ATHLETIC TRAINING EDUCATION JOURNAL

www.nataej.org
ISSN: 1947-380X
DOI: 10.5608/0704187

LITERATURE REVIEW

Quality of Instruments Used to Assess Competencies in Athletic Training

Jim F. Schilling, PhD, LAT, ATC, CSCS

The University of Southern Maine, Gorham, ME

Context: An emphasis on knowledge and skill competency acquisition continues to gain importance in allied health professions. Accuracy and fairness in the summative assessment of competencies are essential to ensure student competence. A positive demonstration of validity, reliability, and authentic quality criteria are needed to achieve evidence-based practice considerations in the assessment of competencies.
Objective: To present a variety of instruments used in the assessment of competencies established in the fifth edition of the athletic training competencies document and judge them based on validity, reliability, and authenticity criteria.
Data Sources: Literature reviewed for this article included published articles pertaining to the assessment of competencies used in health care professional programs.
Data Synthesis: Self, written, and observation assessment methods with specific types of instruments for each category are used in the summative assessment of competencies. Quality of the assessment instruments are considered to ensure score authenticity, validity, and reliability of measures. The type of assessment instrument and its content was recommended depending on the level of competence, which was categorized according to the depth of understanding and complexity of skill in the competencies.
Conclusions: There was no one-size-fits-all assessment method determined. Certain instruments demonstrated greater quality than others and were used depending on assessment goals and resources.
Keywords: Competency, assessment, validity, reliability, authenticity

Dr. Schilling is currently an assistant professor at The University of Southern Maine. Please address all correspondence to Jim F. Schilling, PhD, LAT, ATC, CSCS, The University of Southern Maine, 37 College Avenue, Gorham, ME 04038. jschilling@usm.maine.edu.

Full Citation:
Schilling J. Quality of Instruments Used to Assess Competencies in Athletic Training
Athl Train Educ J. 2012;7(4):186-197.

Quality of Instruments Used to Assess Competencies in Athletic Training

Jim F. Schilling, PhD, LAT, ATC, CSCS

To demonstrate accountability to stakeholders and appease accreditation requirements, athletic training education programs have adopted a standardized competency-based education model whose framework consists of specific pre-determined competencies. Competency-based education programs use competencies consisting of behavioral objectives that are commonly grouped into domains such as skills, knowledge, or attitudes.[1] In regards to program acceptance and retention, an emphasis on the objective demonstration of knowledge and skill competence continues to gain significance in allied health professions including physical therapy, physician assistant, occupational therapy, and athletic training. To obtain objective data for grading, the accuracy and fairness of summative assessments are of the utmost importance. It was previously suggested that the goal of the summative assessment of competence is to act as a barrier and protect the public by upholding high standards and screening out students who are incompetent to practice.[2]

A consistent definition of being competent, or acquiring competence, is very elusive. Epstein and Hundert[3] defined competence as a professional's overall suitability for the profession and the communication abilities, knowledge, technical skills, clinical reasoning, and values he or she is expected to possess and demonstrate. Competencies have been categorized into varying levels of aptitude depending on the depth of understanding and skill complexity requirements. An example of such a model is "Miller's Pyramid," which presents four levels of competence.[4] These levels, along with corresponding assessment instruments, are presented in Table 1. An example of a level one competency listed in the fifth edition of athletic training competencies[5] is "Define evidence-based practice as it relates to athletic training clinical practice."[(p.10)] Assessment of this knowledge competency would ask for the recall of the definition. Level two assessments are meant to determine the depth of understanding an individual has for a specific competency. An example from the athletic training competencies is "Determine the criteria and make decisions regarding return to activity and/or sports participation based on the patient's current status."[5(p.17)] A level three competency asks the individual to perform a specific skill, such as "Perform joint mobilization techniques as indicated by examination findings."[5(p.24)] The highest level of competence requires a professional assessment with real patients. In such situations, the individual is required to address an integration of competencies. Using the fifth edition of the athletic training competencies as an example, this level of competence is referred to as the clinical integration proficiencies.

