Advances in Medical Education and Practice

Dovepress
open access to scientific and medical research

Open Access Full Text Article

PERSPECTIVES

# Deconstructing programmatic assessment



This article was published in the following Dove Press journal:
Advances in Medical Education and Practice

Tim J Wilkinson[1]
Michael J Tweed[2]

[1]Education Unit, University of Otago, Christchurch, New Zealand; [2]Education Unit, University of Otago, Wellington, New Zealand

**Abstract:** We describe programmatic assessment and the problems it might solve in relation to assessment and learning, identify some models implemented internationally, and then outline what we believe are programmatic assessment's key components and what these components might achieve. We then outline some issues around implementation, which include blueprinting, data collection, decision making, staff support, and evaluation. Rather than adopting an all-or-nothing approach, we suggest that elements of programmatic assessment can be gradually introduced into traditional assessment systems.

**Keywords:** assessment, implementation, medicine, decision making

## Introduction

Programmatic assessment offers many advantages over traditional systems of assessment: decisions are made on a wide body of evidence, assessment is used to guide learning, and robustness is obtained through the collation of many pieces of information rather than just depending on highly reliable, but limited snapshots of performance, such as through structured high-stakes examinations. Many implementations of programmatic assessment have been made within renewed or highly refined curricula and often seem labor intensive, relying on staff time, completion of portfolios of learning, and collation of data.

If programmatic assessment were a drug, it would probably be about to enter phase IV trials. That is, the nature of programmatic assessment has been described and it has been shown to be effective in certain circumstances, but many are now wanting to see how it could be applied in wider contexts. How could this "drug" of programmatic assessment work in more real-world and messy contexts? In order to understand this, we need to be clear about what the components of programmatic assessment are, what each component does, how much of each component is required, and the degree to which is it an "all or nothing" intervention. That is, do all components need to be implemented in exactly the same way everywhere or can there be some tailoring to suit context? We suggest that there is considerable scope for flexibility. Implementing part of programmatic assessment as time, space, and feasibility allows may be better than being overwhelmed and doing none of it; evolution rather than revolution.

In making this suggestion, we note that there are parallels with some other education innovations, such as the objective structured clinical examination (OSCE) and problem-based learning (PBL). When the OSCE first came out it was shown to produce scores that were more reliable than any other forms of assessment of clinical skills. It

Correspondence: Tim J Wilkinson
Education Unit, University of Otago, Christchurch, PO Box 4345, Christchurch, New Zealand 8140
Tel +64 3 364 0530
Email tim.wilkinson@otago.ac.nz

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/AMEP.S144449

was assumed that this reliability came from its "objectivity" and, in turn, it was thought that the objectivity came from the use of checklists and specifications within each OSCE station. We now know that is not the case. Instead, the OSCE works because it has multiple stations and multiple examiners; both these aspects improve content representation (by sampling more broadly across the curriculum) and score reliability by using multiple observations. The checklists turned out to be red herrings.[1] In other words, the effectiveness of the "black box" of the OSCE turned out to be due to some, but not all, subcomponents. We suggest that the "black box" of programmatic assessment may also have subcomponents necessary for its success that are not all initially suspected.

Likewise, with PBL, it was initially thought that it could only work if students decided their own learning outcomes and if scenarios were framed around patient problems. It was thought that by doing this, somehow it would also improve problem-solving skills. Many of the benefits of PBL do indeed arise from a more student-centered approach to learning, but they also come from the more integrated manner in which curriculum content is presented. The evidence of any effect on problem solving is weaker.[2] This also illustrates that not all aspects of PBL, as it was initially constructed, are needed to make it work.

We begin with a description of programmatic assessment, then we outline what we believe are the key components and what these components might achieve. We then make some suggestions on ways in which the key components could be implemented.

