education sciences

MDPI

Case Report

From Traditional to Programmatic Assessment in Three (Not So) Easy Steps

Anna Ryan * and Terry Judd

Department of Medical Education, Melbourne Medical School, University of Melbourne, Melbourne 3010, Australia; terry.judd@unimelb.edu.au

* Correspondence: annatr@unimelb.edu.au

Abstract: Programmatic assessment (PA) has strong theoretical and pedagogical underpinnings, but its practical implementation brings a number of challenges—particularly in traditional university settings involving large cohort sizes. This paper presents a detailed case report of an in-progress programmatic assessment implementation involving a decade of assessment innovation occurring in three significant and transformative steps. The starting position and subsequent changes represented in each step are reflected against the framework of established principles and implementation themes of PA. This case report emphasises the importance of ongoing innovation and evaluative research, the advantage of a dedicated team with a cohesive plan, and the fundamental necessity of electronic data collection. It also highlights the challenge of traditional university cultures, the potential advantage of a major pandemic disruption, and the necessity for curriculum renewal to support significant assessment change. Our PA implementation began with a plan to improve the learning potential of individual assessments and over the subsequent decade expanded to encompass a cohesive and course wide assessment program involving meaningful aggregation of assessment data. In our context (large cohort sizes and university-wide assessment policy) regular progress review meetings and progress decisions based on aggregated qualitative and quantitative data (rather than assessment format) remain local challenges.

Keywords: programmatic assessment; assessment programs; assessment reform; formative assessment; feedback; medical education

Citation: Ryan, A.; Judd, T. From Traditional to Programmatic Assessment in Three (Not So) Easy Steps. *Educ. Sci.* **2022**, *12*, 487. https://doi.org/10.3390/educsci12070487

Academic Editors: Marjan J. B. Govaerts, Cees van der Vleuten and Suzanne Schut

Received: 29 May 2022
Accepted: 13 July 2022
Published: 14 July 2022

1. Introduction

Proposed by van der Vleuten and Schuwirth [1,2] as a response to the limitations of standardised assessments and to optimise the learning and decision-making functions of assessment, programmatic assessment is intended to provide an integrated view of assessment design. Aggregation and interpretation of data from across a range of assessment formats is used to direct learning and support judgements of students' developing knowledge, skills, and attributes [3]. Best characterised as an approach rather than a strict method, considerable effort has gone into defining programmatic assessment's key elements. This work culminated during the 2020 Ottawa conference where, through a consensus process, an assembled expert group recognised and formalised 12 core principles [3]. These principles were further distilled into three implementation themes: 1—"Continuous and meaningful feedback to promote dialogue with the learner for the purpose of growth and development"; 2—"Mixed methods of assessment across and within the context of a continuum of stakes"; and 3—"Establishing equitable and credible decision-making processes including principles of proportionality and triangulation" and used to frame a high-level comparison of 15 medical and health science assessment programs [4]. This comparison nicely demonstrates how contrasting elements of disparate assessment programs can successfully address the various principles and aims of programmatic assessment. It also highlights which of the principles are more readily achieved and which appear to more difficult to implement than others.

Our paper builds on the work laid out in the Ottawa consensus papers by providing a more detailed case report of one of the 15 assessment programs included in their comparison (the University of Melbourne, graduate-entry Doctor of Medicine (MD) program). We describe a decade of assessment innovation and reform within this program, representing a transition from a more traditional assessment program to what can clearly be recognized as programmatic assessment (although that was not necessarily our intention from day one). We begin by outlining our medical program's existing assessment existing approach at the start of 2012, situating it in reference to the overarching principles and implementation themes outlined in the Ottawa consensus papers [3,4]. We provide a visual representation of self-assessed alignment with each principle through five levels of shading in Figure 1. No shading correlates with no reflection of that principle within our assessment program through to heavy shading representing full alignment of our program with that principle. We then describe the transitional process, breaking it into three key steps. We consider "step" to be an appropriate metaphor (despite the differences in scale, of duration, complexity, difficulty, and potential impact) as it captures the sequential and additive contributions of the changes involved. At each step we again reference the principles and implementation themes of the Ottawa consensus papers [3,4] and self-assess alignment of our program with those principles to illustrate the key areas of change over time. Step 1 describes a series of innovations around formative assessment and feedback, and how we subsequently extended these more generally within our existing program of assessment. Step 2 describes a fundamental organisational change within our department which facilitated coordinated assessment delivery and reform in conjunction with the adoption of key assessment related technologies, while step 3 details a substantial and ongoing process of curriculum renewal and the assessment reform associated with that renewal. In describing our work in this way, our aim is to contextualise our single-institution case report within the ongoing and important conversations around the nature and value of programmatic assessment versus more traditional assessment programs.

1.1. Case Context

Established in 1853, and with more than 50,000 full and part-time students, The University of Melbourne is consistently ranked as one of Australia's top universities [5]. The university's medical school offers a four-year master's-level graduate entry medical program (Melbourne MD—Doctor of Medicine). Of the approximately 360 students entering the course each year, most have an undergraduate biomedical degree. Each year of the existing MD program (phasing out from 2022) includes a large composite clinical subject, contributing between 43.75 and 93.75 credit points towards a full year of study equivalent of 100 points. Students complete an initial campus-based year primarily focused on core medical skills and biomedical knowledge and then move into clinical years based around core clinical rotations (year 2), specialty health rotations (year 3), and a research-oriented semester and a transition to practice semester (final year). During the three clinical years the students are based in groups of approximately 10–60 students at one of ten clinical school locations (inner metropolitan, outer metropolitan, or rural) where they are supported by a local team of academic and professional staff.

