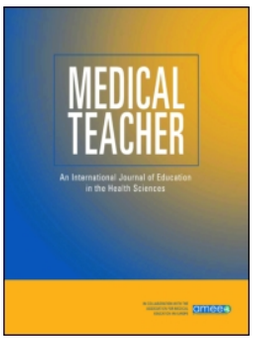

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

2018 Consensus framework for good assessment

John Norcini, M. Brownell Anderson, Valdes Bollela, Vanessa Burch, Manuel João Costa, Robbert Duvivier, Richard Hays, Maria Felisa Palacios Mackay, Trudie Roberts & David Swanson

To cite this article: John Norcini, M. Brownell Anderson, Valdes Bollela, Vanessa Burch, Manuel João Costa, Robbert Duvivier, Richard Hays, Maria Felisa Palacios Mackay, Trudie Roberts & David Swanson (2018): 2018 Consensus framework for good assessment, Medical Teacher, DOI: 10.1080/0142159X.2018.1500016

To link to this article: https://doi.org/10.1080/0142159X.2018.1500016

Published online: 09 Oct 2018.

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2018.1500016

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

2018 Consensus framework for good assessment

John Norcini[a], M. Brownell Anderson[b], Valdes Bollela[c], Vanessa Burch[d], Manuel João Costa[e], Robbert Duvivier[f], Richard Hays[g], Maria Felisa Palacios Mackay[h], Trudie Roberts[i] and David Swanson[j]

[a]FAIMER, Philadelphia PA, USA; [b]NBME, Philadelphia PA, USA; [c]School of Medicine of Ribeirão Preto, Universidade Cidade de Sao Paulo, Ribeirão Preto, Brazil; [d]Groote Schuur Hospital, University of Cape Town and Groote Schuur, Cape Town, South Africa; [e]School of Medicine, University of Minho, Braga, Portugal; [f]Parnassia Psychiatric Institute, Maastricht University, Hague, The Netherlands; [g]Rural Clinical School, University of Tasmania, Burnie, Australia; [h]Cumming School of Medicine, University of Calgary, Alberta, Canada; [i]Medical Education Unit, University of Leeds, Leeds, UK; [j]ABMS, Chicago, IL, USA

ABSTRACT

Introduction: In 2010, the Ottawa Conference produced a set of consensus criteria for good assessment. These were well received and since then the working group monitored their use. As part of the 2010 report, it was recommended that consideration be given in the future to preparing similar criteria for systems of assessment. Recent developments in the field suggest that it would be timely to undertake that task and so the working group was reconvened, with changes in membership to reflect broad global representation.

Methods: Consideration was given to whether the initially proposed criteria continued to be appropriate for single assessments and the group believed that they were. Consequently, we reiterate the criteria that apply to individual assessments and duplicate relevant portions of the 2010 report.

Results and discussion: This paper also presents a new set of criteria that apply to systems of assessment and, recognizing the challenges of implementation, offers several issues for further consideration. Among these issues are the increasing diversity of candidates and programs, the importance of legal defensibility in high stakes assessments, globalization and the interest in portable recognition of medical training, and the interest among employers and patients in how medical education is delivered and how progression decisions are made.

Background

In 2010, the Ottawa Conference produced a set of consensus criteria for good assessment (Norcini et al. 2010). These were well received and in the intervening years, the original working group has continued to monitor their use. As part of the 2010 report, it was recommended that consideration be given in the future to prepare similar criteria for systems of assessment. Recent developments in the field suggest that it would be timely to undertake that task and so the working group was reconvened, with changes in membership to reflect broad global representation.

As a first step, consideration was given to whether the criteria that were initially proposed continued to be appropriate for single assessments (Table 1). The group believed that they were. As a second step, there was discussion about whether the same set of criteria could be applied to both individual assessments and systems of assessment. The group was initially divided on this issue but eventually reached consensus that a separate set of criteria should be developed for the systems of assessment.

This paper reiterates the criteria that apply to individual assessments. With minor exceptions, it duplicates the relevant portions of the 2010 consensus report and an acknowledgement of purpose and stakeholders on the application of the standards. This paper also presents a new set of criteria that apply to systems of assessment and, recognizing the challenges of implementation, offers several issues for further consideration. Among these issues are the increasing diversity of candidates and programs, the importance of legal defensibility in high stakes assessments, globalization and the interest in portable recognition of medical training, and the interest among employers and patients in how medical education is delivered and how progression decisions are made.

To generate the criteria for systems of assessment, the group began by conducting a search of the literature for the purpose of identifying relevant work. We identified five sources that yielded a list of 24 criteria (Office of Academic Planning and Assessment 2001; Clarke 2012; National Research Council 2014; van der Vleuten et al. 2015; St Olaf College 2018). Through discussion we settled on seven criteria drawn from the twenty four, some with modifications. We then compared our criteria to the much more detailed guidelines proposed by Dijkstra and colleagues to ensure they were broadly consistent (Dijkstra et al. 2012).

