Draft 2018 Consensus Framework for Good Assessment

Background

In 2010, the Ottawa conference produced a set of consensus criteria for good assessment (1). 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 generating 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. The group believed they did. 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 after discussion reached consensus that a separate set of criteria should be developed for 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 that should be considered. 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 progress decisions are made.

To generate the criteria for systems of assessment the group began by conducting a search of the literature for purposes of identifying relevant work. We identified five sources that yielded a list of 24 criteria (2-6). Through discussion we settled on seven criteria drawn from the 24, some with modification. We then compared our criteria to the much more detailed guidelines proposed by Dijkstra and colleagues to ensure they were broadly consistent (7).

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 assessments. The context and purpose-priorities of assessment heavily influence the importance of those elements. For example, a good summative 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 the doctors who take it or the government that finances 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 such as student selection. A recent review (8) of selection methods invoked the concept of 'political validity'. First introduced in the occupational psychology literature, political validity recognises that "*there are often many stakeholders (or stakeholder groups) that influence the design of selection processes"* (9). 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 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 perspective of the stakeholder influences 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 that follow are applicable to a single assessment and were included in the previous edition as criteria. Many of the elements

described here have been described before (e.g., 10) and we continue to support their importance. However, in this framework, we place particular emphasis on the educational and catalytic effect of assessment.

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.

The Framework and Assessment Purpose

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 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 ensure a significant effect.

Summative Assessment. Effective summative assessment is typically medium or high stakes and 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 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.

Table 1: Framework and Assessment Purpose

Criteria	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 Framework and 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 the 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

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

Table 2: 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 SSSS	F SSSS	SSSS	F S	FFFF S	FFFF S	FFFF SSSS
Teachers-Educational Institutions	FFFF SSS	FF SS	F SS	FFF SSS	FFFF SSS	FFFF	FFF 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

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 different 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 several cognitive, psychomotor, and attitudinal/relational skills. Single methods of assessment are 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 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 stake) for a variety of stakeholders (e.g., students, faculty, patients, regulatory bodies).

