Research Article https://doi.org/10.12973/eu-jer.12.2.649

European Journal of Educational Research
Volume 12, Issue 2, 649 - 662.
ISSN: 2165-8714
http://www.eu-jer.com/

Development of a Self-Evaluation Instrument with Programmatic Assessment Components for Undergraduate Medical Students

Dina Qurratu Ainin* Universitas Islam Al-Azhar, INDONESIA

Yoyo Suhoyo Universitas Gadjah Mada, INDONESIA

Artha Budi Susila Duarsa Universitas Islam Al-Azhar, INDONESIA

Mora Claramita Universitas Gadjah Mada, INDONESIA

Received: July 19, 2022 ▪ Revised: October 25, 2022 ▪ Accepted: February 17, 2023

Abstract: This study aimed to develop and test a student self-assessment instrument based on the programmatic assessment (PA) components. We applied a series of psychometric research methods by (a) conducting a literature study to find PA constructs, (b) developing the students' self-questionnaires, (c) ensuring content validity, (d) testing face validity, and (e) conducting reliability tests that involve medical students, medical teachers, medical educationalist, and an international PA expert. Face validity (readability test) was conducted with 30 medical students from an Indonesian university who were in their last year of pre-clinical education and had average scores above or equal to their classmates. The confirmatory factor analysis (CFA) was used to report the instruments' validity and reliability. The final instrument was tested on 121 medical students with excellent GPAs from another medical school with a middle-level accreditation. The PA consists of five components: 'learning activities', 'assessment activities', 'supporting activities', 'intermediate evaluations', and 'final evaluations'. These components are conveyed through 41 relevant statements with a four-point Likert scale and three yes/no statements. According to the respondents, there was a lack of 'supporting activities' and 'intermediate evaluation' components in the PA in their universities. This study has developed and tested a five-component evaluation instrument based on medical students' perceptions regarding PA implementation.

Keywords: *Instrument development, medical education, programmatic assessment.*

To cite this article: Ainin, D. Q., Suhoyo, Y., Duarsa, A. B. S., & Claramita, M. (2023). Development of a self-evaluation instrument with programmatic assessment components for undergraduate medical students. *European Journal of Educational Research*, *12*(2), 649-662. https://doi.org/10.12973/eu-jer.12.2.649

Introduction

There has been a shift in paradigm from 'assessment of learning to 'assessment for learning' in medical education. The assessment program should encourage students to conduct further learning instead of judging what has been learned, as in the typical focus of traditional assessment systems (Amin et al., 2006; Martinez & Lipson, 1989; Shumway & Harden, 2003; Wass et al., 2001). The Programmatic Assessment (PA) is a new approach to assessing students' learning. It gathers routine information about students' progress and then analyzes it to provide maximum feedback to students and their mentors to make evidence-based decisions. Designing a comprehensive assessment program is recommended to ensure the excellent quality of a student assessment system (van der Vleuten et al., 2018; Wilkinson & Tweed, 2018).

The PA model initially formulated by van der Vleuten et al. (2010) connected the components of teaching, assessments, supportive learning activities, intermediate evaluation, and final evaluation (Bok et al., 2018). The definitions of each component are elaborated in Table 1. The PA as a progressive assessment aligns with student-centered learning (SCL). SCL is a learning concept that emphasizes the basic concepts of active learning from the students themselves. The PA also emphasizes feedback for the student's learning process and increases motivation (van der Vleuten et al., 2010, 2012, 2018). SCL may give the impression that the lecturer has a minimal role because students are required to be more active in searching for information to support the achievement of their competencies (Attard et al., 2010). However, the SCL concept is equally essential in the PA since it is aimed at giving constructive feedback to the students and assisting them to reflect on their learning is required with the PA approach (Attard et al., 2010; van der Vleuten et al., 2012). Feedback in the PA is expected to be given to each student regarding their strengths and weaknesses (Torre et al., 2020; van der Vleuten et al., 2012). This approach confirms that in addition to teaching and assessing students, lecturers or instructors play an essential role in providing constructive feedback based on the student's learning progress results. Additionally,

* **Corresponding author:**
Dina Qurratu Ainin, Universitas Islam Al-Azhar, Mataram, Indonesia. ✉ dinaqurratu86@gmail.com

lecturers or instructors are obliged to provide input on student follow-up plans and mentor them (Perry et al., 2018; Torre et al., 2020; van der Vleuten et al., 2012, 2018).

Furthermore, the PA is in line with competency-based education. Complex skills such as communication skills, collaboration, and professionalism cannot only be assessed via multiple-choice questions. They should also be done through continuous observations, reflections, and constructive feedback (Heeneman et al., 2021). Moreover, students' character development could be assisted by using the PA that emphasizes student-teacher dialogue throughout the medical curriculum (Heeneman et al., 2015, 2021; van der Vleuten et al., 2012, 2018).

The PA approach is a globally new and rapidly emerging paradigm in medical education (Schut et al., 2018). This approach has been used mostly in western contexts where there are more partnership relationships and mutual dialogue between teachers and students (Driessen et al., 2012; Lacasse et al., 2022; Schut et al., 2018). However, in conditions where student-teacher dialogue is limited, and a summative assessment is routinely used to judge the pass/fail of future doctors, we need a more appropriate way to promote the PA. Such a context could be seen in countries or cultural backgrounds with a wide power distance dimension (Hofstede, 2011). This type of cultural dimension is spread across many regions, such as Asia, Africa, the Mediterranean, and Latin America, including Indonesia, as the setting of this study (part of Southeast Asia). The challenge in this culture is the high acceptance towards power distance between people, including between doctor-patient and student-teacher, with norms of silence traditionally kept to maintain social harmony and avoid conflict.

For cultures with a wide power distance that prefers indirect communication, one of the more subtle ways of endorsing the PA is to evaluate the medical schools' assessment system based on student perceptions. This evaluative method will be more acceptable than if a researcher challenges the teachers in a teacher-training program concerning a new paradigm of student assessment. By using the students' evaluation results and processing the data through a rigorous instrument, teachers in a hierarchical culture may be more supportive of receiving information and input regarding the current student-assessment evaluation in their medical schools. Therefore, this study is important for promoting a more formative principle of learning in the global context, considering that many regions in the world maintain a more hierarchical relationship in their society, with its main consequence being fewer dialogues between students and teachers.