When these ideas were presented as part of a workshop at the 2018 Ottawa Conference, there was a strong sense that the use of the word "criteria" was not optimal since it implied the development of standards against which assessments could be judged. Instead, there was general agreement that the word "framework" more precisely captured our desire to create a structure that might be useful in the development and review of individual assessments and systems of assessment. That change is reflected in the remainder of the document.

Given these shifts in priorities and purpose, the various elements of a framework do not apply universally and equally to all the assessments. The context and purpose-priorities of assessment heavily influence the importance of those elements. For example, a good summative

CONTACT John Norcini jnorcini@faimer.org FAIMER, 3624 Market Street, 4th Floor, 215-823-2170, Philadelphia PA, USA

Table 1. Framework for good assessment: single assessments.

1. Validity or Coherence: The results of an assessment are appropriate for a particular purpose as demonstrated by a coherent body of evidence.
2. Reproducibility, Reliability, or Consistency: The results of the assessment would be the same if repeated under similar circumstances.
3. Equivalence: The same assessment yields equivalent scores or decisions when administered across different institutions or cycles of testing.
4. Feasibility: The assessment is practical, realistic, and sensible, given the circumstances and context.
5. Educational Effect: The assessment motivates those who take it to prepare in a fashion that has educational benefit.
6. Catalytic effect: The assessment provides results and feedback in a fashion that motivates all stakeholders to create, enhance, and support education; it drives future learning forward and improves overall program quality.
7. Acceptability: Stakeholders find the assessment process and results to be credible.

Table 2. Framework and Assessment Purpose.

Elements	Formative				Summative			
Validity or Coherence	X	X	X	X	X	X	X	X
Reproducibility or Consistency	X				X	X	X	X
Equivalence	X				X	X	X	X
Feasibility	X	X	X		X	X	X	
Educational Effect	X	X	X	X	X			
Catalytic Effect	X	X	X	X	X			
Acceptability	X	X	X		X	X	X	

The number of "X"s denote the importance of that element given the purpose of the test.

examination designed to meet the need for accountability for the knowledge of medical graduates (e.g. a medical licensing examination) does not produce detailed feedback that would guide future learning or curricular reform, since it has not been designed to do so.

Similarly, the elements of the framework are not of equal weight for all stakeholders, even, given the same assessment. For example, the validity or coherence of a licensing examination may be of more importance to patients than how much it costs doctors who take it or governments that finance it. Indeed, students may value the educational and catalytic effect of an assessment while regulators might be indifferent. The importance of the various elements will vary with the perspective of the stakeholder.

Interestingly, similar issues have arisen in other high-stakes processes like student selection. A recent review (Prideaux et al. 2011) of selection methods invoked the concept of ""political validity"". First introduced in the occupational psychology literature, political validity recognizes that "*there are often many stakeholders (or stakeholder groups) that influence the design of selection processes*" (Patterson and Zibarras 2011). This is evident in assessment processes too, where a wide group of stakeholders with different perspectives are involved, including current members of the profession (e.g. consultant physicians), professional bodies (e.g. Medical Colleges), regulators (e.g. Medical Council), and the government (e.g. Ministries of Education and Health). Put differently, systems of assessment require both criterion-related (concurrent/predictive) validity (using methods with robust and defensible psychometric properties) and political validity (including the interests of different stakeholders).

To respond to these issues, this paper aims to help determine whether assessments are fit for purpose by introducing and amplifying the concept for systems of assessment and listing a set of elements within the framework for assessment with short definitions of each. We then include sections on purpose (summative, formative, informative), internal stakeholders (examinees, teachers, educational managers/institutions), and external stakeholders (patients, healthcare system, and regulators/community). In these sections, we discuss how the perspectives of the stakeholders influence the design for the Systems of Assessment and the importance of the elements within the framework.

Single assessments

Framework for good assessment

The elements of the framework for good assessment in Table 1 are applicable to a single assessment and were included in the previous edition as criteria. Many of the elements described here have been presented before (for example, American Educational Research Association et al. 2014) and we continue to support their importance. However, in this framework, we place particular emphasis on the educational and catalytic effect of assessment.

The framework and assessment purpose

Table 2 presents a graphical view of the relationship between the elements of the framework and the purpose of assessment.