Assessment instruments need to meet minimum criteria to be considered effective at accurately evaluating competencies. Published literature proposes that to achieve evidence-based practice considerations in assessment design, part of the process needs to feature a positive demonstration of valid, reliable, and authentic methods.[6] Validity is a conceptual term that refers to "the extent to which a measurement actually measures what it is intended to measure."[7(p.1217)] The different types of validity for assessment techniques include predictive, content, construct, convergent, divergent, and face validity.[7] Van der Vleuten and Schuwirth[8] state that assessment instrument validity increases when offering students real-world challenges either on paper, in computerized forms, or in laboratory settings. Reliability measures the reproducibility or consistency of measurement.[9] Intra-rater reliability is the consistency of assessment by a single assessor using a specific instrument and set of circumstances, while inter-rater reliability is the consistency of multiple assessors using the same instrument and set of circumstances.[9] Authentic assessment requires students to use the same knowledge, skills, and attitudes (or the same combination of them) they need to apply in a criterion situation in professional life.[10] The level of authenticity of the assessment is defined by its degree of resemblance to the criterion situation.[10] For the assessment of knowledge and skill competencies to be authentic, the competencies need to be assessed in a context that resembles the instructional context.[11]

Table 1. Framework for Assessment

Competence Level[4]	Outcome	Competency Domain[5]	Instrument
1	Recall of Principles, Theories,Concepts	Knowledge	MC, TF
2	Problem Solving, Decision making	Knowledge	SA, EE, OE, EM, KF
3	Demonstration of Skill in Controlled Setting	Skill	SeA, M-CEX, PMP, OSCE, 360, LC, PA, CELI
4	Real-life Performance	Clinical Integrated Proficiencies	SeA, P, PA, 360 Video with Rating

MC (Multiple Choice), TF (True or False), SA (Short Answer), EE (Essay Exam), OE (Oral Exam), EM (Extended Matching), KF (Key Feature), SeA (Self-Assessment), M-CEX (Mini-Clinical Evaluation Exercise), PMP (Patient Management Problem), OSCE (Objective Structural Clinical Examination), 360 (360° Assessment), LC (Long Case), CELI (Control, Explaining, Listening, Influencing), P (Portfolio), PA (Peer Assessment).

The purpose of this review is to present instruments that can be used in the assessment of the competencies in the 5th edition of athletic training competencies. I have judged these instruments based on validity, reliability, and authenticity criteria. I have presented this information by assessment method (eg, self, written, and observation) and further divide it by the specific type of assessment instrument.

ASSESSMENT INSTRUMENTS

Self-Assessment

Self-assessment has been investigated at both the graduate and undergraduate levels.[12-15] The accuracy and role of self-assessment in attaining technical skills was explored with noncertified student surgeons using a global rating scale instrument, which showed excellent inter-item and inter-rater reliability; however, this study demonstrated a significant difference in self-assessment scores between the student surgeons and the two trained raters. [12] The students consistently scored themselves higher, thereby showing a bias for an overestimation of technical skills. A literature review investigated 17 studies of self-ratings by practicing physicians.[13] The self-assessments conducted in these studies were by questionnaire, checklists, or surveys, and they focused on learning needs, confidence in performance, clinical skills, and critical appraisal knowledge. The studies compared physician self-ratings with an external assessment including objective measures such as objective structured clinical examination scores and ratings by residents, patients, or faculty. The results of this review suggested there was no association between the physicians' self-rated assessments and external assessments, thereby putting the validity of these assessments into question. Self-assessment scores of undergraduate students have also shown a weak correlation with peer and tutor scores,[14] which further questions their validity. Self-assessment of communication skills using medical students on video was found to be feasible and informative; however, reliability and validity were not measured in this study.[15]

Using self-assessment scorings as a summative assessment instrument for competencies at the graduate and undergraduate level suffers in both reliability and validity, but it may hold some promise in athletic training foundational behaviors such as communication skills. It may also be used as a formative assessment tool and support a student's learning experiences through self-reflection. This type of instrument may be authentic by learning and assessing competencies in universal contexts and could even support adequate reliability; however, the accuracy and validity of this type of assessment is in question.

Written Assessment

Multiple Choice. Written tests such as multiple choice exams are frequently utilized for assessing knowledge. This is especially evident with certification or licensure exams where reliability and validity are essential qual-

ities to defending the accountability of such assessments.[16] While the multiple choice exam has demonstrated good reliability, its validity has been criticized.[16] A problem with the multiple choice assessment technique is that correct answers can be achieved through recognizing the answer in a list of options. This phenomenon is referred to as a "cueing effect," which can be a threat to authenticity as well as the validity of the instrument.[17] Great care should be taken when creating the multiple choice questions because writing vague and confusing questions can lead to inaccurate measures.[18]

The authenticity of multiple choice questions may suffer due to the lack of correlation between assessment and practice context.[17] Multiple choice instruments are efficient and cost effective; however, they are not context rich so they may not challenge the more complex cognitive processes needed for understanding certain subject matter. Overall, a multiple choice exam is a good instrument to choose when the goal of the assessment is to examine a large breadth of subject matter while maintaining strong reliability and reasonable validity when providing a significant number of answer choices (eg, 6-10).