However, there is a lack of specific explanation related to the components of the PA that can be used as indicators of the program's implementation in medical schools. These indicators are needed for all stakeholders (including students) conducting self-evaluation to determine whether the application of the assessment system in their medical schools aligns with the recommended literature regarding the PA. Accordingly, this study aimed to develop an instrument that could be used to assess the PA application steps in medical schools based on students' perceptions. We selected the students' perception basis because students are the subjects most impacted by the learning and assessment system. Thus, the student's self-assessment results could provide insights into the actual conditions that need to be reconsidered by medical teachers within the institution.

Literature Review

The questionnaire produced in this study can be used for students in the pre-clinical and clinical stages, with a slight adjustment and modification to its contents according to the activity's objective (van der Vleuten et al., 2012). The most important consideration is that the five components in the PA model are clearly illustrated.

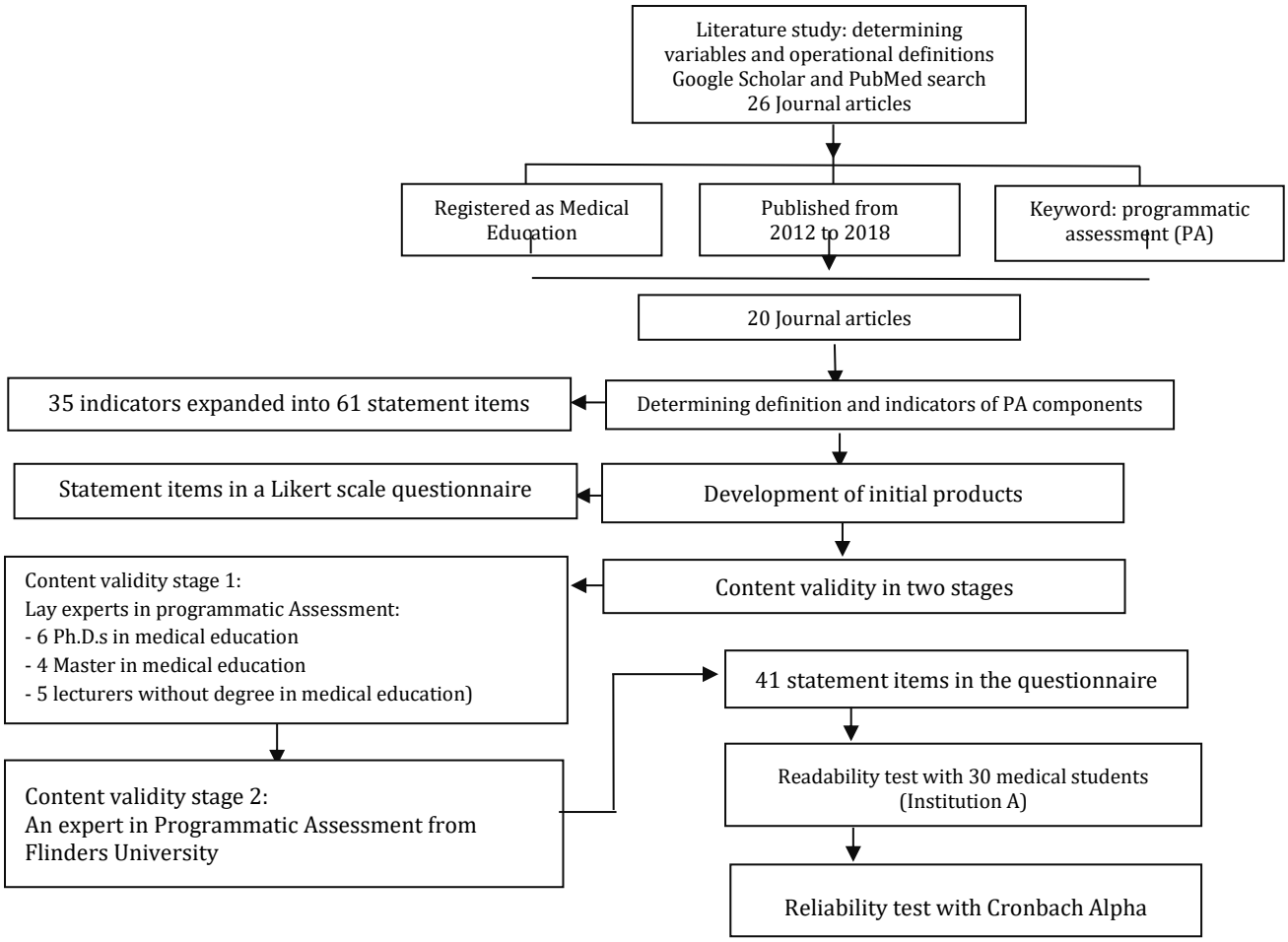

Figure 1. Diagram of Procedures of Instrument Development for Evaluating the Implementation of

Programmatic Assessment

This instrument will make it easier for medical schools to conduct a self-assessment on the quality of their assessment system based on their student's self-evaluation. The feedback generated from evaluations using this instrument will positively impact the quality of instructors and learning facilities as part of the development of the medical faculty, especially where Student-Centered Learning (SCL) is widely introduced. The results of the evaluations using this instrument can also be a reference for directing and developing the SCL system (Attard et al., 2010; Torre et al., 2020).

Many strategies could be used to introduce a novel program in medical education. In this study, we developed an instrument for the following reasons: the setting of this study, the medical students, and the lay experts who validated the instrument have come from a non-western background: Indonesia. In this country, the medical school system has implemented summative assessments for decades (McComas, 2014). There have been a lack of constructive feedback and reflection on the learning process. Therefore, the participants in this study may not be familiar with the PA principles. However, the items developed in this study underwent procedures that validated the instrument into the final stage. These steps involved the students and lay experts or local leaders in medical education to ensure they could gain an understanding of the PA concept. When the teachers and leaders know that the instrument has undergone psychometric procedures, it will increase their trust in it. Ultimately, this approach will enable the principle of PA to be better understood and applied in the appropriate settings.