Formative assessment

Effective formative assessment is typically low stakes, often informal and opportunistic by nature, and is intended to stimulate learning. By definition, the framework element that is most important for formative assessment is the "catalytic effect." Formative assessment works best when it (1) is embedded in the instructional process and/or clinical work flow, (2) provides specific and actionable feedback, (3) is ongoing, and (4) is timely. On the other hand, elements such as equivalence and reproducibility-consistency are of lower priority, although care must be taken to use assessment methods and items of a similar quality to that used in summative assessment. Validity-coherence remains central, while educational effect becomes paramount. Feasibility also increases in importance in response to the fact that formative assessment is more effective if it is ongoing, timely, and tailored to examinees' individual difficulties. Likewise, acceptability, both for faculty and students, is especially important if they are to commit to the process, give credibility to feedback, and achieve a significant effect on the learning outcomes.

Summative assessment

Effective summative assessment is typically medium or high stakes and is intended to respond to the need for accountability. It often requires coherent, high-quality test material, a systematic standard-setting process, and secure administration. Consequently, elements such as validity-coherence, reproducibility-consistency, and equivalence are paramount. Feasibility, acceptability, and educational effect are also important, but not to the same degree as the psychometric criteria, which will to a great extent determine the credibility in the scores and the underlying implications for learners. A catalytic effect is desirable but is less

emphasized in this setting. However, by not providing useful feedback, we miss the opportunity to support the learners in their continuing education.

The framework and stakeholders

Table 3 presents a graphical view of the relationship between the elements of the framework and the needs of stakeholders.

Examinees

Examinees have a vested interest in both formative and summative assessment and they should be actively involved in seeking information that supports their learning. For formative assessment, educational effects, catalytic effects, and acceptability are likely to be of most concern to examinees, since these are the main drivers of learning. Examinees may take validity-coherence for granted, and feasibility will most probably be a consideration based on cost and convenience. Equivalence and reliability-consistency are of less immediate concern.

For summative assessment, issues related to perceived fairness will be most salient for examinees, as will clarity and openness about the content and process of assessment. Hence, elements such as validity-coherence, reproducibility-consistency, equivalence, and acceptability will be most important. The catalytic effect will support remediation, especially for unsuccessful examinees. When successful examinees are not provided with feedback or do not use it, the opportunity to support ongoing learning is missed.

Teachers-educational institutions

These stakeholders have interests in every facet of the assessment of students to fulfill their dual roles in education and accountability. Consistent with what was outlined above, the elements apply differently to these two roles or purposes. Validity-coherence, reproducibility-consistency, equivalence, and acceptability are particularly important to ensure correctness and fairness in decision making. Educational effects, catalytic effects, and acceptability are the cornerstones of successful student engagement and learning based on assessment.

For both teachers and institutions, student assessment information serves an important secondary purpose, namely, it speaks to the outcomes of the educational process. In other words, students' assessments, appropriately aggregated, often serve as benchmarks for comparison and formative assessment for teachers and institutions. For such data, elements like equivalence and reproducibility-consistency are a bit less important while the educational effect and catalytic effect are a bit more important. Validity-coherence is important but should be addressed as part of good student assessment, while feasibility should be straightforward since the data are already available.

Beyond repurposing student assessment, institutions engage in the assessment of individual teachers and the evaluation of programs. These applications can be broadly classified as either formative or summative and the criteria apply as noted above.

Patients

For patients, it is most important that their healthcare providers have good communication skills, appropriate qualifications, and the ability to provide safe and effective care. While patients certainly support the use of formative assessment to help the students and practitioners in the development and refinement of these skills, summative assessment is a more immediate concern since patients need to be assured of their providers' competence. Consequently, elements such as validity-coherence, reproducibility-consistency, and equivalence are of greatest importance. Feasibility, acceptability, educational effect, and catalytic effect are of less concern to this group. In the long term, however, formative assessment that supports and promotes continuous improvement will be important.

Healthcare system and regulators

The most pressing need of the healthcare system and the regulators is to determine which providers are competent and safe enough to enter and remain in the workforce. This need implies correct decisions based on summative assessment, so validity-coherence, reproducibility-consistency, and equivalence are paramount. Feasibility is also important since the healthcare systems and the regulators sometimes bear these costs.

It is becoming more common for health systems to engage in some form of continuous quality improvement (CQI). These systems are often embedded in the clinical work flow and they provide ongoing, specific, feedback to healthcare workers about their activities and outcomes. Validity-coherence is central, along with educational and catalytic effects, feasibility, and acceptability.

Likewise, many regulators are beginning to time limit the validity of their registration-licensure-certification decisions. This is often accompanied by the addition of a CQI component to the revalidation process. As with the healthcare system, such a component would need to emphasize validity-coherence, educational effect, educational quality,

Table 3. The relationship between assessment framework, stakeholders, and the purpose of the assessment.