True or False. As with multiple choice questions, true-or-false questions can be answered quickly and cover a broad domain while providing a reliable assessment. However, true-or-false exams are difficult to construct because the questions need to be defensibly true or absolutely false.[19] Also, when a student answers a question false, all that can be determined is that the student knows the statement is incorrect and not if the student knows the correct answer. This limitation questions the validity of the instrument. True-or-false exams are most suitable if the purpose is to test whether students are able to evaluate the correctness of an assumption.[19] As with multiple choice instruments, authenticity is also inhibited because of a lack of similarity in context with real-life situations.

Short Answer. Written assessment using short answer questions can test knowledge that requires creativity and spontaneity; however, it shows poor reliability.[19] This assessment technique requires time and demonstrates poor reliability due to inconsistency in scoring and an inadequate sample of domains tested within a given period of time.[19] If there is a limited number of realistic response alternatives, multiple-choice questions would supply adequate validity and be most suitable. If the number of alternatives is large, the open-ended nature of a short answer question is more authentic than multiple choice questions. This aspect of the short answer question could improve instrument validity and provide the assessor a greater indication of a student's understanding of the subject matter.[17]

Essay Exam. An essay exam is a reasonable assessment method if the goal is to examine a student's ability to summarize, find relationships, process information, and gain insight into his or her writing ability.[19] However, the reliability of this type of instrument is very poor. In addition, instrument validity may suffer if the questions are too structured or overly explanatory.[19] Authenticity may be demonstrated when the answer to the question is consistent with the learning context, which should resemble criterion situations. Schuwirth and van der Vleuten[19] have suggested that essay questions should be used only when short answer or multiple choice questions are considered inappropriate. The difficulty in consistently scoring essay questions greatly challenges the reliability of this type of assessment. However, scoring rubrics can guide the assessment and improve the scoring consistency.[18] Essay questions can be context rich and inquire into one's depth of understanding; however, accuracy in scoring is difficult and breadth of subject matter coverage is minimal.[19]

Extended Matching Questions. Another written assessment technique referred to as extended matching questions consists of a lead-in question, case description, and a list of options.[17] The rationale for this instrument is to create many possible combinations and minimize the "cueing effect" that occurs with a standard multiple choice exam, which should improve authenticity and assessment validity. By using cases instead of facts, this technique can be used to assess problem-solving abilities. Extended matching exams are not difficult to construct, which results in a lower time and cost commitment while still achieving acceptable reliability.[19] Although validity and authenticity have not been specifically studied, it could be argued that this type of instrument could be very positive if its questions adhere to proper content and structure.

A type of an extended matching question could be to select from a list the most likely diagnosis for a specific case. Below is an example of such a question.

A patient walks into the athletic training room with an antalgic gait, including a bent knee, and states he injured his knee when making a cut on the football field

five minutes ago. Upon observation, moderate effusion is noted. There is no apparent pain with knee palpation. The patient complains of posterior knee pain when passively moving the knee through extension. Results of special testing of the knee include negative findings for valgus and varus stress , Thessaly, patellar apprehension, and posterior and anterior drawer stress tests. A Lachman's stress test was inconclusive due to muscle guarding.

The list of possible diagnoses for this case could include the following: (a) anterior cruciate ligament tear, (b) medial collateral ligament tear, (c) posterior cruciate ligament tear, (d) medial meniscus tear, (e) lateral collateral ligament tear, (f) lateral meniscus tear, and (g) patellar subluxation.

Portfolio. A portfolio assessment technique is a collection of evidence, usually in written form, of both products and processes of learning.[20] It attempts to bring about professional development of the learner through the critical analysis of experiences. The portfolio is based on experiential learning, which promotes authenticity because the student actively integrates theory and practice with real-life situations.[20] McMullan et al[20] stated that portfolios have shown low inter-rater reliability; however, they suggested that reliability would improve by decreasing the number of learning outcomes. The tradeoff with reducing the number of learning outcomes though is that the validity of the assessment instrument will suffer.[20] Suggestions for improving instrument reliability include using trained assessors, clear guidelines, and well-defined goals without describing every minute detail.[21] Portfolios are beneficial in acquiring a student's perception of improvement in real-life skills, which signifies authenticity, and the assessment of foundational behaviors such as professional judgments and ethical issues.[18] Most significantly, portfolios encourage students to develop self-reflection and take charge of their own learning.

Patient Management Problem. A patient management problem (PMP) is constructed to assess decision-making skills. The PMP is composed of a written problem consisting of a clinical scenario followed by items that evoke an injury or illness management plan.[17] Assessors, considered experts in the clinical area, agree on the desired outcome of PMPs based on scoring rubrics but not on the process by which the outcomes are reached. The results from cognitive psychological research have demonstrated that solving complex problems requires more than selecting the correct standard solution.[17] In other words, the efficiency and effectiveness of the process by which a particular case is solved is significant and should be assessed.