Methodology

This research used a psychometric methodology to develop a valid instrument for evaluating the implementation of the PA in undergraduate education. The instrument is related to the five components of the PA and uses student perceptions as a source (Kerlinger, 1986; Netemeyer et al., 2003; van der Vleuten et al., 2010). The steps in this study included recommendations from Davis, 1996. There are four main steps: (a) concept identification, (b) item development, (c) validation, and (d) internal consistency. Figure 1 illustrates the steps taken in this study.

This study was conducted at the Faculty of Medicine, Public Health and Nursing of the University Gadjah Mada, Yogyakarta, Indonesia. A total of 30 pre-clinical undergraduate students in their final year were selected. The inclusion criteria for this study's subjects were students with a GPA of at least 2.5. The subjects were divided into 15 male and 15 female students aged around 21 to 22 years old. The 30 students were divided into three data-collection phases over three days. Ten students were selected for each phase. Improvements were made in stages based on feedback from each trial stage.

Study Instrument Procedure

Instruments related to PA evaluation are new instruments that need to go through a process of developing instruments that are good in content, as well as considered valid and reliable. The instrument was created according to concept identification (developing the definition of the domain contents or constructs of the PA), item constructions (generating and rating measurement items, face validity assessment), validity and reliability test, and testing the instrument on medical students.

Concept Identification: Developing the Definition of the Domain Contents or Constructs of the PA

This step was done by conducting a literature review of the PA components to determine which will be selected according to items that fit the theme of this study. The literature study is the first step for identifying indicators for each component of the PA. Researchers collected and read 26 papers on the PA model proposed by van der Vleuten et al. (2012). A more specific literature search was conducted using the keyword "Programmatic Assessment" for research published between 2012 and 2018 on online medical education journal sources from Google Scholar and PubMed. A total of 20 articles were obtained. Based on the literature review of these 20 articles, five components must be present in the implementation of the PA. The next step was to define the five components and determine the implementation indicators for all components. From this step, 35 indicators were produced and converted into 61 questionnaire statement items.

Item Constructions: Generating and Rating Measurement Items

After collecting the selected items, we assessed content validity. In phase 1, we consulted ten medical educationalists from Indonesia and five teachers from an assessment division who are not in medical education as lay experts in the PA. The corrected items were then validated in the second stage by an international expert in the PA from Flinders University, Australia. Using the Aiken formula (Wass et al., 2001), content validity was estimated based on the calculation results derived from some "n" experts on an item in terms of the extent to which the item represented the measured construct. Aiken's coefficient values (V) typically range between 0-1. They were calculated for each statement item in the questionnaire. The validity test results were used to determine which items from the questionnaire would be retained, revised, or even deleted based on expert opinions and inputs.

Face Validity Assessment

This test was used to assess the readers' understanding of all items and statements in the questionnaire that experts in the previous stage had revised. This readability test was conducted on a sample of final-year medical school students in their pre-clinical education from the highest level of medical school in Indonesia (institution of the second and fourth authors in this study). A total of 30 subjects were divided into three stages of data collection for three consecutive days, and each stage was corrected by 10 sample students. Improvements were made gradually based on input in each step of the readability test until it was considered easy to understand (van der Vleuten et al., 2018).

Validity and Reliability Test

The final stage of the preparation of this instrument was to test its validity and reliability using Confirmatory Factor Analysis (CFA) in Structural Equation Modeling (SEM). This test produces better outcomes when evaluating the reliability and validity of an instrument (Said et al., 2011).

Testing the Instrument on Medical Students

We tested the final items on 121 students of undergraduate medical education in one faculty of medicine in Indonesia (the first author's institution). The sample ranged from year 1 to 4 students with a minimum GPA of 3 out of 4. We selected students with satisfactory academic results to ensure that the opinions collected were from accountable students.

Theoretically, the researchers must complete several steps in the development and preparation of the instrument, which starts with identifying the construct of the research variable, including the definition of the construct and content domain, which is a synthesis of the theories that have been discussed and analyzed and whose presentation has been described in the theoretical review or review of the literature conducted (Netemeyer et al., 2003). Additionally, validity was tested in this study because it is viewed as a structured argument that gathers evidence from various sources to support or reject suggested interpretations of instrument values. It has become a recommendation for creating valid instruments (Cook & Beckman, 2006).

Results

Results and Analysis From the Study's Instrument

There were 20 journal articles collected to identify and describe the five components of the PA based on the van der Vleuten's model. The components consist of 'defined learning activities,' 'assessment activities,' 'supporting activities,' 'intermediate evaluation,' and 'final evaluation.'

'Learning activities' consist of various learning methods to achieve student competencies. Next, the 'assessment activities' component acts as a medium to illustrate the extent to which student competencies are achieved and need to be encouraged through continuous feedback and reflection. 'Supporting activities' help the student's learning process by providing opportunities to reflect on information obtained from learning activities, assessment activities, and feedback received through coaching or mentoring. The 'mid-term evaluation' acts as an intermediate evaluation before the final evaluation so that students and teachers can reassess the previous learning process and plan for better learning progress. The 'final evaluation' is conducted at the end of the program. This evaluation can be used to make high-stakes decisions on student learning outcomes and is based on comprehensive information and many data points to ensure the quality of the decisions made. Table 1 explains the definition of each component of the PA. The development activities of the PA evaluation instrument were done in chronological order to finalize the final set of instruments.