	Validity Coherence	Reproducibility Consistency	Equivalence	Feasibility	Educational Effect	Catalytic Effect	Acceptability
Examinees	FFF	F		F	FFFF	FFFF	FFFF
	SSSS	SSSS	SSSS	S	S	S	SSSS
Teachers- Educational Institutions	FFFF	FF	F	FFF	FFFF	FFFF	FFF
	SSS	SS	SS	SSS	SSS		SSS
Patients	SSSS	SSSS	SSSS	S	S	S	S
Healthcare system	SSSS	SSSS	SSSS	SSSS	S	S	S
Regulators	SSSS	SS	SS	SSS	SSSS	SSSS	SSS

Purpose: F: Formative Assessment; S: Summative Assessment.
The more times the letter appears, the more important the type of assessment is to the stakeholder.

feasibility, and acceptability with less stress on equivalence and reproducibility-consistency.

Systems of assessment

In the 2010 version of this work, the focus was on single-purpose assessment processes, but we noted that systems of assessment required consideration at some point in the future. Such systems integrate a series of individual measures that are assembled for one or more purposes. Over the past several years, there has been considerable interest in this topic and consequently we have developed a second framework for systems.

Education and practice in the health professions typically requires multiple cognitive, psychomotor, and attitudinal/relational skills. Single methods of assessment are generally unable to capture all of these skills so multiple measures are needed. However, these measures are often applied in isolation or at least in an uncoordinated fashion. These uncoordinated measures are often combined to reach an overall decision based on the weights dictated by tradition. A system of assessment explicitly blends single assessments to achieve the different purposes (e.g. formative versus summative; high vs. low stakes) for a variety of stakeholders (e.g. students, faculty, patients, regulatory bodies).

Figure 1 illustrates the various states of assessment around the world. There are some educational and/or regulatory programs that have no assessment (Figure 1.1). This often occurs when the responsible agency does not have the resources or expertise to assess particular skills or abilities. For example, for logistical/financial reasons some countries and educational institutions are unable to mount an OSCE to assess clinical skills.

Figure 1.2 depicts a more common situation, where competence is acknowledged to be complex but only one aspect of it is assessed. For instance, it is not uncommon to mount an assessment of the cognitive aspects of competence since they are relatively easy to examine, while ignoring the performance and attitudinal/relational components.

Many institutions have addressed these deficiencies by incorporating a number of assessments aimed at different dimensions of competencies (Figure 1.3). However, as the figure illustrates, there is a limited attempt to integrate these with the overall purposes of the system. This leads to gaps in what is covered and inefficiencies that might lead to over-assessment.

Figure 1.4 comes closest to a well-functioning system of assessment. It offers the best (though not perfect) coverage of the universe of content and the most efficient use of resources. Properly done, it would offer the opportunity for triangulation based on complementary information and incorporate both formative and summative assessments Thus it would address the multiple needs of the stakeholders, support education, and ensure high quality decisions.

Framework for good assessment

The elements of a framework for good assessment in Table 4 are applicable to a system of assessment. Many of these have been described before and we continue to support their importance here (for example, National Research Council 2001; Schuwirth and Van der Vleuten 2011).