1.2. The 2012 Assessment Program

A decade ago, the predominant goal of assessment within the Melbourne MD was to facilitate robust and defensible progress decisions. The relatively large composite clinical subjects utilised a range of assessment formats across all years of the program to capture the knowledge and skills required by students to progress and eventually graduate. Written examinations (multiple choice [MCQ] and short answer question [SAQ] format), assignments, standardised clinical assessments (objective structured clinical exam [OSCE]) and less standardised assessments such as mini-CEX were used across the four years of the program. Two subjects were excluded from grading—the final 43.75-credit point Transition to Practice subject, and a 6.25-point Student Conference subject (delivered in

all four years). The remainder of the 331.25 credit points contributed to an overall course aggregate (converted to a z-score) which was then used to rank students against the other three state-based medical programs for selection into medical internship.

	Description	2012	Step 1	Step 2	Step 3
1	"Every (part of an) assessment is but a data-point"				
2	"Every data-point is optimised for learning by giving meaningful feedback to the learner"				
3	"Pass/fail decisions are not given on a single data-point"				
4	"There is a mix of methods of assessment"				
5	"The method chosen should depend on the educational justification for using that method"				
6	"The distinction between summative and formative is replaced by a continuum of stakes"				
7	"Decision-making on learner progress is proportionally related to the stake"				
8	"Assessment information is triangulated across data-points towards an appropriate framework"				
9	"High-stakes decisions (promotion, graduation) are made by in a credible and transparent manner, using a holistic approach"				
10	"Intermediate review is made to discuss and decide with the learner on their progression"				
11	"Learners have recurrent learning meetings with (faculty) mentors/coaches using a self-analysis of all assessment data"				
12	"Programmatic assessment seeks to gradually increase the learner's agency and accountability for their own learning through the learning being tailored to support individual learning priorities"				

Figure 1. Progress towards Programmatic Assessment in 2012, and at Step 1, Step 2, and Step 3. Alignment of the assessment program with Henneman and colleagues' 12 Principles [3] is indicated in the far-right columns. Strength of the alignment is indicated by depth of shading. Thematic Groupings [4] are indicated through coloured shading across the principles—those related to theme 1 (continuous and meaningful feedback) are shaded in green, theme 2 (mixed assessment methods) in blue, and theme 3 (decision-making processes) in orange.

Established processes for blueprinting, item development, and robust standard setting methods were used across the various assessment formats. Peer review was seen as an essential part of assessment quality, and written and clinical assessment review panels met regularly to write and review items used in summative assessments. Routine evaluation processes, including entire test and individual item analysis as well as detailed student perception surveys, were used to monitor assessment quality and inform ongoing improvements.

Subject co-ordinators were largely responsible for the blueprinting, development, and selection of individual assessments, which were loosely mapped to subject learning objectives aligned with a selection of the 67 course attributes. End-of-semester and end-of-year examinations were heavily weighted, contributing a significant proportion of the overall grade for each subject. The written assessments reflected elements of a more historic

input/output model of education with a focus on recall (and some application) of recently learned content within a given subject rather than integration of competencies across the program. Individual assessment hurdles were introduced to limit 'gaming' of results (i.e., passing overall despite failing either the clinical or written component of a subject). When critiqued against "criteria for good assessment" at that time [6], the strengths of the assessment program were reliability, equivalence and construct validity representing a previously dominant perception of assessment as a measurement challenge strongly influenced by the field of test psychology [7]. The assessment formats were feasible and acceptable within the local context, but their educational and catalytic effects were weak points—with very little focus on assessment for learning, no routine provision of feedback after high-stakes assessments, and few examples of high-quality low stakes or formative assessment (for example, low quality assessment items with flaws being routinely provided to students "for their learning").

At this time our assessment program could have been argued to have been most aligned with the Ottawa consensus statement's PA principles 3, 4, 5, and 9 (see Figure 1, columns "Description" and "2012"). Pass/fail decisions (principle 3) were not made on a single data point with central university processes around further assessment ensuring that issues of measurement error were considered. Miller's pyramid [8] was used as a guiding framework for assessment design and a mix of methods of assessment (principle 4) across all years of the program ensured assessments focused on the knows, knows how, shows how, and does levels. While educational impact was not a driving force in assessment selection and design, it was recognised and being considered in subject evaluations. High-stakes decisions (principle 9) were made in a credible and transparent manner consistent with University Board of Examiners guidelines, but the approach was arguably more reductionist than holistic—requiring a pass on individual subjects, rather than considering achievements against core competencies. Principle 11—recurrent learning meetings with faculty—was more sporadically achieved thanks to the existing clinical school structure in years 2–4 of the program and through the longitudinal relationships between small groups of students and their clinical skills coaches and professional practice tutors in year 2.