## What is programmatic assessment?

The term "programmatic assessment" reflects a move away from focusing on individual assessment tools or episodes toward considering a suite of assessment components that are part of a larger whole. This has arisen from the acknowledgment that all assessment tools or methods have strengths and weaknesses: some have higher score reliability but lower validity for robust decision making, whereas some are more valid and authentic but have lower score reliability. Others are sufficiently robust to make high-stakes decisions but less feasible. A single assessment or even several equivalent assessments will not provide information for a complete picture. This leads to the conclusion that a combination of methods will be needed so that the limitations of the information provided by one method can be countered by the strengths of another. A consequence of this approach is that decisions need to be separated from the assessment episodes. Each assessment episode will not be sufficient in itself to justify a robust decision. Instead, the decision needs to be deferred until there is sufficient evidence from a variety of sources. This component has profound implications for students and assessors. Instead of always deciding whether a learner has "passed" after each assessment episode, the strengths and weaknesses of that learner are recorded, and may be used to guide future learning and performance. Once there is a sufficient amount of information outlining an individual learner's strengths and weaknesses, from a variety of assessment episodes, a robust decision can be reached. The amount of evidence required for each decision will depend on the potential consequence of that decision for a learner: high-stakes decisions (such as failing a year or not graduating) require more evidence than low-stakes outcomes (such as deciding where next to focus one's learning). As van der Vleuten and colleagues[3] describe, "The summative/formative distinction is reframed as a continuum of stakes, ranging from low-stakes to high-stakes assessment. Decision-making on learner progression is proportionally related to the stakes." Moreover, these assessment episodes are clustered by attributes so that results from a variety of assessment methods are synthesized. Finally, and most importantly, freeing ourselves from always having to make decisions after each assessment episode, and worrying less about the reliability of each assessment, means that we can shift to consider the impact on learning. In other words, we can ensure that each time a learner is assessed, the focus is on guiding learning, rather than on making a decision. The learner receives feedback during this longitudinal journey so that any ultimate decision will not come as a surprise.[4] This has led to the shift from assessment of learning to assessment for learning.

One of the sticking points for some assessors relates to the concepts of reliability and objectivity. There has been a view that global judgments, without structured rubrics, or without sufficiently controlled circumstances, can be unacceptably unreliable. Some of these views have been reinforced by early research related to OSCEs which showed increased score reliability. This was assumed to be due to their greater structure and use of checklists. What has emerged, however, is that their reliability comes from including more cases, thereby reducing case specificity (variation due to the patient seen, rather than the learner's ability); and more examiners, thereby reducing the effect of examiner idiosyncrasies (variation due to examiners, rather than the learner's ability). Provided there are enough examiners and enough observations, and provided the examiners understand the purpose of the assessments, it has been found that achievement of reliable decisions does not require standardization[5] and is

not worsened by the use of global judgments.[1] All methods require sufficient sampling and those methods which are less structured or standardized, such as the oral examination, the long case examination, the Mini-Clinical Evaluation Exercise (Mini-CEX), and the incognito standardized patient method, can be entirely, or almost, as reliable as other more structured and objective measures provided there are sufficient observers and observations.[5] Instead, decision reliability comes from synthesizing multiple observations of a learner's ability, before coming to a decision.

## What problems might programmatic assessment solve?

We outline some problems in assessment where a programmatic approach can offer solutions.

### Confidence to fail is low

Single assessors observing single assessments often recognize that they have insufficient information to be confident to make a judgment. If they are also expected to make a decision on a learner, based on this information, they generally tend to favor the learner and give a pass rather than a fail.[6,7] If a series of such decisions is then aggregated, all that the decision makers see is a consistent series of passes, whereas, in some cases, the learner may persistently have been below the required standards, but it is only in seeing all the information together that such a picture emerges. This contributes to the phenomenon of "failure to fail" and is often seen as an examiner problem; however, often it is more an aggregation of information problem where assessors are asked to make high-stakes decisions on limited information.[8] Instead, if the single observations are captured in narrative or descriptive forms, which are then fed into a larger decision-making process where other such information is also scrutinized, a more confident and defensible decision can be made.

