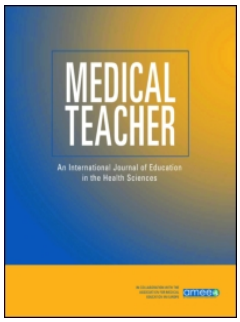

Medical Teacher

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

Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment

Sylvia Heeneman, Suzanne Schut, Jeroen Donkers, Cees van der Vleuten & Arno Muijtjens

To cite this article: Sylvia Heeneman, Suzanne Schut, Jeroen Donkers, Cees van der Vleuten & Arno Muijtjens (2016): Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment, Medical Teacher, DOI: 10.1080/0142159X.2016.1230183

To link to this article: http://dx.doi.org/10.1080/0142159X.2016.1230183

Published online: 19 Sep 2016.

MEDICAL TEACHER, 2016
http://dx.doi.org/10.1080/0142159X.2016.1230183

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment

Sylvia Heeneman[a,c], Suzanne Schut[b,c], Jeroen Donkers[b,c], Cees van der Vleuten[b,c] and Arno Muijtjens[b,c]

[a]Department of Pathology, Maastricht University, Maastricht, The Netherlands; [b]Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands; [c]School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

ABSTRACT

Background: Progress tests (PT) are used to assess students on topics from all medical disciplines. Progress testing is usually one of the assessment methods of the cognitive domain. There is limited knowledge on how positioning of the PT in a program of assessment (PoA) influences students' PT scores, use of PT feedback and perceived learning value.

Methods: We compared PT total scores and use of a PT test feedback (ProF) system in two medical courses, where the PT is either used as a summative assessment or embedded in a comprehensive PoA and used formatively. In addition, an interview study was used to explore the students' perception on use of PT feedback and perceived learning value.

Results: PT total scores were higher, with considerable effect sizes (ESs) and students made more use of ProF when the PT was embedded in a comprehensive PoA. Analysis of feedback in the portfolio stimulated students to look for patterns in PT results, link the PT to other assessment results, follow-up on learning objectives, and integrate the PT in their learning for the entire PoA.

Conclusions: Embedding the PT in an assessment program designed according to the principles of programmatic assessment positively affects PT total scores, use of PT feedback, and perceived learning value.

Introduction

Problem-based learning (PBL) can foster and guide students to self-direct their learning (Dolmans et al. 2005; Loyens et al. 2008). However, it has been shown that the end of block tests are more powerful drivers of learning as compared to the personal learning objectives, and this leads to unwanted, shallow learning approaches (Al Kadri et al. 2009; Cilliers et al. 2012). To avoid typical test-enhanced learning behavior, the progress test (PT) was introduced as a comprehensive test, attuned to the learning objectives at the end of the curriculum (Van der Vleuten et al. 1996; Freeman et al. 2010). The longitudinal and repetitive nature of the PT should encourage knowledge retention, deeper learning behavior and discourage last minute superficial learning (Schuwirth & van der Vleuten 2012). However, as with every assessment method, the design of and choices made in the implementation of the PT in the curriculum and assessment program may affect the perceived positive effects of the PT on performance and learning.

It was previously shown that the introduction of the PT as a formative longitudinal assessment procedure, regularly providing feedback, was associated with better performance of medical graduates on a licensing exam (Norman et al. 2010). A yearly, formative PT with limited feedback was acceptable for postgraduate residents, however the self-reported educational impact was low (Dijksterhuis et al. 2013). In addition, Wade et al. (2012) showed that the design and embedding of the PT in the curriculum significantly affected study patterns and students' perception of the value for learning and preparation time. In a setting of more frequent summative PTs, no end of block tests and feedback on PT performance, students more appreciated the PT as a support for learning and encouraging a deeper learning approach, although more time was used for the preparation (Wade et al. 2012). These studies show that important factors in the design and implementation of the PT are the formative or summative nature of the PT, the PT frequency and whether or not feedback on performance is given. However, more research is needed to optimize the embedding of the PT in a curriculum and assessment program to ensure better performance on the PT and its effects on student learning.

The PT is often used alongside other assessment methods, such as end of unit or block tests, and assessment of

Practice points

- Using the progress test (PT) in a comprehensive program of assessment (PoA) is beneficial in terms of results and educational impact.
- Stimulating the use of PT feedback by students through analysis of patterns, formulation, and follow-up of learning objectives is helpful for further learning and progress in the knowledge domain.
- This study calls on assessment and curricula designers to consider formative use of PTs in a PoA to maximize the benefits of the longitudinal and repetitive features of the PT and use of its feedback.

CONTACT Sylvia Heeneman s.heeneman@maastrichtuniversity.nl Department of Pathology, Maastricht University/MUMC, Peter Debyelaan 25, 6229 HX Maastricht, the Netherlands

clinical skills and communication. In the assessment program, it needs to be clear what the primary purpose of these assessments is, i.e., decision-making, providing information on student progress, or to drive student learning (Ricketts et al. 2010). If the primary purpose of the PT is to drive student learning, feedback on performance is important (Muijtjens et al. 2010). Even more important is that the students will look back at and use this feedback information. It is well known that a pass on a summative, consequential test will not encourage students to look back at and learn from the feedback and use it for future learning (Archer 2010; Harrison et al. 2015). A programmatic assessment model has been proposed to optimize the use of assessment information and feedback for self-regulation of learning by students (van der Vleuten & Dannefer 2012; van der Vleuten et al. 2012). In this approach, assessment information and feedback, originating from different forms of assessment and evaluations, is combined in a high-stakes decision, i.e., on promotion to the next year. Single assessments do not lead to summative pass–fail decisions. In addition, students are asked and encouraged to reflect on the assessment and feedback information to self-direct their learning. The integration of the PT in a comprehensive assessment program, designed according to the principles of programmatic assessment has not been studied yet. This is of interest both from the perspective of optimizing growth in PT performance, and from the perspective of the student, whether the active use of PT feedback aids or encourages a deeper learning approach. Therefore, we performed a study to answer the following three research questions: (1) how does the performance on a summative PT compare to performance on a formative PT embedded in a comprehensive program of assessment (PoA)? (2) how does the use of the online PT feedback system compare in the summative versus the formative setting? and (3) how do medical students perceive the active use of PT feedback on how they prepare for and learn from the PT?