Table 1. PA Components Based on Literature Review in This Study

No	Component	Characteristics
1.	Learning activity	• Consists of several learning tasks that can guide learning and, if properly arranged, can provide a coherent program or curriculum that is structured according to instructional design principles. • Designed and harmonized with the competency framework to be achieved and adapted to learning activities for individual learning needs. • A series of planned learning activities and tasks to optimize the information collection on student learning development.
2.	Assessment activity	• Evaluates artifacts generated from learning tasks to support learning according to the principles of assessment fully, drives learning, and provides feedback on student performance rich in quantitative or qualitative information. • Ensures students receive feedback in various areas of competence with an emphasis on qualitative feedback, which will contribute to the final assessment. • A form of activities designed to obtain comprehensive information regarding student performance progress and use information from feedback received for self-direct learning and learning from assessment results. • Consists of a variety of assessment methods. • Assures that every low-stakes assessment provides meaningful feedback for learning.
3.	Supporting activity	• Supports student learning by allowing students to reflect on information obtained from learning activities, assessment activities, and the feedback received through coaching or mentoring that will be used to plan assignments and new learning goals. • Optimizes providing information and feedback from teachers and peers about their learning progress in developing competencies to be incorporated into the portfolio as a basis for reflection with mentor guidance. • Assists students in supporting their learning through feedback received in various opportunities to reflect on their performance and learning progress through a portfolio and supported by mentoring.
4.	Intermediate Evaluation	• The assessment of student learning progress in the middle or between learning processes by gathering all information from learning artifacts, assessment information, and information from supporting activities to make decisions and recommendations for planning further learning activities. • Intermediary evaluation before the final evaluation in the program assessment to decide whether a student is allowed to continue their work based on information in the portfolio. • The logical longitudinal development of the learner through learning tasks, appropriate feedback, and (supported) self-direction is of key importance
5.	Final evaluation	• Performed at the end of the program to make high-stake decisions with great consequences related to student learning outcomes. The evaluation is based on comprehensive information and many data points to ensure the quality of decisions made. • The high-stake decision is taken by an independent assessment committee of the medical faculty through several actions to ensure the quality of decisions by recommending qualitative information.

Based on the definitions of each component of the PA in Table 1, we established 35 indicators from which a total of 61 statements were then distributed for validation by experts in the initial content validity test based on the Aiken formula. The results of the revised items were then translated into English for validation by an assessment expert from Flinders University, Australia. After expert validation, the final results were 41 four-scale Likert items and 3 yes/no items. In quantitative calculations, all statements were statistically valid, as indicated by the results of the normal Aiken's coefficient V values, ranging from 0-1 for all statements. This result indicates that all items in the instrument have

conformed to the indicators based on expert judgment. However, qualitatively, improvements were made for certain items based on the suggestions and comments from experts.

Statements that were changed were those that needed grammatical improvements and adjustments to the meaning of the sentences in the items, while deleted items were those that were redundant and repetitive. Language selection was a major obstacle for us in adjusting statements with indicators to make it easier for respondents to understand. Nevertheless, we strived to perfect the items based on as much expert input as possible.

One suggestion for improvement was to provide instructions to students before the statement for the mid-term evaluation component, to state whether or not the mid-term evaluation is conducted at the institution. In addition, several choices of terms were adjusted for item improvement. For example, the word "difficulty" was replaced by "weakness" to focus more on the relationship in the preceding item that asks about student learning strengths. Likewise, the statement "a thorough evaluation in the middle of the learning process helps me to make a plan of learning strategies to overcome my learning difficulties" was replaced with the sentence "a thorough evaluation in the middle of the learning process helps me to make plans of what I must do next." Improvements to these items are related to determining the relationship between the meaning of the assessment and students' perceptions that must be explained in a coherent statement. The statements were adjusted to ask about the strengths and weaknesses of student learning and what efforts are their following learning strategies. This approach is consistent with past literature and is why we maintained these items.

An example of input from the international assessment experts was removing any redundant items. For example, "learning activities that I undertake based on the topics I want to master" is in line with what competencies are expected to be mastered by students in the previous section of the questionnaire. Therefore, only one item was kept. Next, for statements such as "I get feedback from all assessment activities," the word "often" was added to ensure students can more easily choose answers on a Likert scale. Other inputs were related to the suitability of the conceptual meaning of the indicators produced. Overall, based on expert input, 41 items were selected for the final validation, then tested for face validity and reliability.

Face validity test results focused on confirming the statement's meaning in the item. There were no changes in the number of items in this readability test. Students also provided input regarding sentence structure and the selection of words for ease of understanding. One example of an item that was improved is "I got a variety of assignments that can help me prepare for assessment activities," in which the "variety of assignments" was replaced by "variety of learning activities." Another example of improvement is the statement, "My academic supervisor gives me continuous feedback," in which the word "continually" was replaced by "continuous."

Results and Analysis From Testing of the Instrument

The test results of the questionnaire items are presented in Table 2. Overall, the students rated the items 3.34 out of 4, which is defined as satisfactory. Components one and two of the PA (learning and assessment activities) were rated satisfactory. However, for supportive activities, only half of the students agreed that there was a mechanism of continuous feedback and reflection in their medical school assessment system with a portfolio. Students who agreed that there are currently supportive activities only in their institution rated this component 2.89, indicating their dissatisfaction. This result suggests that continuous reflection and constructive feedback or mentorship should lie at the heart of the PA system. Moreover, only about half of the students sampled agreed that there was intermediate evaluation in their medical school. Additionally, among the half who agreed they were mostly satisfied with the process. For the final evaluation, most of the students agreed that they understood the final evaluation process (Table 2).

Table 2. The Final Instrument of Students' Self-Assessment Based on the PA Components in This Study, and the Final Results of Students' Self-Perception at a Medical School in Indonesia

Component of PA	Items	Mean	Sd	Min	Max
Overall component of PA (N:121)	Overall Items	3.34	2.58	1.00	4.00
Student with a minimum GPA of 3 from 4-point scale					
Student year 4: N = 1; Male = 1; Female =0		3.35	3.31	1.00	4.00
Student Year 3: N = 40; Male = 10; Female = 30		3.31	0.21	1.00	4.00