### Inappropriate combination of assessment results

Aggregation of results can occur at defined periods of time (cross-sectionally) or longitudinally. If aggregated in a cross-sectional manner, this can lead to several attributes being combined to make a single decision, for example, results on knowledge being combined with results on communication. Many medical courses have clinical attachments, where satisfactory performance on each attachment is required. This is an example of cross-sectional aggregation. The risk of this is greater when assessment results are reported as numbers, as there is a temptation to simply add the numbers to inform decisions. The disadvantage of this is that there can be inappropriate compensation for pass–fail decisions, e.g. good knowledge compensating for poor communication.[9] Such compensation can lead to both an incorrect pass and an incorrect fail. Instead, combining results by attribute may assist with detecting patterns.

An example might be to synthesize all observations of communication across several time points, pull these together, and look to see a learner's trajectory on this attribute and/or if they have attained the required standard. A similar exercise could be undertaken for other attributes, such as physical examination skills. This contrasts with a more traditional assessment approach where the assessments of a learner's abilities in communication and physical examination might be combined at each point, thus obscuring a pattern where a learner might be consistently strong in communication but consistently weak in physical examination.

### Biases in decision making

There are known biases in decision making. Some decision makers place more salience on the strength of evidence (noting a single aberrant behavior) than on the weight of evidence (and pattern of behaviors over time).[7] There is also the well-described halo effect,[10] where a learner's strength in one area makes it harder for an assessor to see their weaknesses in another. A safeguard against these examples of biases is to ensure that decisions are made by a group of people.[11,12]

### Results come as a surprise or with insufficient time to remedy deficits

If a pattern of deficits only emerges once all assessment results are available (such as at the end of an academic year), this may leave insufficient time to remedy any deficits and the learners may be unaware of the issues. Instead, if the feedback from assessment is used to guide learning on a continuous basis, alongside sufficient guidance, many problems can be remedied before they become high stakes.[13]

Finally, and not specific to programmatic assessment, but in situations where staff time and resource are limited, priority should be placed on guiding those learners who have the greatest learning needs. While all learners benefit from feedback and all learners can improve, if we had to prioritize, we should provide most attention to those learners who are at greatest risk of failure.

## Models of programmatic assessment

Programmatic assessment has been described in medical programs,[13–17] postgraduate residency programs,[18] veterinary courses,[19] and nutrition and dietetics.[20] Developed and implemented mostly in the Netherlands,[16,17,19,21] it has also been implemented in the USA,[14,15] Canada,[18] Australia,[20,22] and New Zealand.[13]

It has been observed that programmatic assessment may be expensive and labor intensive as it increases time for feedback, in a quantitative and qualitative form, and may require mechanisms to increase learner support in order to guide feedback uptake and self-directed learning, and a decision-making arrangement that includes groups of experts making a holistic judgments.[3] Some descriptions of programmatic assessment focus primarily on the use of a portfolio as a means of capturing the evidence and contributing to decision making.[14,15,22] However, it may be timely to consider (and encourage) other models. To do this requires us to dissect out the key elements, which we suggest are:

1. Create clear expectations of required learning
2. Undertake purposeful selection of assessments
3. Focus on those learners who need extra attention and/or extra information
4. Separate data from decisions
5. Aggregate by attribute, not method or timing
6. Make decisions on aggregate, not on individual assessments
7. Promote sharing of information and dialogue around narrative rather than numbers
8. Maximize the assessments to guide learning

The first three components should be familiar to those involved in designing any form of assessment, and are not unique to programmatic assessment. The last five components are more specific to programmatic assessment and are responsible for many of its advantages.

Specifically, separating data from decisions helps failure to fail[8] "Just give us the information and we'll make the decisions collectively", as any difficult decision lies with a group of people, not a single assessor.