Methods

Educational context

This study was performed at Maastricht University, the Netherlands. Two medical courses were used: students of the six-year bachelor–master program (6yrBM), and students of the four-year graduate-entry Master program (4yrM). The curriculum of both medical courses uses PBL and is competency-based, using the CanMEDS framework (Frank 2005; van Herwaarden et al. 2009). Table 1 shows the main features of the educational and assessment program of both medical courses and the position of the PT in the program.

Progress test

The PT is composed of multiple choice questions with a single correct answer and a "do not know" option. The test has 200 items, representing a sample from the relevant and functional knowledge domain that graduates are expected to know by the end of their training (Van der Vleuten et al. 1996). The test scores are calculated according to formula scoring (Diamond & Evans 1973; Rowley & Traub 1977), i.e., a correct answer is rewarded by +1 point, and an incorrect answer is penalized with −1/(number of alternatives −1). The "do not know" option yields 0 points, thus allowing a student to avoid the penalty of an incorrect answer when knowledge of the item's subject is lacking. The final score is the sum of all rewards and penalties expressed as a percentage of the maximum attainable score. In the sequel of the paper, "score" will be used to indicate the formula score. The medical PT in the Netherlands is a collaboration between 5 of the 8 medical schools. The jointly constructed PT is taken by all medical students of the five medical schools at the same day and time four times a year (Schuwirth et al. 2010). The benchmark group for the

Table 1. Main features of the curriculum and assessment program of both medical programs.

6yrBM program	4yrM program
Majority of students enter the medical course directly after secondary education/high school, average age at entry is 18 years, at graduation 24 years	Graduate-entry program, student have completed a biomedical bachelor program, average age at entry is 21 years, at graduation 25 years
Three year (preclinical) Bachelor program, three year (clinical) Master program.	1.5 year preclinical program, 2.5 year clinical program
300 students/year	30 students/year in 2010 and 2011, 50 students/year in 2012
Students graduate as Physicians	Students graduate as Physician-Clinical Investigators, emphasis in the curriculum on scholarly topics
Problem/Patient-based learning (PBL) in the pre-clinical years	
Clinical rotations two years: 11 disciplines, two electives; Senior Clerkship Clinical Care in discipline of choice (18 weeks) and Senior Clerkship Clinical/scientific research in discipline of choice (18 weeks)	Clinical rotations 1.8 years (11 disciplines), Senior Clerkship combined clinical care and research in discipline of choice (30 weeks)
Separate educational units such as the PBL units, clinical clerkships, the progress test (see below) are appointed credits/ECTS, adding up to 60 ECTS/year	All assessment information and feedback, including the progress test, is used by the student for reflective activities and setting learning objectives. All information and reflective activities is aggregated in a portfolio. The portfolio is assessed at the end of the year, and appointed the majority of ECTS/year in 2010 and all ECTS in 2011 and 2012
Student has a mentor and portfolio in Year 1 and 5	Students has a portfolio and the same mentor in all years of the program
Four progress tests per year	
Feedback on progress test score/performance – total and by subject area	
The individual total score on each progress test is calculated (as described in the method section). A pass/fail cutoff of the test scores is set at the mean minus the standard deviation (for the cohort of students in the same year-group), and student receives an indication of fail–pass–good for each individual progress test	
Summative assessment: the four scores are combined to determine whether the student has made adequate progress and receives a (summative) overall pass for the progress tests of that year, yielding nine ECTS (European Credit Transfer System)/year (of 60 ECTS/year) in the Bachelor and 12 ECTS/year (of 60 ECTS/year) in the Master	The progress test results and feedback are (1) (in)formative and not connected to a pass–fail decision for the progress test only, and (2) are embedded in the portfolio assessment. The portfolio assessment yields all 60 ECTS/year)

6yrBM program are the combined results of all medical students (of collaborating medical schools) that have taken the test (n on average 1170 students per PT per year group). The benchmark group for the 4yrM students are the students of the same year group (30 students/year in 2010 and 2011, 50 students/year in 2012).

Data collection and analysis

Quantitative and qualitative methods were used to investigate the research questions (see Figure 1).

For Research Question 1, quantitative data was used of the PT scores of both groups of students (4yrM and 6yrBM). Each academic year, four PT were taken by students of all year groups, so throughout the 6yrBM program a student was assessed at 24 measurement moments (MMs). For a student in the 4yrM program it is $4 \times 4 = 16$ MMs. When comparing the scores of the two programs, MMs 9–24 of the 6yrBM students, were compared to the 16 MMs of 4yrM students. Thus the 4yrM students started at the level of MM 9, owing to the fact that these students have a certain level of knowledge acquired in their prior undergraduate/bachelor program. In the time-frame of analysis (academic years 2010–2011, 2011–2012, and 2012–2013), data could be used of cohorts starting the 6yrBM program in 2005–2010, and starting the 4yrM program in 2007–2012.

Between-group differences in mean score were tested using a two-sample t-test and the substantiality of the difference was evaluated by calculating Cohen's d as a measure of effect size (ES). Cohen's criteria (0.2 small effect, 0.5 medium effect, 0.8 large effect) were used to qualify the ES (Cohen 1988). Per MM the overall ES was obtained by calculating the weighted average of the d-indices over the three academic years according to the expression

$$\text{average } d = \frac{\sum_{i=1}^{3} d_i w_i}{\sum_{i=1}^{3} w_i}; w_i = \frac{2(n_{i1} + n_{i2}) n_{i1} n_{i2}}{2(n_{i1} + n_{i2})^2 + n_{i1} n_{i2} d_i^2}$$

where d_i is Cohen's d for the ith academic year, weight w_i is equal to the inverse of the variance of the d-index estimate d_i, and $n_{i1,}$ and n_{i2} are the sizes of the two groups of students (4yrM, and 6yrBM) in the ith academic year (Cooper 2009). To accommodate multiple comparisons, a Bonferroni correction was applied. Bonferroni corrected p-values were obtained by multiplying each single comparison p-value by the relevant number of corresponding multiple comparisons.

