Designing and selecting assessment instruments: *focusing on competencies*

Stanley J. Hamstra, PhD

Dr. Hamstra is acting assistant dean, Academy for Innovation in Medical Education, research director, University of Ottawa Skills and Simulation Centre, and associate professor, departments of Medicine, Surgery and Anesthesiology, Faculty of Medicine, University of Ottawa.

Objectives

After reading this chapter you should be able to:

- describe different criteria used to select assessment instruments
- selectively identify assessment instruments for summative and formative purposes
- use a standard approach to assist in the drafting of a new assessment instrument

Case scenario

You are the residency program director for a large subspecialty in a mid-sized academic health centre. Your program was cited during the recent accreditation review by the Royal College of Physicians and Surgeons of Canada. The reviewers' main concern was that your program lacks comprehensive assessment tools. The only CanMEDs Roles that are being assessed in a valid and reliable way are Medical Expert and Scholar. It is your job to develop and put into place assessment tools to comprehensively measure competence in all the CanMEDs Roles. You decide to form a working group of members of your residency training committee to survey other schools to see what tools they are using and to adapt these to your local context.

A few weeks later, the working group reports that there are some instruments that have evidence for validity and reliability, but these cannot be readily applied to your program. You realize that you will have to create your own. You decide to start by developing a tool to assess the Communicator Role, and you hope that you will be able to "copy and paste" this tool for the other competencies. However, as you are reading the literature in medical education, you realize this task will be more complicated than you had imagined. Your chair is interested in supporting your efforts but has limited resources to offer and sees you as the departmental expert in the education field.

Among the issues that you still need to address are the following:

- how to demonstrate validity of the instrument,
- feasibility issues, including who exactly will do the rating of performance and in what context,
- faculty development (you realize as you work through the steps of developing the instruments that your colleagues will require some training to use this instrument effectively) and

» reporting issues (Where will the assessment data be stored? How exactly will the data be reported? To whom and when will the data be reported?).

Although some of these questions appear to be logistical in nature, your reading of the recent literature suggests something you've never heard from your colleagues: all of these issues are important contributors to the instrument's validity and hence can be critical in determining whether or not it provides accurate scores for each trainee. The risk of providing invalid scores is reflected in the rising number of challenges that are being brought to the Postgraduate Medical Education Appeals Committee. You want to avoid these headaches and build an assessment toolbox that meets the highest standards for validity and reliability.

Background and context

The basics

Assessment is a very important part of medical education. It is also a complex topic. In the world of education, the terms ***assessment*** and ***evaluation*** mean different things. ***Assessment*** refers to judging an individual learner's progress, whereas ***evaluation*** refers to judging the effectiveness of a program or curriculum.

Almost all assessment instruments are made up of a series of individual items. The quality of the assessment instrument depends on the quality of the individual items that comprise it, and the process of developing (or editing) items should not be taken lightly. Two common forms of items are statements, such as "the candidate maintained appropriate eye contact while communicating sensitive information" anchored by a five-point Likert scale, or simple checklist items, such as "keeps edges of the wound everted while closing the skin — done/not done." There is an extensive literature on the process of writing items to test for knowledge or for application of knowledge to a clinical problem. One of the best guides to this process is a manual published by the National Board of Medical Examiners.[1] Item writing is a learned skill, and with practice it can be done efficiently and effectively.

For knowledge tests, one of the key concerns is to ensure that the test takers do not perform well because they are using highly developed test-taking strategies. Although this is less of an issue for tests of performance or skill, it is helpful to realize that any test has a powerful motivational influence on the student, and it is human nature to use any means possible to perform well in a high-stakes test. Given this tendency, it is your responsibility to ensure that the test measures what you intend it to measure, not irrelevant test-taking skills. This concept forms the core of validity.

What construct are you trying to measure?