Aggregating by attribute, and making decisions on the aggregated information, promote appropriate compensation and reduce the risk of inappropriate compensation, such as poor knowledge being compensated by good consultation skills. This can also contribute to the "halo effect," where there is a trap that an assessor may judge a student to have strengths in one area based on observed strengths in a different area.[10]

Describing performance by words, rather than numbers, assists both the supervisors and the learners to guide learning.[23] We recognize that a result, for example, of 5 on a 9-point assessment scale carries much less informative content than a description of what someone did well and needs to work on. Yet, somehow the use of numbers implies greater objectivity, a view not supported by evidence. The use of numbers also makes it too easy just to add them up: the total of the OSCE plus the multiple choice question examination determines the level of performance, thereby increasing the risk of inappropriate compensation.[9] It has been shown that the use of a narrative, without the need for a portfolio, to describe areas a learner might need to work on has shown positive impacts on a system's ability to detect and act on learners displaying poor professional behaviors.[13]

## Issues around implementation

Many models require the use of a portfolio[14,15,22] and large-scale changes to assessment methods. Any system of assessment benefits from a clear articulation of purpose, engagement of personnel with sufficient expertise in assessment, central governance, and in-built evaluation.[24] We suggest that the seven key elements described above can be distilled into the following key components of implementation:

1. Blueprinting
2. Data collection
3. Decision making
4. Staff support
5. Evaluation

### Blueprinting

This starts with being clear on the purpose, or purposes, of assessment. Broadly, there are usually three main purposes: 1) to guide learning; 2) to inform quality improvement of curriculum development; and/or 3) to make high-stakes decisions on learners. The first purpose has a focus on learning, particularly on where the individual learners have difficulty. The second purpose focuses on how might a curriculum change to address difficulties faced by many learners; how is the curriculum driving learning behaviors? The third purpose includes certifying whether someone is ready to proceed to the next stage of their training; this could be certifying competence to practice or deciding that they are ready to move to the next year of the course. Sometimes these broad purposes of assessment overlap, but at other times

trying to achieve one may undermine another. For example, if the assessment is to guide learning, learners should be encouraged to openly acknowledge weaknesses, but if the assessment is to certify competence, learners are likely to actively conceal weaknesses. Another example could be if the curriculum and its assessment goals are to promote collaboration, yet the assessment is also used for competitive selection, the collaboration and competition aspects may be in conflict. Often these issues can be resolved, but the key point is to be explicit about assessment purposes and to be explicit about any compromises that may need to be made. This often leads to the need to be clear about the important attributes of learners that need to be assessed. Such attributes might include, for example, teamwork skills, underpinning knowledge, procedural skills, and interpersonal interactions.

This is then followed by forming a blueprint for the program of assessment, whereby there is clarity about how each assessment tool contributes to each attribute. Often this is documented as a grid, where the rows are the attributes and the columns are the assessment tools. The cells of the grid indicate which tools are designed to assess which attributes. Traditionally, decisions are made by looking at the results of each assessment episode or short time period of assessments, such as within a single clinical attachment (columns), whereas under programmatic assessment, decisions are made from the results for each attribute (rows).

## Data collection

In order to decide whether a student has achieved a satisfactory performance level in each attribute, there needs to be a collection of evidence. Such evidence comprises the results from each assessment episode. Under programmatic assessment, the data decisions are separated from the data collection. This means that the data that are collected need to be framed in ways to explain where a learner is at in their learning. In other words, ideally they document their strengths and weaknesses, rather than any decision. The data also need to be collected in ways that make it easy to collate and synthesize, so that they can be presented in meaningful ways when it comes to decision making. They should also be presented back to the learner in ways that make it easy for them to decide where to focus their learning. This is an area where there may be many innovative mechanisms to collect such data. Decision making should ideally be by attribute rather than by assessment method; for example, have the learners reached a satisfactory level on communication, rather than have they passed a Mini-CEX? The implications of this for data collection are that records of performance should be matched to the attribute of importance (e.g. communication), not just to tool(s) used to assess that performance (e.g. Mini-CEX). Information technology has the potential to be very useful here.

## Decision making

There are two broad areas of decisions that need to be made: lower stakes decisions guiding learning, and higher stakes decisions on progress or certification. The process of collating evidence is similar, but the strength and weight of evidence need to be greater for the high-stakes decisions than for the low-stakes decisions.