For Research Question 2, it was investigated if the students used the Progress Feedback system (ProF) (Muijtjens et al. 2010). This system allows students to view their scores, sub-scores (discipline, organ system) and aggregates of sub-scores (basic sciences, clinical sciences, behavioral sciences) either per test or longitudinally across tests. The logging data of ProF was used to calculate the average number of ProF sessions per student of the 4yrM program and the 6yrBM program, for each period between two successive PT. Logging data were only available for the curriculum year 2012–2013. The data for the two groups consisted of 16-paired observations (the number of sessions per student in the period of the PT corresponding to the MM). Only sessions where a student visited more than five pages were taken into account. A paired samples t-test was used to compare the between-group difference.

Table 2. Set of pre-defined codes, used for the initial template.

1. FEEDBACK
   a. Detect deficiencies
   b. Detect patterns in scoring
2. FOLLOW-UP FEEDBACK
   a. Discuss with mentor
   b. Formulate learning objectives
   c. Discuss with peers
   d. Reflection/analysis
   e. Preparation for the PT
   f. Strategy
3. SCORE
   a. Cut off/standard setting
   b. Pass–fail decision
4. SIGNIFICANCE PROGRESS TEST
   a. Useful
   b. Aggregate with other results
   c. Personal assessment of knowledge level
   d. Induces stress
5. POSITION PROGRESS TEST
   a. In program of assessment

For Research Question 3, interviews were used to explore experiences of the students in their natural context (Malterud 2001), using an interpretative, constructivist approach (Bunniss & Kelly 2010). Individual interviews with $n = 17$ 4yrM students ($n = 7$ male and $n = 10$ female) were conducted. The interview data was part of a larger study on the impact of programmatic assessment on student learning, which was conducted in the Year 2 cohort of the 4yrM program during November–December 2013 (Heeneman et al. 2015). The perception of the students on the use of PT feedback were not used in the previous study and analyzed separately for the purpose of the current study. The design of the interview study is described in (Heeneman et al. 2015). Briefly, Year 2 students of the 4yrM program were invited for individual interviews. Sampling of the students was done using a maximum variation sampling strategy, regarding gender and portfolio assessment result in Year 1. Verbatim transcripts of the interviews were made and analyzed using Template Analysis, which consists a succession of coding templates and hierarchically structured themes, that were applied to the data (King 2004). For the specific research question of the current study, the set of predefined codes was based on literature (Van der Vleuten et al. 1996; Muijtjens et al. 2010; Norman et al. 2010; Schuwirth & van der Vleuten 2012; Wade et al. 2012; Dijksterhuis et al. 2013) and shown in Table 2. Analysis of interview 1–8 (by SH and SS) using the pre-defined codes resulted in the initial template that was discussed and then applied to interview 9–17. The outcome of the analysis was a final template that was discussed with the research team, then analysis was advanced from the themes to an interpretation of the use of PT feedback and perceived learning value, as presented in the results.

Ethical considerations

For the interviews, participation was voluntary, students were ensured of confidentiality and signed an informed consent form. The ethical review board of the Dutch Association for Medical Education approved this study (approval NVMO-ERB-276). The PT and ProF data was collected as assessment information as part of the regular curriculum, and analyzed anonymously. The researchers were

educationalists (CvdV, AM, SS, JD), a psychologist (AOP, who performed the interviews) and a biologist with an educational background (SH). SS, CvdV, AM, JD, and AOP had no direct contact with the students in the program, SH did, as the program director.

Results

The progress in knowledge level as assessed by the PT was compared for the students in the 4yrM and 6yrBM program (Research Question 1). Total PT scores were based on approximately 270 students in the 6yrBM (excluding MM 23 and 24, see below) and 27 students in the 4yrM program 2010–2012 and 50 students in 2012–2013. Standard deviations of the scores were on average 7.5 (range: 6.0–9.2) in the 6yrBM program and 6.8 (range: 5.0–8.9) in the 4yrM program. Total PT scores (Figure 2, Table 3, column 5) showed that the 4yrM students started with a lack of knowledge (MM =9, highly significant difference, average ES= −1.10, a large effect]. However, their knowledge progress was steeper, and after one year there is no difference compared to the 6yrBM students. In the years 2–4 there was a consistent positive difference of medium ES. The differences for MMs 23 and 24 should be considered with caution as the scores for the 6yrBM program were most likely negatively biased as many 6yrBM students already had satisfactory (pass) results and therefore did not attend the last PT.

The sub-domain scores (rows 2–4, Figure 2) showed more irregular patterns compared to the Total PT score (row 1, Figure 2). The Scholarly Topics (row 4) were most irregular, expressing that the noise level in these subtest scores was higher due to the smaller number of items (see Table 3). Nevertheless the score patterns in all three sub-domains were informative for Research Question 1. For Basic Sciences, the knowledge level was comparable for the 4yrM students at the start of their program (MM =9), while in time the difference with the 6yrBM program became substantial with medium to large ESs. For the Clinical Sciences the pattern was different and showed a considerable lag in the first year for the 4yrM students which faded in the second year and remained comparable to the 6yBM program (in general positive but non-significant differences) in the next years. The pattern of the scholarly topics showed a favorable positive difference for 4yrM right from the start and consistent medium to large ES differences throughout all other MMs.