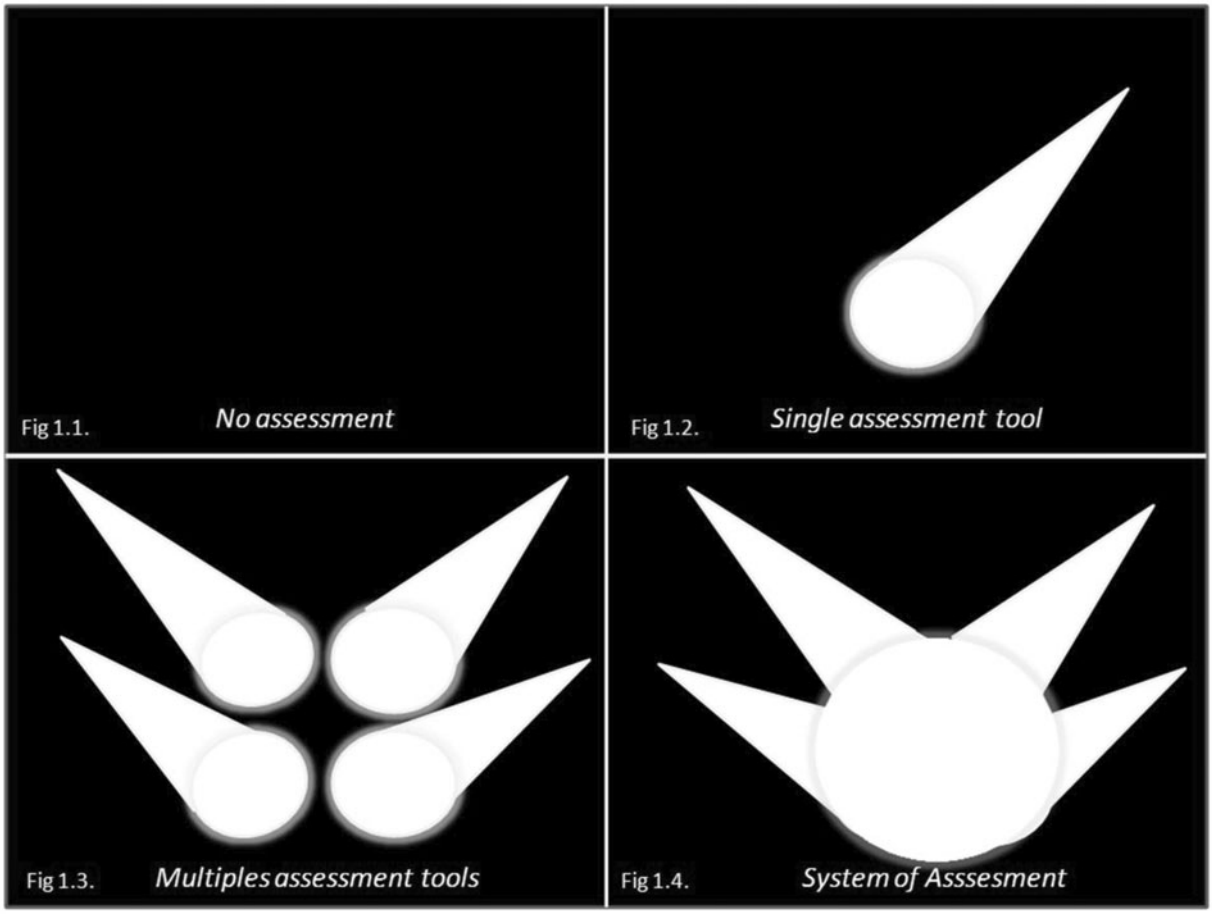

Figure 1. An illustration of the various states of assessment.

Table 4. Framework for Good Assessment: Systems of Assessment.

1. Coherent: The system of assessment is composed of multiple, coordinated individual assessments and independent performances that are orderly and aligned around the same purposes.
2. Continuous: The system of assessment is ongoing and individual results contribute cumulatively to the system purposes.
3. Comprehensive: The system of assessment is inclusive and effective, consisting of components that are formative, diagnostic, and/or summative as appropriate to its purposes. Some or all components are authentic and integrative.
4. Feasible: The system of assessment and its components are practical, realistic, efficient, and sensible, given the purposes, stakeholders, and context.
5. Purposes driven: The assessment system supports the purposes for which it was created.
6. Acceptable: Stakeholders in the system find the assessment process and results to be credible and evidence-based.
7. Transparent and free from bias: Stakeholders understand the workings of the system and its unintended consequences are minimized. Decisions are fair and equitable.

Table 5 presents examples of common systems of assessments in health professions education. For each, the purposes of the assessment system, possible components of the system, and special considerations are identified; the last may apply to individual components or may be emergent characteristics of the system as a whole. The consensus criteria identified in the 2010 report and reiterated above apply to individual assessment components, and the framework outlined in this paper applies to the system as a whole.

Some systems of assessment can reasonably be viewed as consisting of a series of assessments, often coupled with other information, for making certain kinds of multi-faceted decisions. Admissions and Licensure systems provide good examples. The quality of these systems depend heavily on the quality of the individual components, though the system may still have emergent properties resulting from the number of times individual components can be attempted, relationships among the components, and adverse impact on specific groups. Other systems of assessment are best thought of as educational interventions; Progress Testing and Programmatic Assessment provide good examples. We think that the design and evaluation of such systems of assessment should be heavily influenced by their impact on the instructional process and learning outcomes.

For all the examples in the table, feasibility, cost, acceptability, transparency/freedom from bias are important so these have been omitted from "Special Considerations." The importance of other elements of the framework varies somewhat across the types of assessment system.

Considerations in implementation of systems of assessment

While the case for systems of assessment in the health professions is strong, the concept is often not well understood, and implementation can be challenging. It is also a complex and sophisticated approach to assessment that is likely to require substantial expertise to achieve its purposes. This section offers some issues for consideration when implementing such a system, although it is far from an exhaustive list.

Definitions need to be clear and accessible to all the participants (regulators, candidates, teachers, and assessors); this reduces the scope for confusion or misinterpretation. Systems of assessment are NOT necessarily the same as progress testing or continuous assessment, although there may be shared principles. Systems of assessment are more than just combining scores over time to make a decision, for example, that enough has been achieved to "pass".

The purposes of the system need to be clear and consistent with the vision/mission of the program it serves. In an educational setting, those purposes also need to be consistent with the curriculum and the learning outcomes (i.e. constructive alignment) (Biggs 2014).