Table 2. Continued

Component of PA	Items	Mean	Sd	Min	Max
Student Year 2: N = 41; Male = 11; Female = 30		3.38	3.01	1.00	4.00
Student Year 1: N =49 Male = 37 Female = 11		3.32	3.78	1.00	4.00
Overall Component of Learning Activities (N 121 or 100%)		3.38	0.52	2.20	4.00
	Learning activities				
	Learning activities help the students to achieve the expected competencies.	3.53	0.50	3.00	4.00
	Students can plan learning activities based on the results of the received assessments.	3.31	0.59	1.00	4.00
Learning activities	Students have several variations of learning assignments that can help them prepare for assessment activities.	3.29	0.51	2.00	4.00
	The material or topic presented in the learning activities is part of the material being evaluated (assessed).	3.41	0.49	3.00	4.00
	Various learning assignments contribute to the final assessment.	3.36	0.50	2.00	4.00
Overall Component of Assessment activities (N 121 or 100%)	Assessment activities	3.14	0.65	1.43	4.00
	All assessment methods are in alignment with student learning activities.	3.31	0.48	2.00	4.00
	Students receive feedback (written or oral) from all assessment activities.	3.03	0.73	2.00	4.00
	Academic advisors (mentors) provide students with ongoing feedback.	3.29	0.64	2.00	4.00
	Students receive feedback (written or oral) on the positive efforts that they have to maintain.	3.15	0.65	2.00	4.00
	Students receive feedback (written or oral) on specific matters that they need to improve on for better learning activities (in the academic field).	3.02	0.72	1.00	4.00
	Students receive feedback that makes it easier to plan their next learning activities.	3.07	0.66	1.00	4.00
	The implemented assessment system helps students to identify their learning weaknesses.	3.14	0.63	1.00	4.00
	The implemented assessment system helps students to formulate a more specific learning plan to make	3.15	0.60	2.00	4.00
Assessment activities	improvements for the next assessment.				
	The implemented assignment system helps students to identify their strengths (tasks that are well achieved).	3.07	0.66	1.00	4.00
	The implemented assignment system helps students determine how to maintain good learning achievement.	3.24	0.55	2.00	4.00
	Students trust the assessment system implemented in their medical schools because it is based on a collection of information from the various achievements of learning activities.	3.23	0.56	1.00	4.00
	After receiving the results of the assessment, students are assisted by academic advisors (mentors) to formulate learning strategies to improve their learning effectiveness.	3.03	0.80	1.00	4.00
	Students can plan or choose remediation options with support from academic advisors (mentors).	3.21	0.62	1.00	4.00
	Through the assessment results, students received a recommendation from an academic advisor (mentors) regarding future learning plans that are the most appropriate for them.	3.01	0.73	1.00	4.00

Table 2. Continued

Component of PA	Items	Mean	Sd	Min	Max
Overall Component of Supporting activities N answer YES: 64 (52.89%)	A portfolio (compilation of learning documents, study results, student reflection, and feedback from teachers and mentors) is a part of the student-assessment system. Overall Mean of students who answer YES	2.89	0.71	1.33	4.00
Supporting Activities	Students write a portfolio containing reflections on the outcomes of their learning activities.	2.71	0.72	1.00	4.00
	Students write a portfolio based on the feedback they receive.	2.78	0.66	1.00	4.00
	Students know how to write good reflections in a portfolio.	2.65	0.76	1.00	4.00
	Academic advisors (mentors) guide student to write reflections in a portfolio.	2.71	0.82	1.00	4.00
	Students write reflections on the learning aspects they still need to improve.	2.76	0.73	1.00	4.00
	Students write reflections on the achievements that they have made in the learning process.	2.74	0.74	1.00	4.00
	Based on the feedback the students receive, they can plan new learning targets.	3.18	0.64	2.00	4.00
	Students are monitored by academic advisors (mentors) so that all the plans in their reflections are done well.	3.07	0.67	2.00	4.00
	The academic advisors (mentors) and the students have several scheduled meetings.	3.20	0.63	2.00	4.00
	In addition to academic advisors (mentors), students often act as peer mentors.	2.93	0.79	1.00	4.00
	Reflection activities help the students to improve their learning ability.	3.20	0.63	2.00	4.00
	The reflective portfolio I wrote is evaluated by academic advisors (mentors) to generate regular feedback (continuous feedback).	2.77	0.73	1.00	4.00
Overall Component of Intermediate evaluation N answer YES: 82 (67.77%)	There is an evaluation in the middle of the learning process (it can be in the middle of the block/middle of the semester) as a reference to assess my learning progress. Overall Mean of students who answer YES	3.27	0.53	2.00	4.00
Intermediate evaluation	Students use all the information from the feedback they receive to prepare for the evaluation in the middle of the learning process.	3.31	0.54	2.00	4.00
	All information from students' reflection activities is used to see their learning progress.	3.24	0.58	2.00	4.00
	Through the evaluation results in the middle of the learning process, students can identify the competencies that must be mastered to prepare for the final evaluation.	3.28	0.52	2.00	4.00
	The middle evaluation allows the students to identify the strengths in their learning that they need to maintain.	3.26	0.51	2.00	4.00
	The evaluation in the middle of the learning process enabled the students to identify their learning difficulties.	3.30	0.53	2.00	4.00
	The evaluation in the middle of the learning process allowed students to make a follow-up plan that they had to do going forward.	3.28	0.52	2.00	4.00
	The evaluation in the middle of the learning process helps students to predict the outcome of the next learning achievement.	3.24	0.53	2.00	4.00
Overall Component of Final evaluation N answer YES: 115 (95, 04%)	Students know there will be a final evaluation after going through learning activities, assessment activities, supporting activities, and intermediate evaluation. Overall Mean of students who answer YES	3.37	0.51	2.33	4.00

Table 2. Continued

Component of PA	Items	Mean	Sd	Min	Max
Final evaluation	The final outcome of students' assessment is based upon all the assessment information that they have received before.	3.37	0.48	3.00	4.00
	All the assessment data that students receive from the entire learning process contributes to the final assessment.	3.41	0.51	2.00	4.00
	Students know that decisions regarding passing or failing are determined in the final evaluation based on all assessment data.	3.34	0.54	2.00	4.00

Scores based on a 4-point Likert scale: 1: Totally disagree, 2: Disagree, 3: Agree, 4: Totally agree (Ainin, 2018).

Validity and Reliability Test Results

The results of the validity can be seen from the item correlation test. An item can be used as a measuring tool if it has a total-item correlation >0.20. It was found that all items met these requirements so that they were declared valid. Meanwhile, the reliability results are based on Cronbach's alpha >0.6. All components of PA obtained Cronbach's alpha >0.6, so that the instruments could be declared reliable (Table 3).