2. From Assessment Program to Programmatic Assessment

2.1. Step 1: Formative Assessment and Feedback Innovations—2012 to 2018

In 2012 we began a program of work around the introduction of high-quality formative (low stakes) assessments in the form of progress tests coupled with automated and individualised post-test feedback reports. Initiated as part of AR's PhD, this research embedded approach was facilitated by the generous loan of a substantial bank of high-quality clinical knowledge items (MCQ format). Our progress tests were delivered using paper-based tests in conjunction with scannable answer sheets. Individual test items were tagged to a series of domains to facilitate the aggregation and analysis of results and the automated production of individualised post-test feedback reports. Our initial implementation involved delivery of four progress tests to volunteer year 2 students in 2013 and compared three relatively basic feedback variations. This early work [9] provided the momentum and evidence base to continue with the delivery of progress tests and automated feedback, involving the same student cohort as they progressed through the final 2 years of their degree. By the completion of this extended trial, incremental improvements to our feedback saw students receiving detailed reports that aggregated and interpreted their results across organ system, clinical rotation, and physical task (instead of a single mark indicating number of correct items on the test). Brief item summaries and response certainty data were also utilised to highlight knowledge misconceptions. High levels of student engagement, positive student perceptions of learning value (as gathered in routine subject evaluations) and successes in university grant schemes designed to improve formative feedback processes across the university, encouraged and supported us in continuing to refine our program of automated post-test feedback delivery.

During 2015, we extended our feedback delivery approach to include newly introduced Situational Judgement Tests (SJT)—which at that time were adopted in years 2 and 4 of the program as a scenario based summative assessment of knowledge of professionalism. The motivation behind the introduction of SJTs in our program has been described in detail elsewhere [10]. While the different format and focus of these tests presented challenges, our ability to adapt and apply our post-test feedback delivery systems and processes neatly demonstrated their general facility and scalability. In 2016, progress testing was formally adopted throughout years 2 to 4 of the medical program, and in the following year we further expanded our use of automated feedback to support the delivery of OSCEs to first- and second-year students. This latter expansion would not have been possible without our parallel program of work around technology supported assessment for interview and observation-based assessments (OSCE, standardised case-based discussion [SCBD] and multiple mini-interviews [MMI]—see [11]). As with our feedback innovations, this work was supported through a combination of departmental and university funding and was recognised by a faculty award for program innovation in 2017.

Summary of Outcomes

By the end of 2018 (i.e., within 5 years), we had developed capacity to provide automated and individualised feedback reports, highlighting domains of knowledge/skill strength and weakness, after a majority of MCQ, SJT and OSCE assessments for each of our student cohorts. Our key frustration from a pedagogical and technical perspective was the lag between test administration and delivery of feedback reports, which for assessments that involved the use of scannable mark sheets, depended on the timeliness and reliability of the data returned to us by the centralised university scanning service that we were obliged to use. In comparison, the electronic collection of OSCE data resulted in a much quicker turnaround of these results and feedback to students.

The design of our feedback reports was underpinned by a parallel program of research around how students interpreted, utilised, and applied the reports they received. While this work confirmed positive student engagement with high quality formative progress tests and feedback reports, as well as self-reported learning benefits, it also highlighted the negative impact of other aspects of the assessment program. Our own research into student learning behaviours [12] as well as routine program evaluation data consistently revealed that many of our students had a less than ideal relationship with assessment. Students often lamented (in hindsight) their missed opportunity for immersive clinical learning as they prioritised library study and high grades. Targeted interventions such as clinical school orientation to learning opportunities, assessment items which reward application of knowledge rather than factual recall, and routine progress testing (with detailed feedback to direct future learning) appeared to be functioning as temporary and isolated fixes rather than a comprehensive and integrated solution to this problem.

The developments associated with step 1 demonstrate progress most strongly aligned with the Ottawa consensus statement's first theme—continuous, meaningful feedback to create a dialogue with the learner, and Principles 1, 2, 11, and 12 (see Figure 1, column "Step 1"). While our shading in this figure hints at a somewhat limited impact, the first two of these principles—every (part of an) assessment is but a data-point and every data-point is optimised for learning by giving meaningful feedback to the learner—are arguably the foundation on which programmatic assessment is built. Our substantial effort to routinely provide students with direction for learning after written and clinical assessments provided both the basis and impetus for the innovations undertaken during the subsequent steps. Our research data [9,12] supported student use of feedback reports for self-analysis of strengths and weaknesses, and anecdotally, we learned that some students were using feedback reports to make learning plans in consultation with the academic staff at their clinical school (aligning with principles 11 and 12). The introduction of new assessment types—progress testing and situational judgement tests—during step 1 also speaks to theme 2—mixed methods of assessment across within the context of a continuum of stakes—and

within that, principles 4, 5, and 6. The feedback reports also provided an early framework for data collation, albeit it focused on directing and supporting learning rather than decision making (principle 8/theme 3).

2.2. Step 2: A Dedicated Assessment Team, Electronic Assessment Delivery, and a Pandemic—2019–2021

The next phase of assessment innovation was kickstarted by a departmental restructure at the end of 2018. In contrast to the existing flat departmental structure, the new organisational design was team-based, with each team led by a director (who was also a member of the departmental executive). A small assessment team was assembled, and the substantive members were a director (AR), a learning technology specialist (TJ), and an experienced former clinician with a focus on assessment quality. The formation of this team, holding overall responsibility for assessment within the medical program, provided a more cohesive oversight of assessment across all subjects and all years of the program compared to the existing subject-based and subject co-ordinator-led approach.