Once the data have been collected and are able to presented in useful ways, there needs to be a system where they are synthesized and scrutinized. The data related to each attribute should then be considered collectively and a decision made for that attribute for each learner. The process is then repeated for the next attribute before a final decision on each learner is made. This has the potential to be labor intensive, but for many learners the decisions are likely to be very straightforward as many will clearly surpass all the requirements (that is, they will pass) and some may clearly fall short (that is, they will fail). For these learners, the data presentation and ensuing decision will be quick and some of this could even be algorithm driven or automated. There will always, however, be a group for whom decisions are more difficult, particularly where there is uncertainty over whether there is sufficient evidence to determine a learner's ability in a particular attribute, where there is conflicting evidence, or where a learner is clearly stronger in one attribute but weaker in another. This is where expert judgment will be needed. As in many other areas of decision making and judgment where defensibility is needed, such decisions are often better made by a group of experts.[11] As such, for high-stakes decisions, most programmatic assessment systems make use of progress committees. The low-stakes decisions can often be guided by a supervisor, a mentor, and/or, of course, the learner.

Ways to enhance the defensibility and robustness of decisions include training, use of narrative information, and having explicit processes for appeal. This process will be assisted by having a common frame of reference for all assessments (e.g. by domain or attribute) but also needs to be mindful of factors that impact on group decision making, such as the power of individual group members, the place of veto (or loss of veto), and therefore the extent to which individual opinions may carry inappropriate weight. All these processes have parallels with qualitative research and with clinical decision making. Credibility and dependability in qualitative research

can often be achieved through triangulation, prolonged engagement, member checking, audit trail, and dependability audit.[25] Similar processes can be used in assessment decision making.[25] Likewise, health professionals are used to making decisions on the basis of multiple disparate pieces of information (e.g. evaluating the disparate pieces of evidence of heart sounds, venous pressure, and edema, in order to determine the decision of presence or absence of heart failure). Other parallels between assessment decision making and clinical decision making have been made.[26]

## Staff support

Programmatic assessment will require, for many staff, a paradigm shift.[27] This inevitably means that many staff will need provision of information, guidance, and/or training. Part of this this will include ensuring that there is good documentation around the rationale for the program of assessment and how the various components fit together.[11,28] Staff and learners will be particularly interested in the process around decision making: What evidence goes where? How is it synthesized? How are credibility and dependability assured?

Asking staff to change what they do also raises issues around cost and cost-effectiveness, an area that has been addressed elsewhere.[3] While there may be more effort needed in some areas, it is suggested that there could be less effort needed in others so that redistribution of resources may be part of the solution. Such redistribution may be to make increased use of routinely collected data and observations (such as through workplace-based observations) and less use of expensive high-stakes examinations.[3]

## Evaluation

Any system of assessment needs built-in quality improvement processes and therefore evaluation. There is also scope for further research. While we are starting to obtain some evidence on the impact[16] and effectiveness of programmatic assessment,[29] we will need more, particularly around the effectiveness of methods of data collection and decision making, but also on the impact on learning behaviors.

# Conclusion

Programmatic assessment represents a paradigm shift. As such, there is a risk that it may be seen as able to solve all problems, countered by others who may see it as solving none, but causing many. The reality is likely to lie in the middle ground. Also, as with other paradigm shifts, there may be purists who feel it can only be done one way, countered by others who see programmatic assessment more as a way of working whereby there is scope for flexibility in its development, implementation, evolution, and forms.

We have tried to outline the key components of programmatic assessment and to emphasize how these may relate to purpose so that others may see that adopting components of programmatic assessment into their existing assessment systems may be feasible and the start of a process toward quality improvement of assessment in general.

# Disclosure

The authors report no conflicts of interest in this work.