Another term that is of central importance in this field is ***construct***. This term is widely used in some of the social and behavioural sciences, such as psychology. When you are developing an assessment instrument, you start with the construct of interest. That is, what are you trying to measure? In medicine, some constructs of interest are communication skills, technical skills and professionalism. For our purposes, the construct of interest is something that can be measured and that varies between individuals based on experience or training. (This is often called "construct-relevant variance.")

When you decide to develop or revise an assessment instrument for a particular construct, it is important that you first gather stakeholders to discuss the definition and the boundaries of the construct at length and that some consensus be achieved before you move on. If members of your stakeholder groups are not in agreement as to the definition of the construct, this will probably cause uncertainty, and possibly criticism, lack of acceptance and lack of uptake of your instrument in the future. For example, in a palliative care environment, one aspect of the construct "communication skill" could be defined as the ability to effectively and compassionately convey important end-of-life information to a patient and their family. Note that in this example, it would be useful to engage content experts to discuss what exactly is meant by the terms "effectively" and "compassionately," according to some behavioural criteria. If your expert panel does not agree on what is meant by one or both of the terms, it will be difficult to move on productively.

In an ideal assessment instrument, all of the variance in test results will be construct-relevant variance. In other words, the test will only measure qualities related to the construct of interest. As mentioned above, test performance can be influenced also by irrelevant constructs, such as test-taking skills, or other factors such as age, gender or genetic makeup, producing construct-irrelevant variance.

Are you interested in formative or summative assessment?

After you achieve a common understanding among your stakeholders on the construct of interest for assessment, you will need to determine the purpose of the assessment:

- Formative assessment is designed to give feedback to assist the learner to improve.
- Summative assessment entails pass/fail decisions to determine whether a minimum criterion has been achieved and the individual is ready for a next step.

In instrument design, this distinction matters because it determines where you will focus your energy with the individual items that make up the assessment instrument. For example, do you want to be able to simply tell the learner that they have passed a test (as with certification examinations)? Or are you genuinely interested in giving them specific and comprehensive feedback on a variety of domains that they can use to improve? There's a funny paradox here, because as administrators, we always say that we want to give learners feedback for improvement, but we typically don't follow through: we find the task of creating, managing and administering an assessment so onerous and time consuming that we usually don't bother giving comprehensive feedback after an event. On the rare occasions when learners actually undergo formative assessment, they typically treat it as summative – they tend to believe that someone whose opinion they value will be looking at the results and judging them, no matter what they are told. In terms of professional risk assessment, learners are safer to assume that someone will be looking at their results than not. In fact some authors argue that all formative assessment is also summative, in that formative assessment makes use of judgmental terms.[2]

To focus matters slightly more, there are also two special cases of summative assessment that we often use:

- We may discriminate among the highest achievers for awards or reference letters (this can also be thought of as a type of formative assessment, in the sense that it provides the individual with more nuanced feedback than a simple pass/fail decision).
- We may discriminate among the lowest achievers to help in tailoring specific remedial programs (this goes beyond the regular form of summative assessment in that it provides specific feedback for use beyond the current rotation or curriculum).

All of this is important because when we are designing a summative assessment instrument, we need to know whether we are concerned with discriminating between individuals who pass or fail (a relatively simple task), or whether we want to be able to rank the individuals in terms of their performance on the test. Commonly, we start off with the assertion that we are only interested in the pass/fail decision but end up trying to use the instrument for making finer distinctions.