Concerns with this technique include the PMP process and outcomes where the scores for intermediate students surpass the scores of experts.[17] This outcome can be explained through cognitive research, which demonstrates that experts differ from students in that their knowledge is organized more efficiently, thereby enabling them to retrieve relevant knowledge faster and solve a problem more efficiently.[17] Since the PMP rewards thoroughness, an expert's efficiency in diagnosing and managing a clinical case is discouraged and scored negatively. Also, there is a low correlation between simulations, which indicates a poor reliability for this assessment technique.[17] For example, a student's score on a PMP exam of one clinical case is a poor predictor of the same student's score on another case, even within the same domain. To improve the validity and reliability of the PMP assessment technique, a large number of cases and long testing times are needed.[17] While expanding the detail of a scenario will increase a problem's authenticity, increasing the time needed to solve it may result in poor instrument feasibility.

Key Feature. The key-feature approach is used to assess decision-making. The approach consists of a large number of concise, clinical case descriptions that ask for essential decisions and are constructed in a written or computer-based assessment form.[17] Incorporating a large number of cases yields reliable scores. In addition, by focusing only on essential decisions, the approach demonstrates good content and construct validity.[17]

The problem with this assessment method is that preparing a good case and exam is time and labor intensive, and it is difficult to define the key decisions to be made. Another issue is that student problem-solving skills tend to be case specific, as demonstrated by low inter-case correlations.[22] Farmer and Page[23] formulated a guide for using the key-feature approach, which has demonstrated high levels of face and content validity, fair authenticity, and acceptable reliability.

An example of a key feature question that addresses decision-making skills is a case about a basketball player who comes to the athletic training room com-

plaining of ankle pain immediately after inverting his ankle. The individual evaluates the injury and observes mild to moderate swelling, point-tenderness of the soft tissue over the anterior-lateral region of the ankle, and (-) findings using the Ottawa ankle rules. Based on this information, the question could be which of the following is the best decision to make at this time: (a) recommend the patient get x-rays; (b) ice the ankle, then send the patient for x-rays; (c) compression first, then apply ice in one hour; (d) apply ice in 15 minutes; or (e) apply ice immediately?

Observational Assessment

Oral Exam. Traditional oral exams are constructed using one or more examiners who ask a student questions. The exam tends to take the form of an interview or discussion.[24] The rationale for this exam is to assess knowledge and probe for depth of knowledge. Not only can such an exam be highly threatening to students, but the exam's reliability and validity has been challenged.[24] The examiners who are taking an active part in the process can introduce bias, and the exam format lacks standardization that negatively affects reliability.[24] In an attempt to assess a student's knowledge of a particular area, an assessor may actually measure aspects of a student's personality, which also negatively impacts the validity of the test.[24] Furthermore, if the purpose of the assessment is to examine factual knowledge, a written exam can more accurately and reliably measure the knowledge. The authenticity of this instrument is challenged because of the inconsistent context between learning and assessment as well as the imposing, unnatural environment for the assessment. However, an oral assessment could provide information to the assessor if measuring aspects such as appearance, manner, alertness, confidence, and honesty.[24] An oral assessment instrument could also work well with case-based, standardized cases (6 or more) if it is used in a uniform environment with examiner consistency to ensure validity.[18] However, meeting all these criteria is very difficult, time consuming, and unrealistic.

Peer Assessment. Some published medical education literature has demonstrated that students being assessed by peers is a reliable and valid method for assessing clinical management, humanistic, and psychosocial dimensions of clinical performance.[25] Also, students have viewed feedback from peers as more meaningful than feedback from faculty in areas such as developing learning agendas and the importance of professional attitudes and behaviors.[26] Research using second year medical students discovered that six peer raters provided acceptable reliability, and the authors argued this method provided a way to measure interpersonal skills at all levels of medical training.[27]

Entry-level masters' athletic training students assessed their peers with high accuracy in scoring of psychomotor skills, but more than one student was needed to assess a peer for acceptable reliability.[28] Since peers are being exposed to the same competency learning contexts and have a similar level of competence, their judgment in assessment could positively influence this instrument's authenticity. This method of assessment may provide valuable feedback for students as a formative assessment tool for learning along with reasonable reliability and validity when using it as a summative assessment[25] for soft skills or personal attributes such as communication.