This course-wide view of assessment aligned with contemporary views of health professional assessment as demanding a more cohesive and program-wide approach. The 2018 consensus framework for good assessment [13] highlighted the importance of continuous approaches to assessment involving co-ordinated use of multiple assessment formats as part of a deliberate and cohesive assessment plan. The newly form team engaged in rich discussions with Australian and New Zealand colleagues as well as our honorary and visiting scholars around the potential for our assessment program to more accurately reflect a more contemporary view of assessment and reward and support longer term and mastery approaches to learning. At around the same time, a university-wide strategy on flexible academic programming (FlexAP) was initiated. This stratagem aimed to improve student orientation, engagement and learning through assessment. Our team, along with a small group of colleagues, developed a FlexAP project application referencing both the university's graduate attributes as well as expectations (around lifelong and mastery learning) of our health professional regulatory authorities and specialty colleges. This project, titled "Written assessment—from hurdling roadblocks to mastery", articulated a clear plan to move away from high-stakes, summative, end-of-year assessment towards a developmental and more continuous approach to assessment more in line with contemporary systems of assessment. It involved five components:

1. The introduction of a new assessment type—cumulative achievement testing (CAT) [14] in year 1 of the MD curriculum;
2. Expansion of the existing progress testing program, including a shift towards compulsory (but still formative/low-stakes) rather than optional assessments;
3. A transition from paper- to computer-based testing;
4. Development of new and enhanced forms of feedback with a focus on scaffolding of, and guidance for, learning;
5. Development of an integrated assessment and feedback literacy program.

Funding for the project enabled us to temporarily expand our team with the employment of a dedicated part-time assessment reform project officer at the end of 2019. The intention at that time was to leverage the new appointee's computer-based exam delivery expertise to implement a carefully staged rollout of computer-based exams using a bring-your-own-device (BYOD) model. Our plans included consultation visits to all clinical school sites to identify stakeholder perceived challenges and barriers to change and allow planning to support each of the sites in implementing these new assessment innovations.

We had barely begun when COVID arrived, throwing all our longer-term plans into disarray, and requiring a fast-tracked and flexible response to the rapidly and dramatically changing circumstances. Our modest 2020 plan for delivery of just four BYOD exams, for year 1 students in large-group, on-campus invigilated settings, was rapidly expanded to include remote delivery of secure and often high-stakes assessments across all four years of the medical program. This expansion necessitated a major financial outlay by our

medical school to provide access to a secure exam delivery platform and associated video invigilation for all our students. Our expanded BYOD delivery plan for 2020 eventually involved delivery of 21 large cohort remote exams, in addition to numerous practice tests, academic integrity quizzes, special exams, and supplementary exams throughout the year. On top of this, pandemic-related restrictions necessitated an immediate shift from paper-based to electronic capture of our workplace-based assessments (WBAs). At the time—and on the university's advice—we adopted an existing enterprise system for this purpose. However, it was poorly suited to WBA, and as a result, the opportunity to capture some workplace-based assessment and examiner feedback narratives was compromised. Finally, in response to major disruptions to learning, reduced clinical placement opportunities, variations between clinical sites, and our inability to run large-scale standardised clinical assessments, all subjects within the MD program were converted to pass/fail grading.

Ongoing challenges to clinical placements (Melbourne endured some of the longest and most stringent lockdown conditions worldwide during 2020 and 2021) drove the need to quickly identify and implement a more effective and efficient solution for capturing and managing workplace-based assessments. Late in 2020, our medical school supported a 3-year trial of an alternative workplace-based assessment platform. Complementarities and support for integration between this platform and its sister written assessment delivery platform led us to also switch from our existing BYOD exam delivery platform. Not surprisingly, the rapid implementation and adoption of two new and complex systems required a huge effort by stakeholders. During 2021, we utilised the workplace-based assessment platform to capture, aggregate, and report on all workplace-based and some written assessment tasks for third and final year students. Due to time constraints, we were unable to effectively utilise the student-facing assessment dashboard affordances of the new system. The new exam delivery platform was used to deliver all written assessments across all year levels, again involving the use of video invigilation when (frequently) in-person exams were not possible. Pass/fail grading was continued into 2021 (and subsequently 2022), and as a result, our state's medical schools were unable to generate rankings for their 2021 graduating cohorts resulting in a fundamental change to intern selection processes.

Summary of Outcomes

By the end of 2021, all written and clinical examinations as well as year 3 and 4 ward-based clinical assessments were delivered and/or captured electronically. Due to the prolonged lockdowns, and restrictions in conduct of learning activities within our clinical school locations, most written assessments throughout 2020 and 2021 were delivered BYOD in students' own homes, and so, contrary to our original plan, most staff had very little involvement with BYOD assessments and as such only a very small proportion of staff developed knowledge of our new exam delivery systems. As anticipated, electronic data collection across our examinations and assessments substantially improved our ability to deliver automated feedback on performance. Individualised feedback reports were routinely provided after all written and clinical examination formats except for SAQ exams (where cohort-level feedback is provided). Recent enhancements to our feedback development processes have improved the consistency of reports between assessment types and ensured that they are distributed as soon as practicable (i.e., subject to any board of examiners requirements) after each assessment. Routine evaluation of the MD program towards the end of 2021 revealed that the introduction of the new CAT assessment format (extended to year 2 of the program in 2021) was well received. Individualised feedback provided to students after key assessments continued to be highly rated. As expected, the evaluation confirmed the issues we experienced with the rollout of our initial WBA delivery platform in 2020 while also highlighting some of the challenges around rushed implementation of electronic systems. However, despite these challenges, the majority of students felt that the administration and invigilation of BYOD assessments was both appropriate and effective. Our regular stakeholder survey of staff and students highlighted that timely identification of students requiring learning assistance, improving

perceived relevance of assignment tasks, and increasing authenticity of standardised clinical assessments (such as OSCEs) were areas for future improvement. A significant proportion of respondents also felt the existing assessment program did not encourage students to disclose weaknesses and work with supervisors to improve their knowledge and skills.