The importance of variance

Variance is your friend. Or to put it more accurately, construct-relevant variance is your friend. Quite simply, when you are assessing performance in any domain, you need to see variation in scores. This is a simple but critical assertion. If there was no variance in the scores produced by an assessment exercise, all individuals would be deemed to be identical in the domain of interest and there would therefore be no need to have used an assessment instrument. Stated another way, the reason one uses an assessment instrument is to discriminate between individuals with differing levels of skills or performance. It is important to determine what level or aspect of the variance between individuals is of most interest to you, because this will help you to

determine how to spend your energy in writing items and developing the overall assessment instrument. For example, if the purpose of the instrument is solely to inform pass/fail decisions, then most of the focus in developing the individual test items should be on distinguishing between borderline performances (i.e., those just above and just below the criterion reference). In this case, there is no need to spend energy developing items that discriminate among the few individuals at the top of the class, because we are only concerned with whether each student has passed or not. Similarly, we would not concern ourselves with fine distinctions between individuals at the very bottom of the class, as long as we were confident that none of them met the standard. If the purpose of our instrument is for formative assessment or discriminating among the highest achievers or among the lowest achievers, then our ideal instrument would discriminate between individuals at every level of performance.

Note that many measures of competence used in resident assessment, such as end-of-rotation evaluations, produce results with relatively little variance.[3–5] Fortunately, many standardized tests of aptitude and skill (e.g., medical knowledge, psychomotor ability, visuo-spatial ability) show good variation across a wide spectrum of performance and have demonstrated evidence of reliability and validity.

Determining the desired pattern of variance among individuals: hitting the target construct

Ideally, the performance scores among your learners should vary in a meaningful way. If you have selected a construct that can be defined with clear consensus among your stakeholders, you should be able to determine the variables to which this construct relates without too much difficulty. Let us consider again the construct "communication skill in a palliative care environment." This construct might be expected to relate to level of maturity, amount of experience in that clinical setting, or level of training. These variables would all contribute to construct-relevant variance. If you see that cultural and language issues are affecting the learners' performance on your assessment instrument, then the assessment results are exhibiting construct-irrelevant variance (i.e., variance, unrelated to the construct of interest). In this case, you would be advised to try to modify your assessment instrument to more directly assess your target construct. This is essentially a question of a full treatment of validity which is beyond the scope of this chapter but is available elsewhere.[6–8]

In addition to creating conditions for valid assessments — in other words, to obtain the desired pattern of variance in its scores — you also need to examine your instrument's reliability (the reproductibility of the test scores) and feasibility. A test can be reliable and valid but fail because of feasibility constraints. Issues related to data collection, logistics of testing, subject fatigue, motivation, rater training, cost and time all are feasibility concerns. All of these feasibility issues can easily be addressed during pilot testing.

Together, validity, reliability and feasibility determine the quality of your assessment instrument. You can and should measure these elements while you are developing your assessment instrument and you should continuously monitor them once you have implemented it. These elements can also be used as criteria in the selection of pre-existing assessment instruments.

Literature scan

Numerous papers have been written on the topic of assessment in medical education. Early papers focused on larger descriptions of frameworks for assessment and provide excellent introductions to the nature and uses of assessment instruments in medical education.[9,10] Cook and Beckman[6] built on this and described the more recent framework for understanding validity.[7,8] In the last few years, the field has turned its attention to competency-based medical education and assessment of the roles and competencies outlined by the CanMEDs framework and the Accreditation Council for Graduate Medical Education (ACGME). This shift has led to many reviews and survey papers in specialty journals designed to help program directors understand the principles and practices of assessment in a variety

of specialties, including Anesthesia,[11] Surgery,[12–14] Emergency Medicine[15] and Psychiatry.[16] Finally, Cook and colleagues recently reviewed assessment in simulation-based education.[17]

Individual assessment has long been a subject of study in psychology and education. Medical education has embraced developments in these disciplines to refine specific procedures to select candidates for admission to medical schools and residency programs and to create examinations for certification and licensing. In the latter half of the 20th century, advances in the field of psychometric assessment were implemented more widely in medical education, for the purposes of curriculum evaluation, summative assessment of learners at the end of courses and rotations, and formative assessment during training. With the introduction of the CanMEDs and ACGME frameworks and the move toward maintenance of certification and competency-based education, the need for an understanding of the basic principles of testing (psychometrics) has increased in our field. While the competencies associated with the Medical Expert Role have a long history of reliable and valid assessments, there needs to be more research on how to assess some of the more intrinsic competencies, such as communication skills, professionalism and systems-based practice.