Application of the framework for systems of assessment will have two benefits; the first is fitness for purpose. Many "traditional" assessments focus on what can be done easily or has always been done, often resulting in an overemphasis on knowledge and clinical skills, at the expense of the other competencies necessary for good performance. Systems of assessment for educational programs should include a broad range of learning outcomes and assessment methods, including those that assess "difficult to measure" competencies important in clinical practice. Often, assessments based on learning- and workplace-based portfolios will be desirable. Examples include assessments related to reflective assignments, morning rounds and hand offs, record keeping, community projects, and professional behaviors. Learners "respect" what programs "inspect".

Another benefit is efficiency. High-quality assessment is resource-intensive, so information gathered should not "waste" expensive resources. Many assessments are highly predictive of each other and of subsequent similar assessments. Consequently, designing the system of assessment with these redundancies in mind should reduce the resources needed to conduct them and make assessment less resource-intensive and more feasible.

Purposeful blueprinting driven by the desired outcomes is essential for systems, just as it is for individual assessments. This promotes the validity of inferences from assessment results by guiding the selection of a range of appropriate methods, competencies, and learning outcomes, while ensuring that purposes are directly addressed. All assessments are based on a sample of a universe (preferably well-designed) of content and skills; well-constructed systems of assessment are consistent with and can extend that sampling. For example in an educational setting, competencies might be sampled from across a set of learning outcomes, ideally with overlapping scope so that, over time, most are assessed on several occasions.

A system of assessment, over time and by using multiple methods and judges, can provide greater coverage of learning outcomes by sampling different components of the "universe" of attributes and competencies with multiple, sometimes overlapping assessment episodes. A blueprint for a system of assessment can be designed to minimize the gaps in assessment coverage through appropriate sampling on a whole or program approach.

Careful selection and design of individual assessments are also required, ideally according to elements we have

Table 5. Examples of Systems of Assessment.

Type of Assessment System	Purpose(s) of System	Common System Components	Special Considerations
Admissions	1. Identify qualified applicants consistent with the mission of the program 2. Attract highly qualified applicants to matriculate 3. Screen out applicants unlikely to successfully complete training 4. Screen out applicants unlikely to be good doctors	• Completion of prerequisite courses • Letters of recommendation • Biographical information • Standardized admissions test (eg, MCAT, UKCAT, GAMSAT) • Multiple Mini-Interview • Personal interviews with faculty, current students	• Impact on recruitment of highly qualified applicants • Impact on protected groups, under-represented minorities, non-native speakers • Impact on racial/ethnic/gender/socioeconomic diversity of admitted students • Uncontrolled variation in difficulty of assessment components (eg, stringency of interviewers)
Licensure	1. Ensure doctors have knowledge and skills necessary to practice medicine safely and effectively *In some countries, there may be other ("secondary") uses of components (eg, in selection of trainees for postgraduate training)*	• Graduation from accredited medical school • Completion of specified training under supervision (typically including multiple workplace-based assessments) • 1 or more standardized tests of applied medical knowledge • OSCE-type assessment of clinical skills	• Coherence, reproducibility, and equivalence of each assessment component • Ultimate pass rates of protected groups, under-represented minorities, non-native speakers • Impact on racial/ethnic/gender/socioeconomic diversity of doctors entering practice • Success in preventing unqualified doctors from entering practice (difficult to measure) • Uncontrolled variation in difficulty of assessment components (eg, difficulty of individual tests)
Progress Testing	1. Assess student's progress against graduation-level outcomes 2. Identify and provide feedback on individual student's areas of strength and weakness 3. Provide students with a good basis for making self-assessments and judging learning needs 4. Motivate students to remediate areas of weakness 5. Provide information on instructional effectiveness to guide improvement	• Series of tests (typically MCQ but other formats also used) given 1-4 times/year beginning early in course • Feedback system assisting students in understanding performance trends • Decision rules about consequences of poor performance • Faculty advisors to aid in providing feedback and plan remediation	• Accuracy in identification of weaknesses, both for students individually and for the course as a whole • Effectiveness of feedback and remediation strategies in motivating students to address weaknesses • Impact on student learning *Progress testing serves as both an assessment and an educational intervention so it is logical for design and evaluation to include impact on learning outcomes*
Programmatic Assessment *See* van der Vleuten et al. (2015) – *"Twelve tips for programmatic assessment"- for additional information*	1. Optimize the impact of assessments on learning, decisions regarding individual students, and curriculum quality 2. Identify and provide feedback on individual student's areas of strength and weakness 3. Provide students with a good basis for making self-assessments and judging learning needs 4. Motivate students to remediate areas of weakness 5. Provide information on instructional effectiveness to guide improvement	• A "master plan" (generally based on a competency framework) for assessment and the methods to be used in combining assessment results • Multiple assessment methods chosen because of the useful guidance they provide to trainees about their learning • A flexible system for capturing, depicting, and aggregating assessment results • A mentoring system for providing feedback, reflection, remediation, and follow-up • A (committee-based) system for making higher-stakes decisions about progression • Mechanisms for using assessment results for curriculum evaluation	• "Scaling" of programmatic assessment to accommodate large class sizes • Accuracy in identification of weaknesses, both for students individually and for the course as a whole • Effectiveness of feedback and remediation strategies in motivating students to address weaknesses • Impact on student learning • Approaches to aggregating information from diverse assessment methods in provision of feedback and making higher-stakes decisions *Programmatic assessment serves both assessment and educational functions so its utility for each should be carefully considered in design and evaluation*