The use of PT feedback (Research Question 2) was investigated by comparing the use of the PT feedback system ProF by students of the 4yrM program and the 6yrBM program. Analysis showed that the 4yrM program students accessed the ProF system more often as compared (number of sessions per student per test, mean ± SD over all MMs: 1.10 ± 0.55) to the 6yrBM program (0.16 ± 0.05; Figure 3, $p = .0005$, ES [Cohen's d] 1.7).

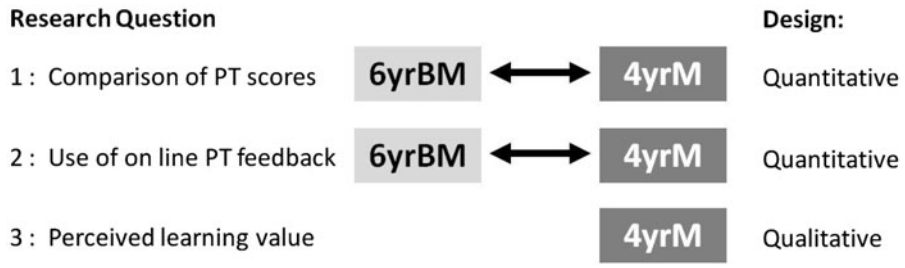

Figure 1. Outline of the study design.

Table 3. Effect size (Cohen's d) and statistical significance for between-group score differences of students in the 4yrM and the 6yrBM programs at Maastricht University, for the total progress test and for sub-domains, in academic years 2010–2011, 2011–2012, and 2012–2013.

	Total test (188 items[b])				Basic sciences (55 items)				Clinical sciences (100 items)				Scholarly topics (16 items)			
MM[a]	2010[c]	2011	2012	Ave[d]	2010	2011	2012	Ave	2010	2011	2012	Ave	2010	2011	2012	Ave
9	−0.97***	−1.27***	−1.09***	−1.10	−0.05	−0.40	−0.19	−0.21	−1.56***	−1.63***	−1.62***	−1.60	0.27	0.23	0.25	0.25
10	−0.16	−0.46	−0.60***	−0.45	1.06***	−0.10	−0.01	0.22	−0.91***	−0.77***	−0.92***	−0.87	0.57	0.76***	0.46**	0.57
11	−0.06	−0.49	−0.32	−0.31	0.85***	0.07	0.40*	0.40	−0.46*	−0.68***	−0.38	−0.49	0.91***	0.51**	0.90***	0.78
12	−0.35	−0.14	−0.06	−0.16	0.31	0.46	0.31	0.35	−0.79***	−0.66**	−0.42*	−0.58	0.77***	0.25	0.47**	0.48
13	0.20	−0.24	−0.08	−0.03	0.51	0.43	0.41	0.45	−0.29	−0.74***	−0.48*	−0.48	0.56	0.45	0.81***	0.61
14	0.71**	0.43	0.41	0.51	1.16***	0.49	0.49	0.70	0.04	0.14	−0.10	0.02	1.28***	0.48	1.35***	1.03
15	0.82**	0.59	0.48	0.62	1.28***	0.78**	0.80***	0.93	0.27	0.20	0.21	0.22	0.36	0.48	0.66*	0.50
16	0.10	0.42	0.43	0.31	0.44	0.95***	0.61*	0.65	−0.26	−0.17	0.08	−0.11	0.64*	0.73*	0.77**	0.71
17	0.49	−0.15	0.15	0.15	0.68*	0.43	0.84***	0.64	−0.22	−0.85***	−0.46	−0.51	1.39***	0.86***	0.64*	0.94
18	0.68**	0.55	0.12	0.45	1.22***	0.65*	0.73**	0.85	0.10	0.12	−0.37	−0.04	0.81**	0.58	0.92***	0.76
19	0.34	0.67*	0.50	0.50	0.78**	0.99***	0.67*	0.81	−0.04	−0.10	0.27	0.04	0.67*	1.08***	0.67*	0.80
20	0.30	0.62*	0.37	0.43	0.84***	1.41***	0.86***	1.04	−0.11	−0.08	−0.10	−0.09	0.60*	0.41	0.36	0.45
21	0.32	0.60*	0.94***	0.62	0.38	0.63*	1.01***	0.67	0.08	0.48	0.60*	0.39	0.74***	0.38	0.94***	0.69
22	0.39	0.55	0.58	0.50	0.70**	0.70**	0.56	0.65	0.14	0.30	0.31	0.25	0.36	0.24	0.62*	0.40
23	0.62*	1.00***	0.74*	0.77	0.89***	1.08***	0.76**	0.90	0.36	0.52	0.54	0.46	0.19	0.71**	0.66*	0.50
24	0.48	1.49***	1.33***	1.02	0.42	1.63***	1.21***	0.99	0.35	1.04***	0.86**	0.71	1.25***	0.91**	1.33***	1.14

*$p < .05$, **$p < .01$, ***$p < .001$ (Bonferroni corrected p-value of between-group difference, two-sample t-test, two-sided).
[a]MM: measurement moment; with four progress tests per year and six-year groups a student throughout the program is assessed at 24 measurement moments.
[b]Average number of items in total test and subtests.
[c]Academic year 2010–2011 (September 2010–June 2011).
[d]Ave: average effect size over academic years (weighted average using weights proportional to the inverse of the variance of Cohen's d); no statistical significance testing at this level.