# References


1. Wilkinson TJ, Frampton CM, Thompson-Fawcett MW, Egan T. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. *Acad Med.* 2003;78(2):219–223.
2. Norman GR. Problem-solving skills, solving problems and problem-based learning. *Med Educ.* 1988;22(4):279–286.
3. van der Vleuten CPM, Heeneman S. On the issue of costs in programmatic assessment. *Perspect Med Educ.* 2016;5(5):303–307.
4. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach.* 2012;34(3):205–214.
5. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ.* 2005;39(3):309–317.
6. Tweed M, Ingham C. Observed consultation: confidence and accuracy of assessors. *Adv Health Sci Educ Theory Pract.* 2010;15(1):31–43.
7. Tweed MJ, Thompson-Fawcett M, Wilkinson TJ. Decision-making bias in assessment: the effect of aggregating objective information and anecdote. *Med Teach.* 2013;35(10):832–837.
8. Wilkinson TJ, Wade WB. Problems with using a supervisor's report as a form of summative assessment. *Postgrad Med J.* 2007;83(981):504–506.
9. Tweed M. Passing assessment should not just be jumping hurdles. *Focus on Health Professional Education: A Multi-disciplinary Journal.* 2010;11(3):85–89.
10. Nisbett RE, Wilson TD. The halo effect: evidence for unconscious alteration of judgments. *J Pers Soc Psychol.* 1977;35(4):250–256.
11. Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, Driessen EW. From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. *Adv Health Sci Educ Theory Pract.* Epub 2017 Oct 14.
12. Giles J. Wisdom of the crowd. *Nature.* 2005;438(7066):281.
13. Wilkinson TJ, Tweed MJ, Egan TG, et al. Joining the dots: conditional pass and programmatic assessment enhances recognition of problems with professionalism and factors hampering student progress. *BMC Med Educ.* 2011;11(1):29.
14. Dannefer EF, Henson LC. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Acad Med.* 2007;82(5):493–502.
15. Fishleder AJ, Henson LC, Hull AL. Cleveland Clinic Lerner College of Medicine: an innovative approach to medical education and the training of physician investigators. *Acad Med.* 2007;82(4):390–396.
16. Heeneman S, Oudkerk Pool A, Schuwirth LWT, van der Vleuten CPM, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. *Med Educ.* 2015;49(5):487–498.
17. Driessen EW, van Tartwijk J, Govaerts M, Teunissen P, van der Vleuten CPM. The use of programmatic assessment in the clinical workplace: a Maastricht case report. *Med Teach.* 2012;34(3):226–231.
18. Chan T, Sherbino J; McMAP Collaborators. The McMaster Modular Assessment Program (McMAP): a theoretically grounded work-based assessment system for an emergency medicine residency program. *Acad Med.* 2015;90(7):900–905.

19. Bok HGJ, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Med Educ.* 2013;13(1):123.
20. Jamieson J, Jenkins G, Beatty S, Palermo C. Designing programmes of assessment: a participatory approach. *Med Teach.* 2017;39(11):1182–1188.
21. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach.* 2011;33(6):478–485.
22. Roberts C, Shadbolt N, Clark T, Simpson P. The reliability and validity of a portfolio designed as a programmatic assessment of performance in an integrated clinical placement. *BMC Med Educ.* 2014;14(1):197.
23. Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Surv Anesthesiol.* 2015;59(4):169.
24. Timmerman AA, Dijkstra J. A practical approach to programmatic assessment design. *Adv Health Sci Educ Theory Pract.* 2017;22(5):1169–1182.
25. Driessen E, van Der Vleuten C, Schuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ.* 2005;39(2):214–220.
26. Schuwirth L, van der Vleuten C, Durning SJ. What programmatic assessment in medical education can learn from healthcare. *Perspect Med Educ.* 2017;6(4):211–215.
27. Ellis R, Hogard E. Programmatic assessment: a paradigm shift in medical education. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education.* 2016;8(3):29501–29515.
28. Dijkstra J, Van der Vleuten CPM, Schuwirth LWT. A new framework for designing programmes of assessment. *Adv Health Sci Educ Theory Pract.* 2010;15(3):379–393.
29. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ.* 2012;46(1):38–48.