Mini-Clinical Evaluation Exercise. The mini-clinical evaluation exercise (mini-CEX) is an assessment instrument used to measure the clinical skills of medical residents.[9] The competencies assessed are communication, clinical examination, diagnosis, and management. The mini-CEX uses one faculty member assessing one resident with one patient for a 15-minute evaluation, and the assessment is done multiple times during the year, in different settings, and with various patients.[9] The mini-CEX demonstrates good reliability with the multiple encounters and encourages a growth in student competence throughout the year, which supports the validity of the method.[29] Also, criterion validity evaluations have shown a strong correlation with other assessment instruments, and its reliability becomes acceptable with a minimum of 10 encounters.[30] The authenticity of this instrument is dependent on the environment and whether the patients are real or simulated.

360º Assessment. A multisource feedback system, or a 360º assessment technique, is a survey-based assessment method that utilizes a specific questionnaire for patients and a separate survey for peers.[31] Health professional students may interact differently with patients than they do with colleagues, faculty, staff, and providers of other disciplines. For this reason, the 360º assessment instrument can be valuable since it includes assessors with expertise in multiple disciplines who provide feedback and guidance on student performances. The survey items need to be carefully struc-

tured and give consideration to the type of information the patient or peers can provide to ensure construct validity. Also, to be a valid assessment technique, the raters need to demonstrate their ability to observe the behaviors being assessed.[31] The reliability of this instrument can be affected by the number of raters. Too few raters may provide data with low reliability; however, too many raters may include individuals with a poor ability to observe the appropriate behaviors, which would negatively affect the validity of the data. This type of instrument creates an authentic context by being assessed in a workplace environment; however, a multi-discipline assessment presents an intimidating environment.

Research that used the 360º method to assess medical students demonstrated general agreement among different categories of assessors for each medical student, except for students assessing themselves.[32] The students just beginning their academic programs graded themselves higher than other assessors, while the senior students rated themselves average, or lower than average, than others assessing them. The findings of this research demonstrated good results when using the 360º instrument to assess interpersonal and communication skills, and it could be used to provide feedback and guidance for the student.[32] The unfortunate drawback to this technique is its extreme time and resource commitment, which makes it unrealistic to conduct under normal conditions.

Long Case. Long-case examinations were originally used to assess clinical skills, but they have been replaced by objective structured clinical examinations (OSCE), which are discussed in the following section.[33] In the long-case exam, students are asked to study real patients given ample, uninterrupted time to complete their evaluations. These exam sessions are often unobserved; therefore, the assessment relies on the student's presentation of the case.[33] Advocates for this test argue that the assessment of a real patient provides an authentic and valid form of assessment.[33] However, there is concern over the reliability of the long-case exam because of its small sample of content/cases and the low number of raters employed.[33] Wass et al[33] argued that with long-case exams, it would take at least three and one-half hours using 10 real patients to produce acceptable test reliability. Although this type of instrument presents an authentic assessment, the longevity of the exam challenges its feasibility.

Objective Structured Clinical Examination. The OSCE is a timed multi-station exam that assesses students on specific tasks typically using standardized patients (SP) to simulate clinical scenarios or real patients.[33] Using SPs assures consistency in assessing students by rating them from checklists that are generated by a panel of experts.[34] However, with SPs the emphasis is placed on standardization and objectivity versus sampling. For example, when assessing the clinical skills of medical students, a real patient's opinion can be very informative when assessing specific behavioral competencies that relate to human qualities. When using this technique, few cases are needed to assess basic technical skills. However, in order to achieve reliable results, a greater number of cases is required to assess interpersonal skills.[33] On average with the OSCE technique, stations are fairly short (ie, 10 minutes) so many stations can be presented. Measurement is structured by using a predefined checklist or rating scale, and the sources of variance (eg, bias) are averaged out by using many different stations (10–15), examiners, and simulated patients.[34]

Due to inter-patient variability and issues of inter-rater reliability, students need to be subjected to multiple exams and assessed using standardized rating forms for reliable patterns to emerge as they are observed.[33] With isolated competencies, the OSCE is limited in its ability to measure what the student would do in real-life situations that require the integration of different skills, which challenges authenticity and validity.[34]

In assessor scoring, using a checklist is thought to strengthen inter-rater reliability; however, studies have shown that inter-rater reliability is a relatively small source of error compared to inter-case variability.[17] The use of global rating scales, or holistic judgments, by assessors did not inhibit reliability when employing direct observation and repeated measurements.[35] The advantage of a global rating scale is its ability to allow assessors to include more qualitative observations such as the efficiency and ease with which skills are performed by a student.[17] In fact, when using expert examiners, a global rating scale demonstrates stronger reliability, construct validity, and concurrent validity than the checklist in assessing technical clinical skills.[35] The decision to use checklists or global ratings could depend on the tasks being scored. For example, checklists may be more appropriate for scoring stations that measure specific, practical, and technical skills. Global rating scales could be used for sta-

tions testing communication skills or when assessing diagnostic tasks that utilize various routes to achieve the same outcome.[36]

Control, Explaining, Listening, Influencing Instrument. The Control, Explaining, Listening, Influencing (CELI) instrument is used to assess a physician's ability to educate a patient regarding their condition.[37] The instrument is composed of four sub-competencies: rapport, explaining, active listening, and influencing. The CELI instrument demonstrates adequate reliability, strong validity, and adequate authenticity as an assessment tool for physician competency in patient education.[37] The 5th edition of athletic training competencies[5] does include patient education as a skill under the psychosocial strategies and referral content area, which could warrant the use of this instrument.