identified above. The use of methods aimed at different aspects of the same competence can be helpful as it will facilitate triangulation and the efficient assessment of a wide range of knowledge, skills, and behavior content (Wilkinson 2007).

The timing and sequencing of individual assessments requires careful planning regardless of the purposes of the system. This is particularly important for systems designed to reflect the learning trajectories of the individual students in an educational program. Knowledge, skills, and behaviors all evolve over the time, but competence can be achieved before the endpoint of the program. There are two broad approaches to this issue. The first, and more traditional approach, is to calibrate assessments to the expected learning outcomes for each stage/phase of a program. An example would be the organization of entrustable professional activities (EPAs) in a matrix identifying the expected level of entrustment at different stages of training. This follows the evolutionary development of competence. The second is to calibrate assessment to endpoint learning outcomes, so that at the end of a training program the expectation is that the learner has achieved the highest level of entrustment in all EPAs. This ensures readiness for independent practice, recognizing that some learners will achieve these earlier and that all learners may benefit from knowing how their performance, at all stages, relates to expectations at the endpoint. In both approaches, it is possible to "tailor" assessments to individuals and to use adaptive approaches, whereby assessment is based on a small sample of learning outcomes, with more assessments added to improve confidence, reliability, and precision for examinees close to a predefined level of mastery.

Increasing the frequency of individual formative assessments reduces the pressure created by a small number of high-stakes events, but this can also create feasibility issues. In educational programs, many competencies can be achieved at different times and in different sequences so this approach allows for some flexibility. Further, slower learning might trigger the need for remediation/additional resources. For example, systematic "progress review meetings" could be scheduled every 3–4 months. Potential outcomes from the progress review may be "on track", "needs focused learning plan", or "needs to be referred to a training committee for remediation".

Some observers are concerned about the potential impact on reliability of using the broader range of assessment methods, some of which, when used alone, demonstrate lower reliability. While this would be a concern if feedback or decisions were based on the individual measures, aggregation over assessment methods and occasions will address this reliability concern. The use of multiple methods and multiple judges on multiple occasions is sufficient to provide evidence for the achievement across a range of attributes.

Where summative decisions are needed, standard setting may be complex and require a variety of methods to make an overall decision based on the aggregated results of individual assessments. Combining these decisions in a purely quantitative and mechanical way, especially when there are numerous assessments (e.g. as part of an educational program), is challenging and may not yield a satisfactory outcome. This strategy may also trivialize important individual assessments when they contribute less to an overall decision. Where it fits the purposes of the system, it may be reasonable to make a series of non-compensatory decisions, although this poses challenges as well when the number of assessments is large. Finally use of a committee judgment process, which takes all of the measurement information into account in coming to a conclusion, may be the best alternative. This has the added virtue of allowing the use of both qualitative and quantitative information in reaching a conclusion.

In some systems of assessment, individual measures are used for both formative and summative purposes. This contributes to improved efficiency, potentially making assessments helpful in both, providing feedback and making decisions. However, we believe this dual purpose needs to be handled cautiously. Assessments designed for formative purposes often have characteristics that make them less than ideal for the summative purposes and vice versa. Moreover, trainees react differently to formative and summative assessments and using the same event(s) for both purposes may influence their effectiveness (Hodges and McIlroy 2003). In an educational setting, one approach to this challenge is to create a committee that is responsible for making decisions based on assessment results, separately from individual faculty providing feedback. Members of the committee are not those who are close to the students along the way and those who teach and give feedback do not make decisions. When multiple institutions are responsible for training, the committee will have the responsibility to implement and oversee the system of assessment in each institution, respecting local values and context. The members would have been trained appropriately and represent the various stakeholder groups. They would have the task of studying and evaluating individual assessments and how they combine to produce an acceptable result. These committees would work closely with others to optimize the individual assessments and their contribution to the overall system.

Recommendations for future work

Through the development and vetting of these frameworks, several important ideas for future work were suggested. The following list is a sample of the ideas that were generated.

- The adaptability of the frameworks to technology and artificial intelligence (AI)
- The costs and the return on investment of assessment methods
- The interaction of assessments with educational and health care systems
- The relationship between these frameworks and others reported in the literature (for example Michie et al. 2011).