Tips and pitfalls

» **Don't try to do this by yourself.** Many novices to psychometric assessment develop instruments on their own. An assessment tool is not a simple questionnaire. You will need help from content experts and possibly psychometricians (yes, psychometrics is an occupation).

» **Give yourself enough time to develop your instrument properly.** You should typically budget about a year to produce a good-quality instrument. You will need this time to review the content of the instrument with your expert panel, write and revise the items, pilot test the instrument and revise the final draft. If you are revising an existing instrument rather than creating a new one, you might be able to shorten the time to six months.

» **Familiarize yourself with the literature.** Too often, program directors are so busy managing the operational aspects of their program that they spend little (or no) time reading up on best practices in medical education. There are many journals dedicated to this field, and there is a thriving sub-field on psychometric assessment in medical education. An assessment instrument relevant to your needs may have been published in the literature. You don't want to "reinvent the wheel."

» **Consult Textbox 12.1**, which is based on a recently published checklist that you can follow when designing your own assessment instrument.[18]

» **Don't start with too narrow a definition of your construct.** It is harder to open up the discussion once you are in the process of developing the instrument than it is to narrow it down. If you start with too narrow a definition, some members of your expert panel will feel like they aren't being heard and their annoyance will colour the rest of the deliberations.

- **Develop test items that reflect the competencies you choose to assess** and carefully consider the levels of performance of your target population. (Consider this question: What does borderline performance look like for this competency?)
- **Review the nationally recommended objectives for your (sub)specialty** (the documents from your specialty committee containing the objectives of training requirements and specialty training requirements) to find relevant language for creating and refining items.
- **Use a representative sample when developing the instrument.** If you only consult content experts or residents from your own institution, you will probably miss some of the variables that are important to your construct. Ensure you get others from outside your geographic area to review the content. An instrument that is highly specific to your location will have limitations that will be obvious to the accreditation review committee.
- **Don't use the instrument for high-stakes assessment before you have done the pilot test.** Pilot testing often reveals a need for critical revisions for feasibility: the content may be solid, but the instrument may prove to be impractical because it is too costly or time-consuming to implement. A pilot test involving trainees with a similar background to that of the target group should allow for an adequate assessment of feasibility.
- **Each time you administer your assessment instrument, collect data for continuous quality assurance** on student performance, individual item statistics, reliability and validity.
- **Don't be afraid to revise the items on your assessment instrument** if the results they generate are not telling you anything useful.

Textbox 12.1: Seven-step checklist for developing a good assessment instrument

1. **Determine the purpose of your assessment.**
   - Will the instrument be used for formative or summative (standard setting/criteria) assessment or research?
   - Do you want to assess knowledge, skills or attitudes (e.g., performance, teamwork, anxiety)?
2. **Identify the main construct of interest and stakeholders to help establish content validity.**
3. **Review the construct with content experts using a consensus method such as focus groups.**
   - Obtain a representative sample from different institutions and disciplines.
   - Work toward thematic saturation and address political issues.
   - Set preliminary standards: What does perfect/borderline performance look like?
4. **Develop and write the items, drawing on related existing tests if applicable.**
5. **If necessary, train the raters (and assess inter-rater reliability).**
6. **Pilot test the instrument (with a representative sample) for validity.**
   - Check the feasibility of the instrument (length, clarity, cost).
   - If necessary, go back to step 4 (modify the items) and then pilot test again.
7. **Implement the modified test and measure its reliability and validity with a larger sample.**
   - Assess construct validity.

Final note: We can never achieve perfect validity, so consider this to be an ongoing process whereby you are constantly checking performance statistics for reliability and validity.

Adapted from Hamstra SJ. Keynote address: the focus on competencies and individual learner assessment as emerging themes in medical education research. *Acad Emerg Med.* 2012;19(12):1336–1343.