The developments throughout our step 2 strengthen alignment across all three implementation themes and a number of principles (see Figure 1, column "Step 2"). The push towards electronic delivery of assessments substantially extending and enhancing our ability to aggregate, integrate and apply assessment data in support of both feedback provision and decision-making (speaking to principles across themes 1 and 3). While initially a temporary change in response to pandemic-related disruptions to learning, the introduction of pass/fail grading was justified based on establishing equitable decision making (theme 3) within the constraints of the pandemic and arguably further blurred the distinction between low- and high-stakes assessments. Consistent tagging of data from a wider range of assessment types and formats, along with the adoption of an effective workplace-based assessment platform (despite all features not being fully realised), also aligns with the third theme through principle 8—"assessment information is triangulated across data-points towards an appropriate framework" [4]. The challenges of the pandemic exerted a strong influence on our Board of Examiners decision making with careful consideration of the challenges and affordances of different clinical placements and increased consideration of assessment performance over time (principle 7) and introduced a more holistic focus on decision making (principle 9) compared to the existing assessment format and marks-based approach.

2.3. Step 3: Curriculum and Assessment Renewal—From 2022

The beginning of 2022 marked the launch and progressive rollout of our new Doctor of Medicine (MD) Program. Pre-requisites for entry into the new program have been removed to promote diversity in student admission. The course has been designed to improve vertical and horizontal integration of core knowledge streams, increase clinical connections, and offer increased flexibility and choice for students. Curriculum mapping software is being used to improve clarity and transparency of teaching and learning materials and their alignment with new course intended learning outcomes (CILOs).

While published guidelines and principles of good assessment had been used to inform assessment design within the program for many years, the curriculum redesign necessitated development of an explicit assessment strategy that was approved by the various departmental, school, and faculty committees and then circulated widely. This assessment strategy made explicit the plan for an integrated system of assessment consisting of authentic assessment formats, designed to support and direct student learning as well as assure patient safety and reflect the increased flexibility and improved vertical and horizontal integration of the new program. It also formally articulated a commitment to programmatic assessment with explicit alignment of all assessment criteria (across all assessment formats) against subject and course intended learning outcomes. Another essential component of our redesigned assessment strategy is the adoption of pass/fail grading (beyond being part of our pandemic response). Our expectation is that this change will improve both student's relationship with assessment [15] and their overall wellbeing [16]. We also anticipate that it will increase students' intrinsic motivation [17], encourage them to disclose areas of weakness [18,19], support students entering the program from diverse (non-bioscience) backgrounds [15], and promote collaboration and co-operation [18,19]. This plan has been approved by faculty and is currently progressing through central university review processes.

Feedback continues to be a core focus of our assessment program. The assessment and feedback processes outlined at the previous two steps mean that systems for routinely providing rich feedback on assessment are already in place and can be continually improved to more explicitly reflect the vertical and horizontal alignment of course content. We have already taken a large step in this direction through overt alignment of our assessment

criteria (across formats) against 12 subject intended learning outcomes (SILOs) within our core clinical subjects (these map directly to the 12 CILOs). While this represents a significant piece if work, it has not required a fundamental shift in our existing assessment culture or processes. Our existing two- and three-dimensional subject blueprints have simply been expanded to incorporate an extra SILO dimension.

This work has also facilitated representation of student achievement within the comprehensive student facing learning dashboard within our workplace-based assessment platform. While this work is very new (just commenced in 2022), we anticipate this dashboard will play a critical role in facilitating and normalising different rates of achievement of learning outcomes throughout the course (supporting students from different backgrounds, options for diversity, and interruptions in course progression). Following our very rushed and pandemic-impacted 2021 implementation of this system, we have recently expanded staff user access to these platforms so that subject co-ordinators and clinical school staff are able maintain effective oversight of the progress of their students, from the largest to the most dispersed of our clinical sites. While representation of student achievement by attribute is a substantial step towards decision making by attribute (a key feature of many programmatic assessment adoptions—[4,20]), full adoption would not be permissible within our current university-wide assessment policy.

Our university assessment policy also mandates that assessment design should promote and reward excellence. Our planned implementation of pass/fail grading aims to balance and accommodate assessment *of* and *for* learning though careful strategic design. We continue to provide numerical or other grades as a component of routine feedback after all standardised assessments. For other assessment types (e.g., written assignments and workplace-based assessments), performance against relevant assessment domains as well as an overall or global judgement is measured against a standard 5-point Likert rating scale (unsatisfactory, borderline, satisfactory, good, excellent). The three pass-levels (satisfactory and above) within this scale are designed to allow differentiation between levels of competency and to promote and support a focus on further development and mastery learning. Standards for overall judgement of performance, as well as against individual assessment criteria, are being explicitly linked to end of subject learning outcomes (which focus on end of year competencies) and include clear guidance for assessors.

Summary of Outcomes

The redesigned assessment strategy and documentation around pass/fail grading have been distributed widely and appear to be increasing stakeholder understanding and buy in. We have significantly expanded staff access and training in our new systems, and all subject co-ordinators and clinical school staff have access to their students' data within the assessment platform. The use of consistent rating scales and tagging by learning outcomes to differentiate performance by assessment domain in support of feedback for learning has been implemented across almost all our workplace-based assessments as well as a number of written assignments. With a few exceptions, this work will be extended to our remaining assessment types as the new MD program is rolled out. When viewed in relation to the principles and implementation themes described by the Ottawa consensus group (Figure 1, see column "Step 3"), these changes can be seen to have further strengthened alignment across nearly all areas.