Table 3. Validity and Reliability Test Results

Component of PA	Items	Item – Test Correlation	Cronbach' Alpha
Learning activities (LA)	LA1	0.83	0.81
	LA2	0.66	
	LA3	0.78	
	LA4	0.76	
	LA5	0.76	
Assessment Activities (AA)	AA1	0.63	0.95
	AA2	0.62	
	AA3	0.81	
	AA4	0.58	
	AA5	0.72	
	AA6	0.78	
	AA7	0.81	
	AA8	0.71	
	AA9	0.70	
	AA10	0.63	
	AA11	0.72	
	AA12	0.73	
	AA13	0.74	
	AA14	0.76	
	AA15	0.81	
	AA16	0.71	
	AA17	0.73	
	AA18	0.69	
	AA19	0.79	
Supporting Activities (SA)	SA1	0.95	0.99
	SA2	0.96	
	SA3	0.97	
	SA4	0.91	
	SA5	0.96	
	SA6	0.97	
	SA7	0.97	
	SA8	0.96	
	SA9	0.96	
	SA10	0.94	
Intermediate Evaluation (IE)	IE1	0.97	0.97
	IE2	0.96	
	IE3	0.98	
	IE4	0.94	
Final Evaluation (FE)	FE1	0.96	0.96
	FE2	0.97	
	FE3	0.96	

Confirmatory Factor Analysis (CFA)

The CFA used to test whether the indicators that have been grouped based on a certain construct are consistent in the construct or not. In other words, in CFA the researchers want to test whether the data fit the model that has been formed before. The CFA results of five components of PA showed that the data has a good fit.

1. Learning Activities (LA)

The results of the CFA test showed that the items LA1 (b= 4.34; p<0.001), LA2 (b= 3.77; p<0.001), LA3 (b= 4.21; p<0.001), LA4 (b= 4.10; p<0.001), and LA5 (b = 4.08; p<0.001), which provide a significant factor loading or contribution to the learning activities instrument and are statistically significant.

2. Assessment Activities (AA)

The results of the CFA test showed that the items AA1 (b= 3.94; p<0.001), AA2 (b= 3.51; p<0.001), AA3 (b= 3.25; p<0.001), AA4 (b= 3.75; p<0.001) , AA5 (b= 3.34; p<0.001), AA6 (b= 3.60; p<0.001), AA7 (b= 3.38; p<0.001), AA8 (b= 2.64; p<0.001), AA9 (b= 2.34 ; p<0.001), AA10 (b= 2.19; p<0.001), AA11 (b= 2.70; p<0.001), AA12 (b= 3.42; p<0.001), AA13 (b= 3.21; p<0.001), AA14 (b= 3.30; p<0.001), AA15 (b= 3.43; p<0.001), AA16 (b= 3.60; p<0.001), AA17 (b= 3.47; p<0.001), AA18 (b= 3.45; p<0.001), and AA19 (b= 3.25; p<0.001), which provide a significant factor loading or contribution to the instrument assessment activities and are statistically significant.

4. Supporting Activities (SA)

The results of the CFA test showed that the items SA1 (b= 1.42; p<0.001), SA2 (b= 1.21; p<0.001), SA3 (b= 1.13; p<0.001), SA4 (b= 1.11; p<0.001) , SA5 (b= 1.34; p<0.001), SA6 (b= 1.30; p<0.001), SA7 (b= 1.30; p<0.001), SA8 (b= 1.32; p<0.001), SA9 (b= 1.23; p<0.001), and SA10 (b= 1.40; p<0.001), which provide a significant factor loading or contribution to the instrument supporting activities and are statistically significant.

3. Intermediate Evaluation (IE)

The results of the CFA test show that the items IE1 (b= 2.17; p<0.001), IE2 (b= 2.15; p<0.001), IE3 (b= 2; p<0.001), and IE4 (b= 2.08; p<0.001), which provide a significant factor loading or contribution to the intermediate evaluation instrument and are statistically significant.

5. Final Evaluation (FE)

The results of the CFA test show that the items FE1 (b= 2.96; p<0.001), FE2 (b= 3.08; p<0.001), FE3 (b= 2; p<0.001), and FE4 (b= 3.47; p<0.001), which provide a significant factor loading or contribution to the final instrument evaluation and are statistically significant.

Discussions

There are five essential components which should be used in assessing the application of the PA. This approach is central because the components are complementary and related to each other to achieve the expected student learning goals.

Good instruments must be valid (Perry et al., 2018). Thus, experts in their fields conducted a content validity test on the developed instrument and then distributed the results to medical school students. The PA is a student-centered program. Hence, the effects of its implementation are evaluated directly by students. The results can also evaluate the curriculum in action (van der Vleuten et al., 2012). However, one of the challenges for researchers in compiling items in this questionnaire is its adaptation to a language more easily understood by students in the faculty of medicine, particularly in the pre-clinical stage.

All components of the PA in its implementation significantly affect the validity results of the evaluation of the PA implementation in a medical school. It is expected that medical training programs should have at least these five components in the PA to maximize the assessment goals (Netemeyer et al., 2003; van der Vleuten et al., 2012).

The 'learning activities' component focuses on the suitability of the learning process with the applied competency framework. 'Learning activities' consist of several learning methods that can guide learning and, if organized appropriately, provide a coherent program or curriculum that is designed according to instructional design principles. For this section, the students in this study showed that they were satisfied with the learning activities of the medical institution they evaluated. Furthermore, the students also showed that they were satisfied with the 'assessment activities' component. 'Assessment activities' are expected to provide information about student learning in different activities. These satisfactory findings are most likely factual.

However, half of the students disagreed with the results for the 'supporting activities' component or the dialogical mentorship process between student and teacher with continuous reflection and constructive feedback recorded in a portfolio. Therefore, the low stakes assessment emphasized by the PA principle was missing from these results. The

students may have perceived the 'learning and assessment activities' components in the previous part of the questionnaires as 'tasks' and 'marks'. They did not focus on the continuous dialogue between students and teachers, which provides further learning strategies based on previous data points from learning and assessment activities.