Professional Assessment. To assess performance or the highest level of competence, medical education researchers have suggested moving away from simulated situations to the examination of real practice settings using global ratings and expert judgments of clinical work samplings and practice video recordings.[17] With observations made in real practice, the examinations are more authentic and greater attention is given to adequate sampling than simulations.[17] However, a large number of judges need to make many observations in order to produce reliable results.[17]

Another assessment technique at this level is called the incognito SP-based examination. For this technique, participants are trained to portray patients with certain signs and symptoms for specific injuries or illnesses.[17] The SP visits a student practitioner who, without realizing the patient is a simulator, must diagnose the injury or illness and develop a plan of care for the SP. After the completion of the consultation, the SP completes a pre-defined checklist or rating scale to score the student's performance. The advantage of this method is that the outcome is not influenced by the measurement because the student is unaware they are being tested. However, this method is expensive and labor intensive, which is partially due to the fact that the SP needs to be trained well to produce sufficient authenticity.

To assess a student's integrated clinical proficiencies, an evaluator may use either SPs, actual patients, or simulations.[38] Published research suggests that assessing student performance with computer-enhanced mannequins and virtual reality simulators can achieve a high degree of reliability and evidence of validity.[39] For example, researchers have reported strong validity and reliability ratings when assessing resident performance during microscopic examination using virtual microscopy with computer-assisted case simulators.[40] The assessment of more abstract skills, such as the demonstration of professional behaviors, may be accomplished best by using a combination of tools (eg, using portfolios along with mini-CEXs).[41] Such assessments may be useful to athletic training educators. For example, professionalism is a foundational behavior of athletic training professional practice that must be assessed throughout the educational program.

DISCUSSION

The purpose of summative assessments of competencies is to ensure that students have reached an acceptable, pre-determined level of competence in all domains when they complete their educational experiences.[42] Assessing all levels of competence, however, requires a combination of written exams and observations. Creating quality yet feasible instruments that verify student expertise with their competencies is very challenging.

To ensure the validity of an assessment, some general guidelines for the summative assessment of competencies need to be reviewed. One suggestion noted in the literature is to test items of knowledge and those measuring clinical skills equally since this positively influences assessment validity.[16] Another suggestion to ensure validity is that the higher the level of competence assessed, the more clinically authentic the assessment needs to be.[8] Also when examining higher levels of competence, assessments of professional judgment using global ratings may be more valid than detailed checklists giving objective scores.[35]

The greatest threat to the reliability of both written and clinical simulation examinations does not appear to be from their structure, but rather errors in sampling.[17] Sampling, in this case, means the items are selected from a range of possible knowledge or skill questions. Errors in sampling may occur when the number of items are limited or too focused on a single element.[8] Inter-rater reliability may be improved through multiple examiners across different cases. Although a larger number of assessors with varied strengths may improve reliability, the assessors should share the same assessment criteria or rating scale to be sure they are

working from the same standards. Inter-case reliability refers to the consistency of a student's performance across clinical cases or stations.[33] There is a need for many stations and sufficient testing time for adequate inter-case reliability.[33] Since competence is highly dependent on context (ie, situation or task) and content (ie, knowledge, skill, or attitude), an effective assessment method must gather a large sample of observations (eg, items, patients, stations or essays) across the content area, which are then tested with a careful sampling of examiners.[8] When combining several assessment methods, overall reliability becomes acceptable.[16]

This review provides examples of several promising competency assessment instruments when judged by quality criteria. When considering knowledge competencies, certain techniques come to the forefront. Knowledge tests, such as multiple-choice exams, are very efficient in handling large numbers of students and can cover a wide range of domains while maintaining strong reliability.[43] If the goal of the assessment is to gain an impression of how well students understand specific concepts, short answer questions may be useful but they may also be less reliable than multiple choice exams. A very promising alternative with the assessment of knowledge competencies is the extended matching question instrument. With proper construction and content, this assessment method can potentially provide acceptable quality with all assessment criteria. In a written examination, the type of instrument does not determine what level of competence is being tested; instead, the content of the question determines the level of competency assessed.[11] For example, multiple choice questions can be constructed to solve problems versus just recalling facts.