We will continue to evaluate the potential for support of student learning through aggregation of assessment data by attribute, and once we have accumulated two or three years of data correlated in this way, we will be in a position to undertake comparative modelling of our current format-based approach to decision making versus decision making by attribute. It is this modelling that will ultimately determine whether we—subject to our university's agreement—take that final and critical step most aligned with principle 8 (and influencing principles 1,3, 6, and 9). It is important to note that traditional measures of assessment quality, particularly reliability, will not be sufficient to justify our approach. As part of this work, we will need to develop additional measures and procedures to build the

validity argument for our approach to PA [21] and work to translate these in relation to existing university assessment policy and procedures.

The collection, aggregation, and triangulation of previously disparate assessment data—both qualitative and quantitative—required to model decision making by attribute [22–24] will also help us determine the volume of assessment required (and hence feasibility) of decision making in this way. In the meantime, the permanent transition to pass/fail marking, the routine mapping and tagging of assessments by domain and learning outcomes, and the use of those tags to structure feedback for learning to students and represent achievement in the student learning dashboard, align with and support additional principles of programmatic assessment (principles 6 and 7).

While smaller programs have instigated longitudinal mentoring as part of their programmatic assessment innovations, our cohort size makes this aspect of programmatic assessment hugely challenging. Recent innovations in electronic data collection and collation have considerably enhanced our capacity to monitor, interpret, and respond to student performance and progress. The introduction of student and administrative dashboards in combination with regular and routine reporting of data should substantially improve oversight by subject coordinators and strengthen relationships with clinical schools and their teaching and administrative staff. Staff are better placed to identify students who need additional support, while students have ready access to data designed to guide and support their leaning—representing a potentially effective large cohort variant speaking to principles 10, 11, and 12. We will continue to monitor the impact of this through our ongoing quality assurance and evaluation processes

3. Discussion

Programmatic assessment has a strong theoretical/pedagogical underpinning [1], and while the evidential base is still developing, indications are that it has a net positive impact on learning [25,26]. However, at a practical level, implementation of programmatic assessment does bring several challenges. Common ones include having to collect and collate large volumes of often disparate assessment data, increases in staff and student assessment-related workloads, and the potential for students to feel they are under constant assessment-related scrutiny [20]. Many are exacerbated in large cohort settings such as ours.

The collection and collation of assessment data for our 1500 students has been an enormous piece of work for our medical program, and even with our dedicated and highly motivated team, this has taken a long time … 10 years and counting from our initial tentative step in 2012. That "step 1" initiative, to improve the quality and impact of our formative assessment and feedback—and to shift focus towards assessment for rather than of learning—seems relatively modest now but it laid strong foundations for what followed. The introduction of progress tests, (and later CATs), involving regularly spaced, high quality assessments with feedback oriented towards learning, highlighted the advantages of more continuous and vertically integrated program of assessment. Lessons learned around item blueprinting and tagging for feedback delivery were also instrumental in allowing us to continue to refine and improve this work during the pandemic impacted years of 2020 and 2021. They also provided a strong basis for our work around the collation of data by attribute, which remains a focus of our current efforts.

While our shift toward programmatic assessment has been incremental and (apart from a burst of activity forced on us by the pandemic) slow, the lengthy implementation timeframe has generally enabled us to better identify and solve adoption challenges as they have occurred. It has also enabled us to build in ongoing research and evaluation processes to capture and respond to stakeholder concerns and feedback and ensure that any changes that are made are both acceptable and sustainable. We strongly believe that our investment in these processes has been a key factor in ensuring stakeholder support and continued willingness of our medical school to make financial investment in our assessment innovations (particularly those innovations that were rushed during

the pandemic-impacted years of 2020 and 2021) and has likely been a major factor in our ongoing success in faculty and university grant schemes.

While our step 1 work was foundational in terms of subsequent development, the limitations of that approach in light of contemporary approaches to assessment are clearly evident. Prior to the establishment of a dedicated team having overall responsibility for assessment in the program, our work constituted innovation around the edges, rather than an integrated and cohesive system of assessment. Establishment of the team was a huge facilitator for the changes outlined in step 2 and beyond. The evidence base for best practice assessment design in the health professions is continually expanding. It is impractical to expect that disparate team members with diverse responsibility across a program can all keep up to date with contemporary expectations. Formation of a dedicated team with significant assessment expertise is essential for building consistency across a program.

In our opinion, a comprehensive embrace of electronic assessment systems and platforms is likewise virtually essential given the volume of assessment data that needs to be captured, aggregated, analysed, and applied to effectively support full programmatic assessment, particularly for large-cohort medical programs That is no small task however, and involves considerable costs, both human and financial (e.g., technology licensing, specialist staff, and user training). The COVID pandemic forced our hand in rapidly transitioning to the electronic delivery and capture of assessments and assessment data, and while that transition has been largely successful, the pace of change, particularly for a relatively small team with limited university level support, was brutal. Adoption failures are not uncommon (as evidenced by our initial attempt to introduce a workplace-based assessment platform) and implementations need to be carefully planned and evaluated, and well-resourced and supported.