The unsatisfactory findings in the 'supporting activities' component were also in line with findings of the 'intermediate evaluation', where only slightly more than half of the students agreed. It can be presumed that students were puzzled with this part because perhaps there was an 'intermediate evaluation' as an intermediate 'test' in their institution instead of intermediate feedback based on previous data points. Next, the 'final evaluation' was the last component of the PA, which was highly rated by most students in this study. However, students may have perceived these final questions as the final exam they usually did instead of a final decision of the previous 'learning and assessment activities.' Providing decisions from stakeholders through various sources plays a vital role in this PA (Heeneman et al., 2015; Hofstede, 2011; van der Vleuten et al., 2010).

Similar research was conducted by Heeneman et al. (2015 with a qualitative study approach. They interviewed 17 students (7 male and 10 female) in the second-year pre-clinical stage regarding their perceptions of the assessment approach. The programmed learning research was undertaken during the first year with semi-structured interviews. Sampling was conducted with a maximum sampling variation to ensure their perceptions were represented. The data were examined using theory-based thematic analysis. The results showed that the components of a comprehensive PA, such as feedback, portfolio, assessment and assignments, are said to have supporting and inhibitory effects on learning.

Furthermore, Driessen et al. (2012) investigated final-year students undergoing their internship (clerkship). This study used the PA model from van der Vleuten to develop a workplace-based assessment program to assess students in a more feasible and less bureaucratic manner. They used a questionnaire and focus group discussion (FGD) on 670 students (from 2007-2009). The students were asked to complete a web-based questionnaire to explore their learning experiences. The results obtained from their first implementation of the PA showed positive results.

Perry et al. (2018) also tried to develop a PA blueprint for assessing the competency progress of emergency medicine residencies at the University of Michigan. The results of this study indicate that residents accepted this rating system well. From a practical point of view, this strong PA showed that all residents have successfully progressed through this residency program. The similarity of this research with the authors' research is that they focus on the student's perspective concerning the implementation of the PA. The difference lies in the research subjects, since the Perry et al. research was conducted for residencies in America while the authors of the current study worked with undergraduate medical students (pre-clinical) in Indonesia.

Based on the several studies mentioned above, the evaluation of the implementation of PA was limited to a qualitative approach which primarily considered the student's perspectives. Therefore, the authors aimed to create valid and reliable instruments for evaluating the implementation of a PA based on the characteristics of each component so that it can be used as feedback for health education among other stakeholders.

The instrument developed in this study is structured based on psychometric principles, which are typically used in developing summative assessments such as Multiple-Choice Questions. However, the construct of this questionnaire refers to the principles of PA. Psychometric principles were used in this instrument's development to convince medical teachers in the context of the hierarchical culture in this study to trust the rigorousness of the instrument, use it, and evaluate their student assessment system.

This study initiates the use of multiple data points of student assessment, reflection on students' learning, and the importance of constructive feedback to promote mutual dialogue and longitudinal systematic assessment in hierarchical and collectivistic cultural settings. As a pioneering study, there should be more rigorous faculty development programs following the use of the PA in this study in many other clinical settings to collaborate our findings.

There are twelve consensus and qualitative tips regarding the PA, one being the collaboration between stakeholders (Heeneman et al., 2015). However, our questionnaires in this study can specifically evaluate the components of the PA. Continuous feedback and reflection are the main messages of student-centered learning. For institutions that are far from receiving feedback and implementing reflection as their cultural learning habit, our questionnaire will be very helpful and can be used to build a good assessment towards competence-based education (Ajzen, 1991; Hofstede, 2011; RG Netemeyer et al., 2003; Perry et al., 2018).

Student-centered learning (SCL) will occur when lecturers and students actively engage in the learning process. In the PA, the students are facilitated to explore teaching materials and discuss various information while the lecturers actively assist them during the process, including encouraging students to search, discuss, and draw clear conclusions from the results of their discussions. The coherence between learning activities and assessment methods while maximizing the quality of feedback will increase student learning achievement in a SCL concept. This principle is also reflected in the main objective of the PA, which is to encourage self-initiative and motivate student learning (Hofstede, 2011; Kerlinger, 1986; Martinez & Lipson, 1989; van der Vleuten et al., 2010).

The PA is another way of thinking about enhancing the role of evaluating students' learning progress. However, it is challenging to translate the elements of the theory into the specifics of implementation. Implementing the PA requires a different mindset from many of its stakeholders, but having a set of indicators developed for the PA might help schools more easily accept and implement this assessment approach. Having students evaluate the assessment system may also further support the institution in having a better learning program.

Conclusions

A five-component evaluation instrument based on medical students' perceptions of implementing the PA was developed and tested in this study using confirmatory factor analysis. Most students were satisfied with their institution's 'learning activities' and 'assessment activities' but rated 'supportive activities,' i.e., the mentorship system, as unsatisfactory. These results indicate that there needs to be more student-teacher two-way dialogue that focuses on the students' reflections on their learning progress and plan. At this point, student participation in their learning process is minimal. Therefore, when the students rated 'intermediate and final evaluation' as satisfactory, they may have perceived these two activities as 'examinations' rather than as feedback on their learning progress. Moreover, teachers still determine the results of students' learning, so general medical education in this study was still part of a teacher-centered learning curve. The instrument can also be further developed for multipurpose users to be applied to other health-related students at various levels of health professions education.

We encourage future studies to use more systematic, longitudinal, and multiple data points as the principle of the PA. The challenge of implementing this kind of assessment should also encourage researchers to discover the qualities and traits that ought to exist in each element of the PA model and to create a good instrument for its evaluation.

Recommendations

Further studies should be done to provide instruments for students completing their clinical stage and lecturers. Therefore, the instrument produced in this study should be further tested in other medical schools at various education levels, preferably in different countries, to determine its general applicability. Practice implication for clinical practitioners we promote a more-two-way dialogue with the students when doing the workplace-based observations or assessments in clinical settings as indicated in the results of this study. Continuous feedback and reflection are the key messages of student-centered learning and PA, especially for institutions that are far from accepting feedback and adopting reflection as a habit of their cultural learning.

Limitations

The instruments produced from this study have limitations as they can only be used to evaluate the implementation of the PA by assessing students' perceptions at the pre-clinical level. In addition, this instrument is only suitable for medical students, while its application for other health-related students at various levels of health professions education has not been determined.