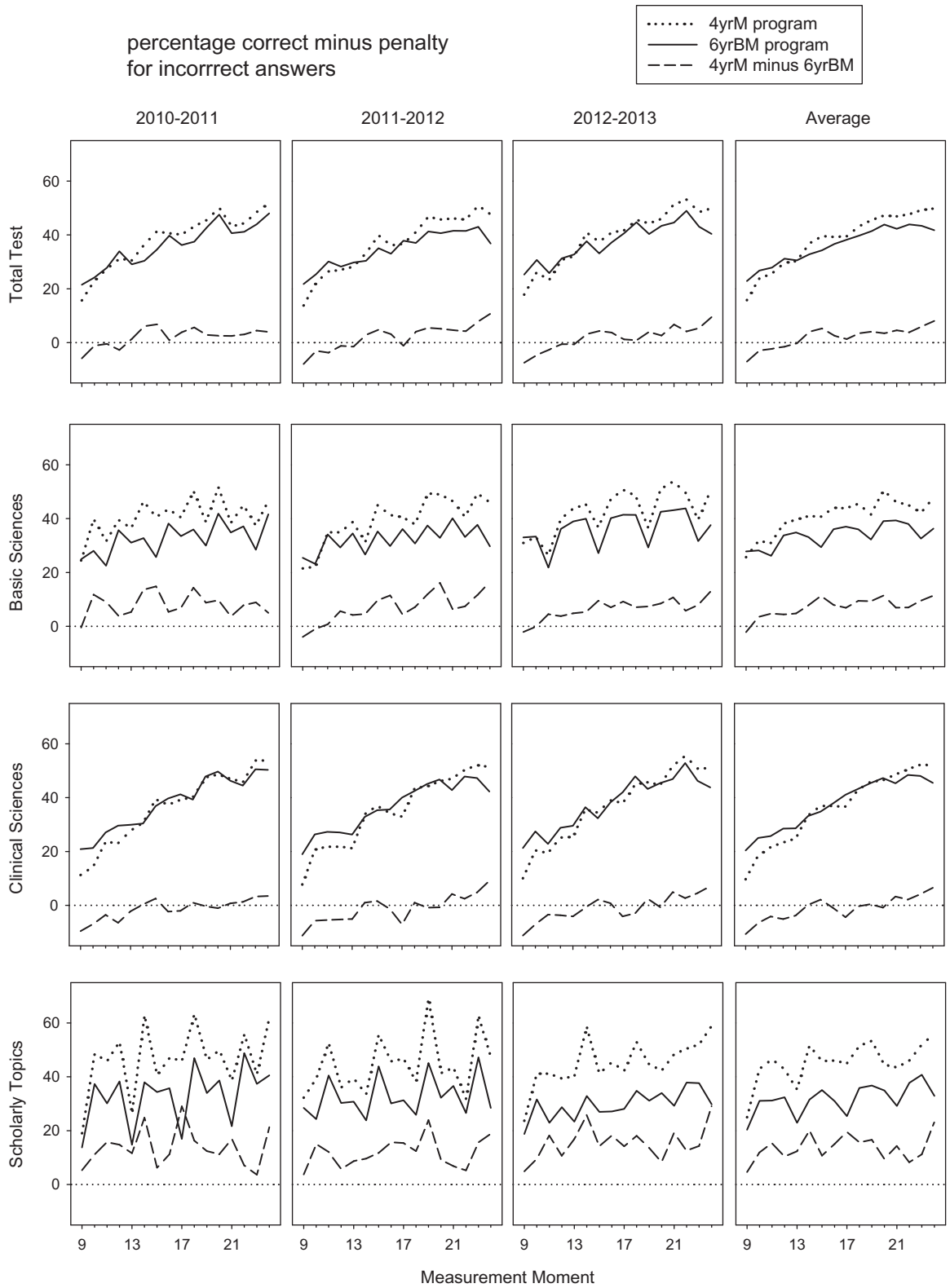

Figure 2. Progress test scores (formula score: percentage correct minus penalty for incorrect answers) for students of the four-year graduate-entry Master program (4yrM) and students of the (traditional) six-year bachelor–master program (6yrBM) at Maastricht University, the Netherlands. Shown are average scores per measurement moment for each program, and for the between program difference (see legend), for the total test, and for sub-domains basic sciences, clinical sciences, and scholarly topics. Scores of three academic years (2010–2011, 2011–2012, and 2012–2013), and the average over academic years are presented. Per academic year four progress tests were taken by students of all year groups resulting in scores at $6 \times 4 = 24$ measurement moments for the 6yrBM program. When comparing the scores of the two programs, the 16 measurement moments of 4yrM students were considered to correspond with measurement moments 9 up to 24 of the 6yrBM students.

The active use of feedback and perceived learning value of the PT was further investigated in the interview study (Research Question 3). Students of the 4yrM program indicated that the PT provided valuable information on both the level and progress of knowledge. Students were able, through the use of feedback, to monitor their progress and search for patterns in the scores of certain topics, to judge whether the score was comparable to the group level or showed evidence of deficiencies. It was indicated that the longitudinal feature of the PT was vital for this monitoring: this could not be done on the basis of 1 PT.

> *I think it [the progress test] is useful, as you can really see progress if you are doing well, or if your knowledge level is unchanging and you cannot establish this on one progress test,*