As outlined in step 3, the curriculum renewal associated with the launch of our new medical program has been of fundamental importance in our journey towards PA. The refinement of 67 course attributes to form the 12 elements expressed in our course and subject intended learning outcomes has allowed us to make constructive alignment explicit and this seems critical if assessment is to fully realise its potential to direct and support learning. Clear documentation of the assessment strategy and particularly the rationale behind pass/fail grading has also improved stakeholder understanding and apparent buy in (rather than these key directions being somewhat limited to the dedicated assessment team).

Our rapid (and unavoidable) adoption of new assessment technologies highlights how the transition from traditional to programmatic assessment can precipitate clashes with existing university assessment policy and procedures. In our case, this manifested as an almost total lack of central support for BYOD exams and the invigilation of remote online exams during lockdowns, which were the mainstay of our written exams during 2020 and 2021. The current challenge is around course-wide implementation of pass/fail grading, and the upcoming one, should we continue on the path to full programmatic assessment, will be around decision-making by attribute. That is not to say that these sorts of differences cannot be resolved, but institutional policy is often slow and resistant to change. Assessment reform of this type and scale, particularly for 'flagship' curricula such as medicine, require careful planning, strong and sustained advocacy, and ongoing evidence of success over an extended period.

Our transition to programmatic assessment began with a focus on improving the learning potential of assessments in our medical program and the experience gained and lessons learned were foundational to future assessment innovation and change. Establishment of a dedicated assessment team and a significant investment in technology (inadvertently facilitated by a pandemic) were key enablers of our PA approach, yet without the significant curriculum renewal associated with a redesigned curriculum, it is unlikely we would have been able to fully embrace the cohesive and outcomes focused approach consistent with expectations for good systems of assessment [13] and typically seen in PA implementations [3,4]. Norcini and colleagues [13] highlight that systems of assessment must be

feasible and acceptable—given the work involved in initial implementations and the need for ongoing evaluation we are probably too early in our implementation to make definitive statements about these aspects of our assessment program. Our self-assessment against the 12 Ottawa PA principles demonstrates gradual strengthening of our alignment with all principles, with principles 1, 3, 6, and 10 less strongly aligned. Interestingly, while initially focused on improving assessment for learning, our self-assessment does not reveal a strong emphasis on one or more of the three implementation themes described by Torre et al. [4] (such as theme 1 "continuous and meaningful feedback") and instead demonstrates progress against principles clustered within each implementation theme at each step of our journey. Henneman and colleagues [3] highlight the importance of context and alignment of education and assessment within individual curricula. In our context (with large cohort sizes and university-wide assessment policy), regular progress review meetings and progress decisions based on aggregated qualitative and quantitative data (rather than assessment format) remain local challenges.

Author Contributions: Conceptualization, A.R. and T.J.; methodology, A.R. and T.J.; investigation, A.R. and T.J.; writing—original draft preparation, A.R.; writing—review and editing, A.R. and T.J.; All authors have read and agreed to the published version of the manuscript.

Funding: A.R.'s PhD research was supported by receipt of a University of Melbourne Australian Postgraduate Award Scholarship. Progress test items were supplied by the National Board of Medical Examiners (NBME) and were used in accordance with the terms of the agreement between the NBME and the University of Melbourne. Other elements of work described in this paper were enabled through funding support from The University of Melbourne Learning and Teaching Initiative Grant Program, The University of Melbourne Flexible Academic Programming (FlexAP) Project, and through the ongoing financial support of The Melbourne Medical School, University of Melbourne.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We acknowledge our colleagues in the Department of Medical Education and the Melbourne Medical School (including our honorary professors) for their involvement with, and support of, the changes described in this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Der Vleuten, C.P.; Schuwirth, L.W. Assessing professional competence: From methods to programmes. *Med. Educ.* **2005**, *39*, 309–317. [CrossRef] [PubMed]
2. Van Der Vleuten, C.P.; Schuwirth, L.W.; Driessen, E.W.; Dijkstra, J.; Tigelaar, D.; Baartman, L.K.; Van Tartwijk, J. A model for programmatic assessment fit for purpose. *Med. Teach.* **2012**, *34*, 205–214. [CrossRef] [PubMed]
3. Heeneman, S.; De Jong, L.H.; Dawson, L.J.; Wilkinson, T.J.; Ryan, A.; Tait, G.R.; Rice, N.; Torre, D.; Freeman, A.; Van Der Vleuten, C.P.M. Ottawa 2020 consensus statement for programmatic assessment—1. Agreement on the principles. *Med. Teach.* **2021**, *43*, 1139–1148. [PubMed]
4. Torre, D.; Rice, N.E.; Ryan, A.; Bok, H.; Dawson, L.J.; Bierer, B.; Wilkinson, T.J.; Tait, G.R.; Laughlin, T.; Veerapen, K.; et al. Ottawa 2020 consensus statements for programmatic assessmen—2. Implementation and practice. *Med. Teach.* **2021**, *43*, 1149–1160. [CrossRef] [PubMed]
5. Times Higher Education 2022 World University Rankings. Available online: https://www.timeshighereducation.com/world-university-rankings/2022/world-ranking# (accessed on 29 March 2022).
6. Norcini, J.; Anderson, B.; Bollela, V.; Burch, V.; Costa, M.J.; Duvivier, R.; Kent, A.; Perrott, V.; Roberts, T. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Med. Teach.* **2011**, *33*, 206–214. [CrossRef] [PubMed]
7. Schuwirth, L.W.T.; Van Der Vleuten, C.P.M. A history of assessment in medical education. *Adv. Health Sci. Educ. Theory Pract.* **2020**, *25*, 1045–1056. [CrossRef]
8. Miller, G.E. The assessment of clinical skills/competence/performance. *Acad. Med.* **1990**, *65*, S63–S67. [CrossRef]
9. Ryan, A.; McColl, G.J.; O'Brien, R.; Chiavaroli, N.; Judd, T.; Finch, S.; Swanson, D. Tensions in post-examination feedback: Information for learning versus potential for harm. *Med. Educ.* **2017**, *51*, 963–973. [CrossRef]
10. Goss, B.D.; Ryan, A.T.; Waring, J.; Judd, T.; Chiavaroli, N.G.; O'Brien, R.C.; Trumble, S.C.; McColl, G.J. Beyond selection: The use of situational judgement tests in the teaching and assessment of professionalism. *Acad. Med.* **2017**, *92*, 780–784. [CrossRef]