Acknowledgments

We would like to express our highest appreciation to Professor Lambert Schuwirth, from Flinders University, Australia, who provided comments for the validation process of the instrument in this study, and Professor Cees van der Vleuten, from Maastricht University, who provided helpful comments to this manuscript. A special thanks for dr. Astrid Pratidina Susilo, Sp.An., MPH, Phd for her scientific suggestion regarding factor analysis in this study. We are grateful to the Faculty of Medicine at Universitas Islam Al-Azhar for funding this study publication. We are also very thankful to the medical teachers and students who participated in developing the instrument in this study.

Ethical Approval and Informed Consent

This study was approved by The Medical and Health Research Ethical Committee of the Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia, with the reference number KE/FK/0303/EC/2018.

Conflict of Interest

We declare that there is no conflict of interest.

Author Contribution Statement

We declare that all authors contributed significantly to the work reported, whether it was in the conception, study design, execution, data acquisition, analysis, and interpretation, or all of these areas; took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; agreed on the journal to which the article was submitted; and agreed to abide by the journal's policies.

References

Ainin, D. Q. (2018). *Penyusunan instrumen evaluasi penerapan programmatic assessment di institusi pendidikan kesehatan* [Development of evaluation instruments for implementation of programmatic assessment in health professions education institutions] (Publication No. 164087) [Master's thesis, Universitas Gadjah Mada]. UGM Campus Repository. http://bit.ly/3RuRTe0

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Amin, Z., Seng, C. Y., & Eng K. H., (2006). *Practical guide to medical student assessment*. World Scientific. https://doi.org/10.1142/9789812773586

Attard, A., di Iorio, E., Geven, K., & Santa, R. (2010). *Student-centred learning - Toolkit for students, staff and higher education institutions in learning.* The European Students' Union Student-Centre.

Bok, H. G. J., de Jong, L. H., O'Neill, T., Maxey, C., & Hecker, K. G. (2018). Validity evidence for programmatic assessment in competency-based education. *Perspectives on Medical Education*, *7*(6), 362–372. https://doi.org/10.1007/s40037-018-0481-2

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, *119*(2), 166.e7-166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036

Davis, A. E. (1996). Instrument development: Getting started. *Journal of the American Association of Neuroscience Nurses*, *28*(3), 204–207. https://doi.org/10.1097/01376517-199606000-00009

Driessen, E. W., van Tartwijk, J., Govaerts, M., Teunissen, P., & van der Vleuten, C. P. M. (2012). The use of programmatic assessment in the clinical workplace: A Maastricht case report. *Medical Teacher*, *34*(3), 226–231. https://doi.org/10.3109/0142159X.2012.652242

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. (2021). Ottawa 2020 consensus statement for programmatic assessment – 1. Agreement on the principles. *Medical Teacher*, *43*(10), 1139–1148. https://doi.org/10.1080/0142159X.2021.1957088

Heeneman, S., Pool, A. O. , Schuwirth, L. W. T., van der Vleuten, C. P. M., & Driessen, E. W. (2015). The impact of programmatic assessment on student learning: Theory versus practice. *Medical Education*, *49*(5), 487–498. https://doi.org/10.1111/medu.12645

Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, *2*(1), 1–26. https://doi.org/10.9707/2307-0919.1014

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). Holt, Rinehart and Winston.

Lacasse, M., Renaud, J. S., Côté, L., Lafleur, A., Codsi, M. P., Dove, M., Pélissier-Simard, L., Pitre, L., & Rheault, C. (2022). Développement et mise à l'essai du guide de rétroaction francophone pour l'observation directe des résidents en médecine familiale au Canada [Feedback guide for direct observation of family medicine residents in Canada: A francophone tool]. *Canadian Medical Education Journal*, *13*(1), 29–54. https://doi.org/10.36834/cmej.72587

Martinez, M. E. , & Lipson, J. I. (1989). Assessment for learning. *Educational Leadership*, *46*(7), 73–75.

McComas, W. F. (Ed.). (2014). *Summative assessments bt - the language of science education: An expanded glossary of key terms and concepts in science teaching and learning*. SensePublishers. https://doi.org/10.1007/978-94-6209-497-0_93

Netemeyer, R. G., Bearden, W., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. SAGE Publications. https://doi.org/10.4135/9781412985772

Perry, M., Linn, A., Munzer, B. W., Hopson, L., Amlong, A., Cole, M., & Santen, S. A. (2018). Programmatic assessment in emergency medicine: Implementation of best practices. *Journal of Graduate Medical Education*, *10*(1), 84–90. https://doi.org/10.4300/JGME-D-17-00094.1

Said, H., Badru, B. B., & Shahid, M. (2011). Confirmatory factor analysis (CFA) for testing validity and reliability instrument in the study of education. *Australian Journal of Basic and Applied Sciences*, *5*(12), 1098-10103. https://bit.ly/3EE2VIO

Schut, S., Driessen, E., van Tartwijk, J., van der Vleuten, C., & Heeneman, S. (2018). Stakes in the eye of the beholder: An international study of learners' perceptions within programmatic assessment. *Medical Education*, *52*(6), 654–663. https://doi.org/10.1111/medu.13532

Shumway, J. M., & Harden, R. M. (2003). AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher*, *25*(6), 569–584. https://doi.org/10.1080/0142159032000151907

Torre, D. M., Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2020). Theoretical considerations on programmatic assessment. *Medical Teacher*, *42*(2), 213–220. https://doi.org/10.1080/0142159X.2019.1672863

van der Vleuten, C., Lindemann, I., & Schmidt, L. (2018). Programmatic assessment: The process, rationale and evidence for modern evaluation approaches in medical education. *The Medical Journal of Australia*, *209*(9), 386–388. https://doi.org/10.5694/mja17.00926

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239

van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice and Research: Clinical Obstetrics and Gynaecology*, *24*(6), 703–719. https://doi.org/10.1016/j.bpobgyn.2010.04.001

Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet*, *357*(9260), 945–949. https://doi.org/10.1016/S0140-6736(00)04221-5

Wilkinson, T. J., & Tweed, M. J. (2018). Deconstructing programmatic assessment. *Advances in Medical Education and Practice*, *9*, 191–197. https://doi.org/10.2147/AMEP.S144449