When assessing specific psychomotor skills, the OSCE provides adequate reliability and validity when incorporating an adequate number of stations, expert judges, and global rating scales. The mini-CEX instrument can be utilized for the assessment of integrated competencies, such as with clinical evaluation and management skills. Reliability improves with multiple encounters; however, validity suffers with simulated patients. Decision-making skills can be adequately assessed using key feature questions on essential decisions and a relatively large number of cases. Multiple peer assessments may provide accurate assessments and provide collaborative learning with colleagues on specific types of competencies. Another instrument that has shown promise in the assessment of soft skills is the CELI instrument, which is presently designed for assessing a physician's skill in educating patients. It has shown strong reliability and validity in this specific area, and with the adjustment of questions, could arguably be useful in the assessment of all patient-practitioner communication skills.

Integrated clinical proficiencies can be assessed with the help of virtual reality simulators, mannequins, SPs, or real patients. However, feasibility issues such as time constraints and resources affect the ability to use these techniques. Consequently, a combination of assessment methods may be needed. For example, the use of portfolios with clear guidelines that provide for clinical reflection in combination with mini-CEXs could provide adequate assessment quality. Optimally, assessment techniques incorporating authentic work settings with adequate sampling across different contexts and various assessors with video assessment and incognito patients could be used to examine students at the performance level.

Although not of direct concern, it is critical to realize that the type of assessment instrument persuades students to engage in specific learning strategies. Student learning is driven by the content, format, and programming of assessment.[44] Given the influence of testing, assessors need to decide how they want students to learn and use the appropriate exam to steer their learning in the desired direction. For example, a particular test format may encourage students to memorize information to maximize their test score; however, this could inhibit the student's understanding of the material. Various assessment techniques can lead to different learning behaviors, which reinforces the view that how the assessment takes place, and in what context it occurs, affects student learning.[8]

In summary, an ongoing evaluation and adjustment in the assessment of competencies is imperative.[8] The literature has demonstrated that there is no one-size-fits-all assessment method for competencies. Additional research is needed that investigates sources of variance in areas such as assessment techniques, assessors, competency domains, tasks, patients, contexts, timeframes, authenticity, educational consequences, and competency level.[10] When comprising assessment techniques for competencies in a competency-based health profession education program, there are a number of things to consider in the assessment planning process. Schuwirth and van der Vleuten[11] recommend assessors construct a document

to aid the assessment planning process, and this document should include the purpose of the assessment, the goals, the techniques to be used and why they are used, how sampling will take place, the quality control mechanisms to use, and how results from measurements are to be compiled to examine assessment efficiency and effectiveness. The content of the test rather than its method primarily determines what is being measured.[44] For example, multiple choice questions could test problem-solving competence while an oral exam could test factual knowledge. Also, the assessment plan should be a carefully composed combination of techniques that provide an overall judgment of competence.[8]

References

1. Bogo M, Regehr C, Hughes J, Power R, Globerman J. Evaluating a measure of student field performance in direct service: testing reliability and validity of explicit criteria. *J Soc Work Educ*. 2002;38(3):385-401.
2. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387-396.
3. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002; 287(2):226-235.
4. Miller GE. The assessment of clinical skills/competencies/performance. *Acad Med*. 1990;65(9): S63-S67.
5. National Athletic Trainers' Association. Athletic Training Education Competencies. 5th ed. Dallas, Texas: NATA; 2011.
6. Thomas A, Saroyan A, Dauphinee DW. Evidence-based practice: a review of theoretical assumptions and effectiveness of teaching and assessment interventions in health professions. *Adv Health Sci Educ*. 2011;16(2): 253-276.
7. Van der Vleuten CP. Validity of final examinations in undergraduate medical training. *BMJ*. 2000;321(7270):1217-1219.
8. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3)309-317.
9. Lee AG, Beaver HA, Greenlee E, et al. Teaching and assessing systems-based competency in ophthalmology residency training pro grams. *Surv Ophthalmol*. 2007;52(6):680-689.
10. Gulikers JTM, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. *Educ Technol Res Dev*. 2004;52(3):67-86.
11. Schuwirth LW, Van der Vleuten CP. Changing education, changing assessment, changing research? *Med Educ*. 2004;38(8):805-812.
12. Sidhu RS, Vikis E, Cheifetz R, Phang T. Self-assessment during a two-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? *Am J Surg*. 2006;191(5):677-681.
13. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence. *JAMA*. 2006;296(9):1094-1102.
14. Lew MDN, Alwis WAM, Schmidt HG. Accuracy of students' self-assessment and their beliefs about its utility. *Assess Eval High Educ*. 2010;35(2):135-156.
15. Zick A, Granieri M, Makoul G. First-year medical students' assessment of their own communication skills: a video-based, open-ended approach. *Patient Educ Couns*. 2007;68(2):161-166.
16. Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Med Educ*. 2001;35(4):326-330.
17. Schuwirth LW, Van der Vleuten CP. The use of clinical simulations in assessment. *Med Educ*. 2003;37(1), 65-71.
18. Leigh IW, Bebeau MJ, Nelson PD, et al. Competency assessment models. *Prof Psychol Res Pract*. 2007;38(5):463-473.
19. Schuwirth LW, Van der Vleuten CP. ABC of learning and teaching in medicine: written assessment. *BMJ*. 2003;326(7390):643-645.
20. McMullan M, Endacott R, Gray MA, et al. Portfolios and assessment of competence: a review of the literature. *J Adv Nurs*. 2003;41(3): 283-294.
21. Driessen E, Tartwijk JV, Van der Vleuten C, Wass, V. Portfolios in medical education: why do they meet with mixed success? A systematic review. *Med Educ*. 2007;41(12):1224-1233.
22. Page G, Bordage G. The medical council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Acad Med*. 1995;70(2):104-110.

23. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ*. 2005;39(12):1188-1194.
24. Davis MH, Karunathilake I. The place of the oral examination in today's assessment system. *Med Teach*. 2005;27(4):294-297.
25. Ramsey PG, Carlene JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. *Acad Med*. 1996;71(4):364-370.
26. Asch E, Saltzberg D, Kaiser S. Reinforcement of self-directed learning and the development of professional attitudes through peer- and self-assessment. *Acad Med*. 1998;73(5):575-576.
27. Dannefer EF, Henson LC, Bierer SB, et al. Peer assessment of professional competence. *Med Educ*. 2005;39(7):713-722.
28. Marty MC, Henning JM, Willse JT. Accuracy and reliability of peer assessment of athletic training psychomotor laboratory skills. *J Athl Train*. 2010; 45(6):609-614.
29. Norcini JJ, Blank LL, Duffy D, Fortna GS. The mini-cex: a method for assessing clinical skills. *Ann Intern Med*. 2003;138(6):476-483.
30. Pelgrim EAM, Kramer AWM, Mokkink HGA, Van den Elsen L, Grol RPTM, Van der Vleuten CPM. In-training assessment using direct observation of single-patient encounters: a literature review. *Adv Health Sci Educ*. 2011;16(1):131-142.
31. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof*. 2003;23(1):4-12.
32. Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate resident's competency in interpersonal and communication skills. *Acad Med*. 2004;79(5):458-463.
33. Wass V, Van der Vleuten C, Shatzer J, Jones, R. Assessment of clinical competence. *Lancet*. 2001;9260:945-949.
34. Barman A. Critiques on the objective structured clinical examination. *Ann Acad Med*. 2005;34(8):478-482.
35. Regehr G, MacRae H, Reznick RK, Szalay, D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an osce-format examination. *Acad Med*. 1998;73(9):993-997.
36. Wilkinson TJ, Fontaine S. Patient's global ratings of student competence. Unreliable contamination or gold standard? *Med Educ*. 2002;36(12):1117-1121.
37. Wouda JC, Zandbelt LC, Smets EMA, Van de Wiel HBM. Assessment of physician competency in patient education: reliability and validity of a model-based instrument. *Patient Educ Couns*. 2011;85(1):92-98.
38. Walker SE, Weidner TG, Armstrong KJ. Eval uation of athletic training students' clinical proficiencies. *J Athl Train*. 2008;43(4):386-395.
39. Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. *J Gen Intern Med*. 2007;23(1):46-49.
40. Bruch LA, Young BR, Kreiter CD, Haugen TH, Leaven TC, Dee FR. Competency assessment of residents in surgical pathology using virtual microscopy. *Hum Pathol*. 2009;40(8):1122-1128.
41. Van Mook WNKA, Gorter SL, O'Sullivan H, Wass V, Schuwirth LW, Van der Vleuten C. Approaches to professional behavior assessment: tools in the professionalism toolbox. *Eur J Intern Med*. 2009;20(8):153-157.
42. Prescott LE, Norcini JJ, McKinlay P, Rennie JS. Facing the challenges of competency-based assessment of postgraduate dental training: longitudinal evaluation of performance (LEP). *Med Educ*. 2002;36(1):92-97.
43. Ram P, Van der Vleuten C, Rethans J, Schouten B, Hobma S, Grol R. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. *Med Educ*. 1999;33(3):197-203.
44. Van der Vleuten CP, Scherpbier AJ, Dolmans DH, Schuwirth LW, Verwijnen GM, Wolfhagen HA. Clerkship assessment assessed. *Med Teach*. 2000;22(6):592-600.