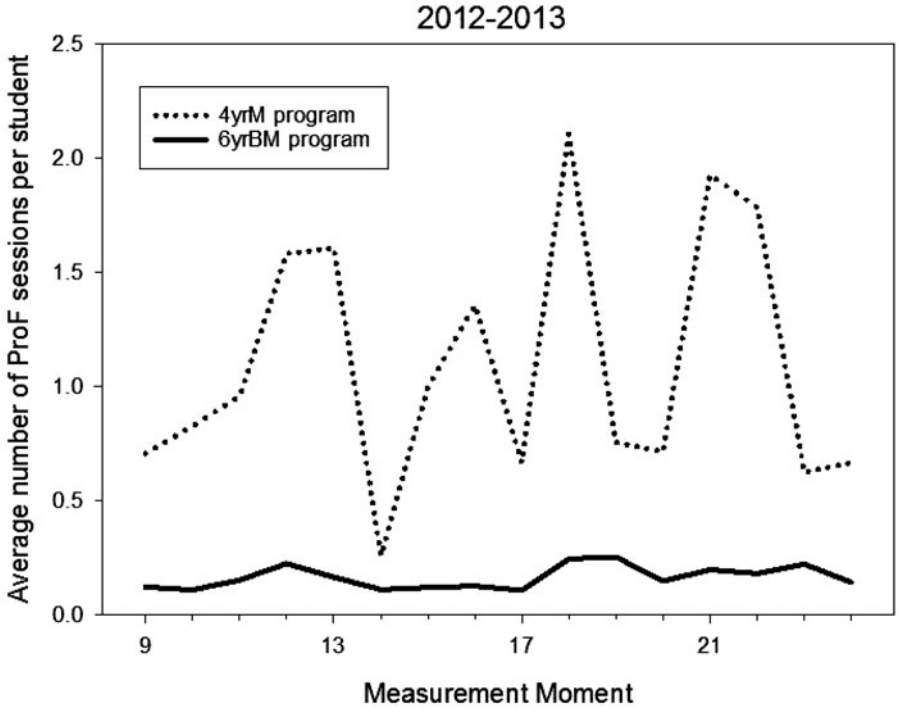

Figure 3. Use of the online Progress test feedback system (ProF) by 4yrM and 6yrBM student. Figure shows the results for measurement moments 9–24 in 2012–2013 for both programs, each point on a curve represents the average number of ProF sessions per student in the period of the progress test corresponding to the measurement moment, and the subsequent test.

but if you have three progress test scores and these all go down or up, then you are able to draw a conclusion (Int.7).

In addition, the analysis of PT feedback was used in combination with other results of the PoA. This gave added value to the use of information from PT results, as indicated by this student:

I have kept an eye on it and coupled to other results. So as an example, I scored low in Epidemiology, OK, but if it was also evident from other assessments or feedback that I score low, then you take it together to act upon. Look, if only the progress test shows a low score on let's say, clinical genetics, than I will do nothing. When it consistently occurs as a failure in my knowledge I, then I can potentially follow-up on it. But a single progress test result means nothing (Int. 7).

The analysis of feedback led to learning objectives, and follow-up of these learning objectives improved subsequent results, this could be the results of subsequent PT or other tests in the PoA. Learning objectives could also serve to get more control on the strategies used to answer questions during the PT, leading to more confidence and better scores. The analysis of feedback and the learning objectives were often discussed with the mentor.

At some point I made the learning objective to, eh, pay more attention [on clinical skills]. Just to read one of the clinical skills books every now and then or, yes, I made a learning objective that for a progress test I would repeat, eh, 3 clinical skills procedures and I changed how I prepared myself for the OSCE. Yes, this was helpful, in any case, let me put it like this, my score in the progress test improved, but if this was really was caused by the learning objective, I don't know, but I think it did [.]. I do have the feeling that I, yes, you do improve, as you are working on it continuously. So, I do think it helped (Int. 10).

Some students indicated to have less or no benefit from the analysis of PT feedback. Their sense of skepticism was partly caused by the fact that the group for comparison of the PT results was relatively small, the benchmark being based on the results in their year group in the 4yrM program and not the students in the 6yrBM program. It was also indicated that certain topics could be represented by many different questions in each PT, leading to variation in scores that was difficult to explain, or the PT results were not in line with the results of other assessment. In addition, it was felt that it was difficult to articulate specific learning objectives given the broad content of certain topics.

I analyze the results, yes, as I said my scores are really above average, so many of the topics have plus, plus, plus indications and there is little wrong with it. Then you can see small points you can play around with, but some topics are like, you get a score of 60%, then 10, then 45 and again 10. Then I think: OK, there is no pattern in that, and I really think this is caused by the sampling of questions in each progress test. Each topic has 5–10 questions and then it is just luck if these are the things you know. Off course you cannot blame everything on the sampling of questions, but it does play a role. This makes me question the reliability, as yes, on the one hand you have good results for all your assessments, and then the progress test tells you a week later that you are far below average on that competency, and then you think: how can I explain that? (Int. 11).

The embedding of the PT results in the portfolio was perceived as helpful, because it stimulated a critical analysis of PT results, even if the overall result was satisfactory, thereby adding to the perceived learning value of the test.

In that way, the portfolio is really good. You are forced, oh well, I also do it for my own benefit, but you have to analyze the test to see where you have deficiencies. And that gave me further insights. I am not sure if I would have done it anyhow, as my progress test result was a really good percentage, so I did not know that that topic was not so good, I was not thinking about that. So, it is not like that if you have a high score, that all is well. No, that is something I discovered through the portfolio (Int. 14).

Discussion

In this study, it was explored how embedding of the PT in a comprehensive PoA affected total test scores (Research

Question 1), how often online PT feedback was used by students (Research Question 2) and how do students perceive the active use of PT feedback on how they prepare for and learn from the PT (Research Question 3). In response to Research Question 1, it was shown that the total scores were higher, with considerable ESs, in three cohorts of students in the 4yrM program where the PT is embedded in a comprehensive PoA, compared to the 6yrBM program where each PT has a summative pass–fail decision. In response to Research Question 2, it was shown that compared to students in the 6yrBM program, the 4yrM program students made significantly more use of the ProF system. In addition (Research Question 3), the 4yrM students indicated that the analysis of feedback in the portfolio stimulated, or sometimes even compelled them, to look for patterns in PT results, link the PT to other assessment results, articulate and follow-up on learning objectives, and integrate the PT in their learning for the entire PoA.

It has been suggested that the PT will stimulate a deeper approach to learning (Van der Vleuten et al. 1996; Wade et al. 2012), although a recent study by Chen et al. (2015) did not show significant changes in approaches to learning in time after introduction of the PT. In the current study, the approach to learning was not determined by a standardized questionnaire, rather the students' perceptions on the learning value were sought via interviews. Students of the 4yrM program indicated that the analysis and follow-up of PT feedback was valuable for learning. This post-assessment learning effect was also shown for other elements of the PoA in our previous study (Heeneman et al. 2015). The longitudinal nature of the PT was helpful and important for the follow-up of information that was gained by analysis, it was clear that a single PT was considered as one data-point and learning objectives were based on the information of multiple PT. This fits with the theory of programmatic assessment (van der Vleuten et al. 2012) and shows that the PT can be a valuable component of an assessment program designed according to the principles of programmatic assessment. The importance of the overall design of the assessment program and how the PT can be used to advance learning, resonates with other studies (Wade et al. 2012; Pugh & Regehr 2016).