Case resolution

After reviewing the checklist in Textbox 12.1, you feel empowered to move forward and create a new assessment instrument or adapt an existing one for your purposes. You enlist the help of a psychometrician from your faculty of education or department of psychology. You decide your instrument should be used for summative assessment of residents in your program, to inform pass/fail and remediation decisions. While assembling your expert panel, you begin work on achieving consensus on the definition of the construct of interest. You decide that you will focus on communication skills. You enlist the help of the research librarian at your institution and find that very little has been published on the assessment of communication skills in your field. You work with your expert panel and the psychometrician to write and revise items that will allow you to assess your specialty's objectives. You then decide on a grading method on the basis of how your expert panel views progressive levels of expertise. Following a pilot test, you revise the items and begin regular administration of the assessment instrument. You gain confidence that your new assessment instrument will meet national standards for best practices in competency assessment in your field. You submit the new instrument and your initial results to the accreditation team and they grant you full approval.

Take-home messages

- Review the purpose of your proposed assessment instrument: Will you use it for a summative or formative purpose?
- Make liberal use of content experts in developing items for your assessment instrument. It is important that the definition and the boundaries of the construct be discussed at length and that some degree of consensus be achieved before moving on.
- Put some effort into pilot testing. You will learn a lot about the construct and the particular CanMEDS Role you are trying to assess from doing this.
- Keep an eye on reliability, validity and feasibility as you develop and work with your assessment instrument. Collect data for continuous quality improvement.

References

[1] Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences.* 3rd ed. Philadelphia (PA): National Board of Medical Examiners; 2002.

[2] Brown SA, Glasner A. *Assessment matters in higher education: choosing and using diverse approaches.* Buckingham (UK) and Philadelphia (PA): Society for Research into Higher Education and Open University Press; 1999.

[3] Dauphinee WD. Assessing clinical performance: Where do we stand and what might we expect? *JAMA.* 1995;274:741–743.

[4] Maxim BR, Dielman TE. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. *Med Educ.* 1987;21:130–137.

[5] Streiner DL. Global rating scales. In: Neufeld VR, Norman GR, eds. Assessing clinical competence. New York (NY): Springer; 1985. p. 119–141.

[6] Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119(2):166.e7–16.

[7] American Educational Research Association, American Psychological Association and National Council of Measurement in Education. *Standards for educational and psychological testing*. Washington (DC): American Educational Research Association; 1999.

[8] Messick S. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995;50:741–749.

[9] Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001;357(9260):945–949.

[10] Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA.* 2002;287:226–235.

[11] Boulet JR, Murray D. Review article: assessment in anesthesiology education. *Can J Anaesth.* 2012;59(2):182–192.

[12] Hamstra SJ, Dubrowski A. Effective training and assessment of surgical skills, and the correlates of performance. *Surg Innov.* 2005;12(1):71–77.

[13] Sidhu RS, Grober ED, Musselman LJ, Reznick RK. Assessing competency in surgery: Where to begin? *Surgery.* 2004;135:6–20.

[14] Fried GM, Feldman LS. Objective assessment of technical performance. *World J Surg.* 2008;32(2):156–160.

[15] Farrell SE. Evaluation of student performance: clinical and professional performance. *Acad Emerg Med.* 2005;12(4):302e6–10.

[16] Manring J, Beitman BD, Dewan MJ. Evaluating competence in psychotherapy. *Acad Psychiatry.* 2003;27(3):136–144.

[17] Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med.* 2013;88(6):872–883.

[18] Hamstra SJ. Keynote address: the focus on competencies and individual learner assessment as emerging themes in medical education research. *Acad Emerg Med.* 2012;19(12):1336–1343.

Other resources

Many medical schools in Canada now have medical education research units, typically with an expert in assessment (i.e., a psychometrician). Textbox 12.1 provides a checklist for developing a good assessment instrument.