problem-based learning and conventional tracks at the University of New Mexico. Academic Medicine 1993; 68:616-624
194. Iputo JE, Kwizera E. Problem-based learning improves the academic performance of medical students in South Africa. Medical Education 2005; 39: 388-393.
195. Kwizera EN, Igumbor EU, Mazwai LE. Twenty years of medical education in rural South Africa – experiences of the University of Transkei medical school and lessons for the future. South African Medical Journal 2005; 95: 920-924.
196. Breier M, Wildschut A. Doctors in a divided society. The profession and education of medical practitioners in South Africa. Cape Town: Human Sciences Research Council Press; 2006.
197. Woodward GA. Some reflections on evaluation outcomes of innovative medical education programmes during the practice period. Paper presented at a workshop: Evaluating the outcome of the undergraduate medical course. Newcastle, Australia: Faculty of Medicine, University of Newcastle; 1991.

198. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.
199. Killen R. Validity in outcomes-based assessment. Perspectives in Education 2003; 21: 1-14.
200. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297-334.
201. Cronbach LJ. My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement 2004; 64: 391-418.
202. Royal College of Physicians and Surgeons of Canada (RCPSC). Handbook for Chairs and Members of Examinations Boards. Ottawa: RCPSC; 2000.
203. Van der Vleuten CPM, Norman GR, de Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Medical Education 1991; 25: 110-118.
204. Elstein AS, Shulman LS, Spafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge, Massachusetts: Harvard University Press; 1978.
205. Wimmers PF, Splinter TA, Hancock GR, Schmidt HG. Clinical competence: general ability or case-specific? Advances in Health Sciences Education 2006. In press.
206. Crossley J, Davies H, Humphris G, Jolly B. Generalizability: a key to unlock professional assessment. Medical Education 2002; 36: 972-978.
207. Boulet JR. Generalizability theory: the basics. In: Everitt BS, Howell DC, editors. Encyclopaedia of statistics in behavioural science. Chichester: John Wiley & Sons, Ltd; 2005.p 704-711.
208. Brennan RL. Generalizability theory. New York: Springer-Verlag; 2001.
209. Norman GR, Streiner DL. Biostatistics. The bare essentials. St. Louis, Missouri: Mosby; 1994.
210. Kane MT, Crooks T, Cohen A. Validating measures of performance. Educational Measurement: Issues and Practice 1999; 18: 5-17.
211. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd ed. New York: Macmillan; 1989