The total scores of the PT were higher in the 4yrM program. We can only speculate if this is caused by the embedding of the PT in a comprehensive assessment program where there was no summative pass–fail decision coupled to the PT. Theoretically this would be plausible, and described as the catalytic effect of assessment (Norcini et al. 2011) or the post-assessment effect (Dochy et al. 2007). Thus, assessments that generate feedback, in a setting where use of feedback is enhanced and supported, could move learning forward and in turn may lead to better assessment results. As for better performance as a result of the use of feedback, it has been shown that giving the opportunity to voluntarily complete in-course assessment tasks that generate feedback, improved the grades of the final exam (Gijbels et al. 2005). Thus, the higher scores for the PT in the 4yrM program could be partly due to the active use of PT feedback. However, the student population of the 4yrM and 6yrBM program is different, the students of the 4yrM program are older and have completed a bachelor program prior to entry into medical school. It has been shown that compared to a traditional five/six-year program, graduate-entry medical students performed at least as well, or marginally better on various bioscience knowledge and clinical assessments with ESs (Cohen's d) varying between of 0.21 and 0.38 (Dodds et al. 2010; Byrne et al. 2014). In our study, the ESs for total scores varied, but were close to 0.5 in the last 1.5 years, and even higher in the last 2 MMs. In addition, the patterns in the scoring on sub-topics were different (see Figure 2). It seemed plausible that the 4yrM students outperformed the 6yrBM students on Scholarly topics from the start, as the 4yrM students have completed a biomedical bachelor that featured a fair amount of scholarly topics. Nevertheless, the difference with the 6yrBM program increased as the 4yrM curriculum puts more emphasis on the Scholarly Topics, especially in the first 1.5 years. The Basic Science topics are comparable at the start of the 4yrM program, but the difference becomes larger during the years, this cannot solely be explained by the 4yrM curriculum, as in the first 1.5 years there is limited time available for the Basic Sciences. Clinical and communication skills are taught and need to be mastered in only 1.5 years, instead of three years in the 6yrBM program. This could suggest that the 4yrM students self-direct their learning more, guided by the analysis and use of information generated by the PoA, to keep up and expand their knowledge of the Basic Science topics, which may have led to higher PT scores on this topic.

In terms of self-direction of learning, both the 4yrM and 6yBM program are PBL curricula. In contrast to the 6yrBM program, several elements were implemented in the 4yrM program that align with the definition of Knowles for self-regulated learning (Knowles 1975), with learners being involved in the identification of learning needs, involved in implementing a learning process, committed to a learning contract and an evaluation/self-assessment of the learning process. Students used the electronic feedback system ProF more frequently, and indicated that they indeed analyzed the feedback and assessment results, formulate learning objectives and self-assess if their strategies were successful. In a systematic review, it was shown that self-direction of learning was associated with a moderate increase in performance in the knowledge domain, a marginal, yet insignificant increase in the skills domain and no change in the attitudes domain (Murad et al. 2010). The PT is an assessment method that is aimed at the knowledge domain and may thus benefit more from a self-directed learning approach. More advanced learners, that is students in later years or residency, benefitted more from self-directed learning (Murad et al. 2010). In the graduate-entry programs, students are older and have completed a bachelor degree, and thus may be more efficient in self-direction of learning, especially in a setting of programmatic assessment where learning from feedback and assessment is encouraged, supported, and stimulated. This may also have had an effect on the difference in PT results.

The findings of this study should be regarded given certain limitations. Firstly, this study was conducted at a single university, so the results may not be transferable to other settings. On the other hand, this may have enhanced comparability in terms of infrastructure (library, resources), and educational vision (PBL) when comparing the two medical courses in our setting. Secondly, the qualitative interview study was only performed in the 4yrM program and we

have no information on the perceived learning value of the PT in the 6yrBM program. The (quantitative) results of the use of the ProF system however, confirmed the results of the interview study that in the 4yrM setting where the PT is used formatively in a comprehensive PoA, students make more use of the feedback as compared to the 6yrBM program in which a summative PT is used.

In conclusion, this study suggests that embedding of the PT in a PoA designed according to the principles of programmatic assessment is associated with higher PT total scores, and increased use of PT feedback for learning. This study may suggest that the principles of programmatic assessment, i.e., to use assessment and feedback as informative for learning, are very compatible with the principle of longitudinal and repetitive assessment in the PT. Educators in the field of health profession education could consider a combination of these principles to yield beneficial effects in terms of results and educational impact.

Glossary

Programmatic assessment: An integral approach to the design of an assessment program with the intent to optimize its learning function, its decision-making function, and its curriculum quality-assurance function.

Progress test: A comprehensive test sampling knowledge across all content areas of medicine reflecting the end objectives of the curriculum.

Acknowledgements

The authors want to thank Andrea Oudkerk Pool (AOP) for conducting the interviews.

Disclosure statement

The authors report no declarations of interest.

Notes on contributors

Sylvia Heeneman, PhD, is Professor of Medical Education at the School of Health Profession Education, Department of Pathology, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

Suzanne Schut, MSc, is an educationalist at the Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht, the Netherlands.

Jeroen Donkers, PhD, is an assistant professor, and knowledge engineer at the Department of Educational Development and Research, Faculty of Health, Medicine, and Life Sciences, Maastricht University, the Netherlands.

Cees van der Vleuten, PhD, is Professor of Education, and Scientific Director of the School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

Arno M. Muijtjens, PhD, is an associate professor and a statistician-methodologist at the Department of Educational Development and Research, Faculty of Medicine, Maastricht, the Netherlands.