11. Judd, T.; Ryan, A.; Flynn, E.; McColl, G. If at first you don't succeed ... adoption of iPad marking for high-stakes assessments. *Perspect. Med. Educ.* **2017**, *6*, 356–361. [CrossRef]
12. Ryan, A.; Kulasegaram, K.; Mylopoulos, M. Challenges and Tensions in the Transition to Clinical Learning: Influence on Learning Behaviour. Presented at the Australian and New Zealand Association of Health Professional Educator Annual Conference, Adelaide, South Australia, Australia, 13 July 2017; Available online: https://anzahpe.org/resources/Documents/Conference/Past%20Conference%20documentation/Oral%20Proceedings%20-%20ANZAHPE%202017.pdf (accessed on 24 May 2022).
13. Norcini, J.; Anderson, M.B.; Bollela, V.; Burch, V.; Costa, M.J.; Duvivier, R.; Hays, R.; Mackay, M.F.P.; Roberts, T.; Swanson, D. 2018 Consensus framework for good assessment. *Med. Teach.* **2018**, *40*, 1102–1109. [CrossRef] [PubMed]
14. Swanson, D.B.; Holtzman, K.Z.; Butler, A.; Case Western Reserve University School of Medicine Cumulative Achievement Testing Study Group. Cumulative achievement testing: Progress testing in reverse. *Med. Teach.* **2010**, *32*, 516–520. [CrossRef] [PubMed]
15. White, C.B.; Fantone, J.C. Pass-fail grading: Laying the foundation for self-regulated learning. *Adv. Health Sci. Educ. Theory Pract.* **2010**, *15*, 469–477. [PubMed]
16. Bloodgood, R.A.; Short, J.G.; Jackson, J.M.; Martindale, J.R. A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. *Acad. Med.* **2009**, *84*, 655–662. [CrossRef] [PubMed]
17. Wilkinson, T. Pass/fail grading: Not everything that counts can be counted. *Med. Educ.* **2011**, *45*, 860–862. [CrossRef]
18. Reed, D.A.; Shanafelt, T.D.; Satele, D.W.; Power, D.V.; Eacker, A.; Harper, W.; Moutier, C.; Durning, S.; Massie, F.S., Jr.; Thomas, M.R.; et al. Relationship of pass/fail grading and curriculum structure with well-being among preclinical medical students: A multi-institutional study. *Acad. Med.* **2011**, *86*, 1367–1373. [CrossRef] [PubMed]
19. Spring, L.; Robillard, D.; Gehlbach, L.; Moore Simas, T.A. Impact of pass/fail grading on medical students' well-being and academic outcomes. *Med. Educ.* **2011**, *45*, 867–877. [CrossRef]
20. Schut, S.; Maggio, L.A.; Heeneman, S.; van Tartwijk, J.; van der Vleuten, C.; Driessen, E. Where the rubber meets the road—An integrative review of programmatic assessment in health care professions education. *Perspect. Med. Educ.* **2021**, *10*, 6–13. [CrossRef]
21. Schuwirth, L.W.; Van Der Vleuten, C.P. Programmatic assessment and Kane's validity perspective. *Med. Educ.* **2012**, *46*, 38–48. [CrossRef]
22. Pearce, J.; Prideaux, D. When I say ... programmatic assessment in postgraduate medical education. *Med. Educ.* **2019**, *53*, 1074–1076. [CrossRef]
23. Tweed, M.; Wilkinson, T. Student progress decision-making in programmatic assessment: Can we extrapolate from clinical decision-making and jury decision-making? *BMC Med. Educ.* **2019**, *19*, 176. [CrossRef] [PubMed]
24. Wilkinson, T.J.; Tweed, M.J. Deconstructing programmatic assessment. *Adv. Med. Educ. Pract.* **2018**, *9*, 191–197. [CrossRef] [PubMed]
25. Bierer, S.B.; Dannefer, E.F.; Tetzlaff, J.E. Time to loosen the apron strings: Cohort-based evaluation of a learner-driven remediation model at one medical school. *J. Gen. Intern. Med.* **2015**, *30*, 1339–1343. [CrossRef] [PubMed]
26. Heeneman, S.; Oudkerk Pool, A.; Schuwirth, L.W.; Van Der Vleuten, C.P.; Driessen, E.W. The impact of programmatic assessment on student learning: Theory versus practice. *Med. Educ.* **2015**, *49*, 487–498. [CrossRef]