Importantly, we van stralen are in a period of rapid growth in terms of technology and its impact on the acquisition and analysis of large datasets (Ellaway et al. 2014). Systems of assessment developed for local uses may need to interface with larger systems designed for similar (e.g.

national assessment systems) or dissimilar (e.g. performance support, revalidation) purposes (Pusic and Triola 2017). Moreover, they may ultimately draw on data embedded in such systems. These trends have implications for our framework and ongoing development is required to ensure that the elements we identified remain relevant.

Conclusions

The framework for systems of assessment is similar to the framework for individual assessments, for which much of the original 2010 Consensus Statement remains relevant. Some contemporary issues have emerged since that time, including an increasing appetite for transparency and meaningful feedback, consideration of increasing diversity of candidates and programs, and increasing interest among employers, regulators and patients in how medical education is delivered. For systems of assessment there are some additional elements, or at least some additional aspects, that should be considered. These relate not so much to the way individual assessment episodes are implemented, but more to the sampling, timing and decision-making, the means of combining different kinds of information from different sources, and how progression decisions are made. There is a need for careful documentation and evaluation of current attempts at developing systems of assessment to provide an evidence base to support further development.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

John Norcini, PhD, FAIMER, Philadelphia, Pennsylvania, USA

M. Brownell Anderson, National Board of Medical Examiners, Philadelphia, Pennsylvania, USA

Valdes Bollela, MD, PhD, School of Medicine of Ribeirão Preto, University of Sao Paulo Brazil

Vanessa Burch, MMed, PhD, FRCP, University of Cape Town, South Africa

Manuel João Costa, MD, PhD, University of Minho, Portugal

Robbert Duvivier, MD, PhD, Maastricht University and Parnassia Psychiatric Institute, The Netherlands

Richard Hays, MD, PhD, MBBS, University of Tasmania, Australia

Maria Felisa Palacios Mackay, MD, PhD, University of Calgary, Canada

Trudie Roberts, MB, ChB, PhD, University of Leeds, UK

David Swanson, PhD, American Board of Medical Specialists, Chicago, Illinois, USA

ORCID

John Norcini http://orcid.org/0000-0002-8464-4115
Manuel João Costa http://orcid.org/0000-0001-5255-4257
Richard Hays http://orcid.org/0000-0002-3875-3134
David Swanson http://orcid.org/0000-0002-0862-944X

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. Standards for educational and psychological testing. Washington (DC): American Educational Research Association.

Biggs J. 2014. Constructive alignment in university teaching. HERDSA Rev High Educ. 1:5–22.

Clarke MM. 2012. What matters most for student assessment systems: a framework paper. Working Paper no 1. Systems Approach for Better Education Results (SABER). Washington (DC): World Bank.

Dijkstra J, Galbraith R, Hodges BD, McAvoy PA, McCrorie P, Southgate LJ, van der Vleuten CP, Wass V, Schuwirth LW. 2012. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. BMC Med Educ. 12:20.

Ellaway RH, Pusic MV, Galbraith RM, Cameron T. 2014. Developing the role of big data and analytics in health professional education. Med Teach. 36:216–222.

Hodges B, McIlroy JH. 2003. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 37:1012–1016.

Institutional Research and Effectiveness Office, St. Olaf College. Assessment of Student Learning. [accessed 2018 Jun 17]. https://wp.stolaf.edu/ir-e/assessment-of-student-learning-2/

Michie S, Van Stralen MM, West R. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci. 6:42.

National Research Council. 2001. Knowing what students know: the science and design of educational assessment. Washington (DC): National Academy of Sciences.

National Research Council. 2014. Developing Assessment for the Next Generation Science Standards. Washington (DC): The National Academies Press.

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V et al. 2011. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 33:206–214.

Office of Academic Planning and Assessment, University of Massachusetts Amherst. 2001. Program-Based Review and Assessment. [accessed 2018 Jun 17]. https://www.umass.edu/oapa/sites/default/files/pdf/handbooks/program_assessment_handbook.pdf.

Patterson F, Zibarras LD. 2011. Exploring the construct of perceived job discrimination in selection. Int J Select Assess. 19:251–257.

Prideaux D, Roberts C, Eva K, Centeno A, Mccrorie P, Mcmanus C, Patterson F, Powis D, Tekian A, Wilkinson D. 2011. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 33:215–223.

Pusic MV, Triola MM. 2017. Determining the optimal place and time for procedural education. BMJ Qual Safety. 11:863–865.

Schuwirth LW, Van der Vleuten CP. 2011. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 33:478–485.

Van Der Vleuten CP, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. 2015. Twelve tips for programmatic assessment. Med Teach. 37:641–646.

Wilkinson TJ. 2007. Assessment of clinical performance: gathering evidence. Intern Med J. 37:631–636.