References

Al Kadri H, Al-Moamary M, van der Vleuten C. 2009. Students' and teachers' perceptions of clinical assessment program: a qualitative study in a PBL curriculum. BMC Res Notes. 2:263.

Archer J. 2010. State of the science in health professional education: effective feedback. Med Educ. 44:101.

Bunniss S, Kelly DR. 2010. Research paradigms in medical education research. Med Educ. 44:358–366.

Byrne A, Arnett R, Farrell T, Sreenan S. 2014. Comparison of performance in a four year graduate entry medical programme and a traditional five/six year programme. BMC Med Educ. 14:248.

Chen Y, Henning M, Yielder J, Jones R, Wearn A, Weller J. 2015. Progress testing in the medical curriculum: students' approaches to learning and perceived stress. BMC Med Educ. 15:147.

Cilliers FJ, Schuwirth L, Herman N, Adendorff HJ, van der Vleuten CPM. 2012. A model of the pre-assessment learning effects of summative assessment in medical education. Adv Health Sci Educ. 17:39–53.

Cohen J. 1988. Statistical power analysis for the behavioural sciences. London: Lawrence Erlbaum.

Cooper H. 2009. Research synthesis and meta-analysis: a step-by-step approach, vol. 2. Thousand Oaks (CA): Sage Publications.

Diamond J, Evans W. 1973. The correction for guessing. Rev Educ Res. 43:181–191.

Dijksterhuis M, Schuwirth L, Braat D, Scheele F. 2013. An exploratory study into the impact and acceptability of formatively used progress testing in postgraduate obstetrics and gynecology. Perspect Med Educ. 2:126–141.

Dochy F, Segers M, Gijbels D, Struyven K. 2007. Assessment engineering. In: Boud D, Falchikov N, editors. Rethinking assessment in higher education: learning for the longer term. Oxford: Routledge p. 87–100.

Dodds A, Reid K, Conn J, Elliott S, McColl G. 2010. Comparing the academic performance of graduate- and undergraduate-entry medical students. Med Educ. 44:197–204.

Dolmans D, De Grave W, Wolfhagen I, Van Der Vleuten C. 2005. Problem-based learning: future challenges for educational practice and research. Med Educ. 39:732–741.

Frank JR. 2005. The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care. Ottawa: The Royal College of Physicians and Surgeons of Canada (report).

Freeman A, van der Vleuten C, Nouns Z, Ricketts C. 2010. Progress testing internationally. Med Teach. 32:451–455.

Gijbels D, van de Watering G, Dochy F. 2005. Integrating assessment tasks in a problem-based learning environment. Assess Evaluat Higher Educ. 30:73–86.

Harrison C, Könings K, Schuwirth L, Wass V, van der Vleuten C. 2015. Barriers to the uptake and use of feedback in the context of summative assessment. Adv Health Sci Educ. 20:229–245.

Heeneman S, Oudkerk Pool A, Schuwirth L, van der Vleuten C, Driessen E. 2015. The impact of programmatic assessment on student learning: theory versus practice. Med Educ. 49:487–498.

King N. 2004. Using templates in the thematic analysis of text. In: Cassel C, Symon G. editors. Essential guide to qualitative methods in organizational research. London: Sage.

Knowles M. 1975. A guide for learners and teachers. England Cliffs: Prentice Hall/Cambridge.

Loyens S, Magda J, Rikers R. 2008. Self-directed learning in problem-based learning and its relationships with self-regulated learning. Educ Psychol Rev. 20:411–427.

Malterud K. 2001. Qualitative research: standards, challenges, and guidelines. Lancet. 358:483–488.

Muijtjens A, Timmermans I, Donkers J, Peperkamp R, Medema H, Cohen-Schotanus J, Thoben A, Wenink A, van der Vleuten C. 2010. Flexible electronic feedback using the virtues of progress testing. Med Teach. 32:491–495.

Murad M, Coto-Yglesias F, Varkey P, Prokop L, Murad A. 2010. The effectiveness of self-directed learning in health professions education: a systematic review. Med Educ. 44:1057–1068.

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

Norman G, Neville A, Blake J, Mueller B. 2010. Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. Med Teach. 32:496–499.

Pugh D, Regehr G. 2016. Taking the sting out of assessment: is there a role for progress testing? Med Educ. 50:721–729.

Ricketts C, Freeman A, Pagliuca G, Coombes L, Archer J. 2010. Difficult decisions for progress testing: how much and how often?. Med Teach. 32:513–515.

Rowley G, Traub R. 1977. Formula scoring, number-right scoring, and test-taking strategy. J Educ Measure. 14:15–22.

Schuwirth L, Bosman G, Henning R, Rinkel R, Wenink A. 2010. Collaboration on progress testing in medical schools in the Netherlands. Medl Teach. 32:476–479.

Schuwirth L, van der Vleuten C. 2012. The use of progress testing. Perspect Med Educ. 1:24–30.

van der Vleuten C, Dannefer E. 2012. Towards a systems approach to assessment. Med Teach. 34:185–186.

van der Vleuten C, Schuwirth L, Driessen E, Dijkstra J, Tigelaar D, Baartman L, van Tartwijk J. 2012. A model for programmatic assessment fit for purpose. Med Teach. 34:205–214.

Van der Vleuten C, Verwijnen G, Wijnen W. 1996. Fifteen years of experience with progress testing in a problem-based learning curriculum. Med Teach. 18:103–110.

van Herwaarden C, Laan R, Leunissen R. 2009. The 2009 Framework for Undergraduate Medical Education in the Netherlands. Available from: http://www.vsnu.nl/Media-item/Raamplan-Artsopleiding-2009.htm.

Wade L, Harrison C, Hollands J, Mattick K, Ricketts C, Wass V. 2012. Student perceptions of the progress test in two settings and the implications for test deployment. Adv Health Sci Educ Theory Pract. 17:573–583.