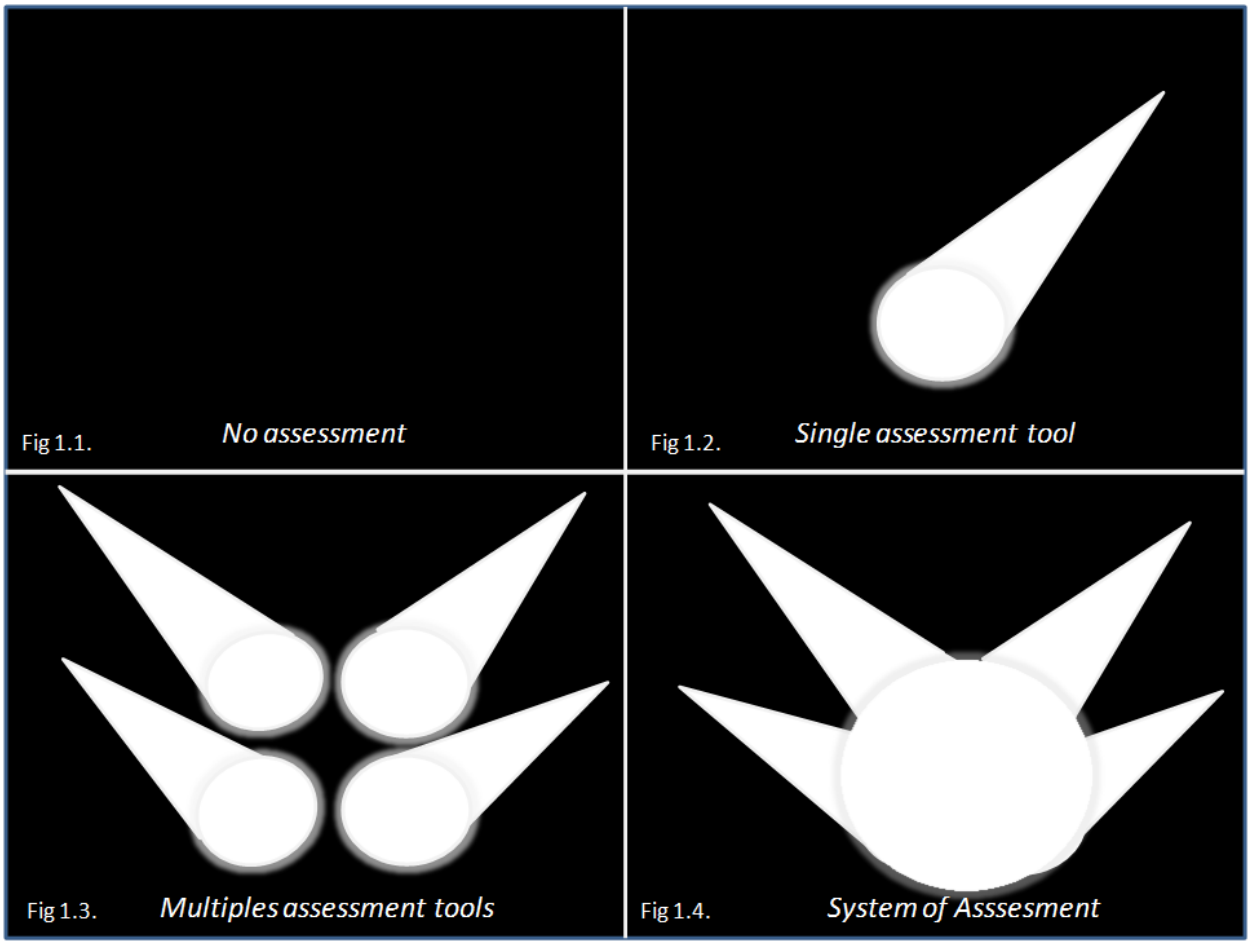
Figure 1 illustrates the various states of assessment around the world. There are some educational and/or regulatory programs that have no assessment (Figure 1.1). This often occurs when the agency does not have the resources or expertise to assess particular skills or abilities. For example, for logistical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

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

Many institutions have addressed these deficiencies by incorporating a number of assessments aimed at different dimensions of various 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.

Figure 1



Framework for Good Assessment

The elements of a framework for good assessment that follow are applicable to a system of assessment. Many of these have been described before (for example, 11,12) and we continue to support their importance here.

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.
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.

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 far from an exhaustive list.

Definitions need to be clear and accessible to all 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 the 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) (13).

Application of the framework for systems of assessment will have two benefits and 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 to good performance. Systems of assessment for educational programs should include a broad range of curriculum content and methods, including those that assess 'more difficult to measure' competencies that are important in clinical practice. Examples include reflective assignments, morning rounds and hand offs, record keeping, community responsiveness through projects, and assessing professionalism through portfolios. Learners ‘respect’ what programs ‘inspect’.

The other benefit is efficiency. All 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 run them and make assessment more feasible.

Purposeful blueprinting driven by the desired outcomes is essential for systems, just as it is for individual forms of assessment. It ensures validity by guiding the selection of a range of appropriate methods, competencies, and learning outcomes, while ensuring that all purposes are directly addressed. All assessment is based on a sample of the universe of content and well-constructed systems of assessment can extend that sampling. For example in an educational setting, competencies might be sampled from across an entire curriculum, ideally with overlapping scope such that over time most are assessed several times.

A system of assessment can, over time and using multiple methods and judges, provide greater coverage of a curriculum 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 gaps in assessment content through appropriate sampling on a whole or program approach.

Careful selection of individual assessments is also required, ideally according to elements we have 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 content (14).

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 individual students in an educational program. Knowledge, skills, and behaviors all evolve over 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 of each stage/phase of the 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 assessment approaches, whereby assessment is based on a small sample of learning outcomes, with more assessment added to improve confidence, reliability and precision.

Increasing the frequency of individual formative assessments reduces the pressure created by a small number of high stakes events but might 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".

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 methods will address this reliability concern. The use of multiple methods and multiple judges on multiple occasions is sufficient to provide evidence of achievement across a range of attributes.

Where summative decisions are needed, standard setting may be complex and require a variety of methods for individual assessments that must be aggregated to make an overall decision or a decision might be based only on the aggregated results. 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 faces limitations as well when the number of assessments is large. Finally use of a committee judgment process, which takes account of all of the measurement information 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 all assessment helpful in both providing feedback and making decisions. However, we believe this dual purpose needs to be handled cautiously. Assessments designed for formative purposes have characteristics that make them less than ideal for summative purposes and vice versa. Moreover, trainees react differently to formative and summative assessments and combining them into the same event may influence their effectiveness (15). In an educational setting, one approach to this challenge is to create a committee that is responsible for making decisions based on assessment results, as well as the feedback from the faculty. 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. 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 (e.g., 16).

Importantly, we are in a period of rapid growth in terms of technology and its impact on the acquisition and analysis of large datasets (17). 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) purposes (18). 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.

Conclusion

The framework for systems of assessment is similar to the framework for individual assessments, for which much of the 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, which 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 progress 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.

References

1. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher. 2011 Mar 1;33(3):206-14.

2. C.P.M. Van Der Vleuten, L.W.T. Schuwirth, E.W. Driessen, M.J.B. Govaerts & S. Heeneman (2015) Twelve Tips for Programmatic Assessment, Medical Teacher, 37.7, 641-646.

3. Office of Academic Planning and Assessment, University of Massachusetts Amherst, Program-Based Review and Assessment, http://www.umass.edu/oapa/oapa/publications/online handbooks /program based.pdf

4. National Research Council. 2014. Developing Assessment for the Next Generation Science Standards (2014) Chapter 6: Designing an Assessment System. Washington, DC: The National Academies Press. https://doi.org/10.17226/18409.

5. Institutional Research and Effectiveness Office Staff, “Assessment of Student Learning”, St. Olaf College, Northfield, MN https://wp.stolaf.edu/ir-e/assessment-of-student-learning-2/

6. Clarke, Marguerite, 2012, What Matters Most for Student Assessment Systems: A Framework Paper. Working Paper 1. SABER - Systems Approach for Better Education Results, Washington, DC, World Bank

7. Dijkstra J, Galbraith R, Hodges BD, McAvoy PA, McCrorie P, Southgate LJ, Van der Vleuten CP, Wass V, Schuwirth LW. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. BMC medical education. 2012 Dec;12(1):20.

8. Prideaux D, Roberts C, Eva K, Centeno A, Mccrorie P, Mcmanus C, Patterson F, Powis D, Tekian A, Wilkinson D. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher. 2011 Mar 1;33(3):215-23.

9. Patterson F, Zibarras LD. Exploring the construct of perceived job discrimination in selection. International Journal of Selection and Assessment. 2011 Sep 1;19(3):251-7.

10. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, Psychological Testing (US). Standards for educational and psychological testing. Amer Educational Research Assn; 2014.

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

12. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. Medical Teacher. 2011 Jun 1;33(6):478-85.

13. Biggs J. Constructive alignment in university teaching. HERDSA Review of higher education. 2014 Jul;1(1):5-22.

14. Wilkinson TJ. Assessment of clinical performance: gathering evidence. Internal Medicine Journal. 2007 Sep 1;37(9):631-6.

15. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Medical Education. 2003 Nov 1;37(11):1012-6.

16. Michie, Susan, Maartje M van Stralen, and Robert West. “The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions.” *Implementation Science : IS* 6 (2011): 42. *PMC*. Web. 14 Mar. 2018.

17. Ellaway RH, Pusic MV, Galbraith RM, Cameron T. Developing the role of big data and analytics in health professional education. Medical teacher. 2014 Mar 1;36(3):216-22.

18. Pusic MV, Triola MM. Determining the optimal place and time for procedural education *BMJ Qual Safty* Published Online First: 09 August 2017. doi: 10.1136/bmjqs-2017-007120