2012; 34: 683–697 MEDICAL TEACHER

AMEE GUIDE

A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71

WILLIAM WRIGLEY, CEES PM VAN DER VLEUTEN, ADRIAN FREEMAN & ARNO MUIJTJENS

Department of Educational Development and Research, The Netherlands

Abstract

There has been increasing use and significance of progress testing in medical education. It is used in many ways and with several formats to reflect the variety of curricula and assessment purposes. These developments have occurred alongside a recognised sensitivity for error variance inherent in multiple choice tests from which challenges to its validity and reliability have arisen. This Guide presents a generic, systemic framework to help identify and explore improvements in the quality and defensibility of progress test data. The framework draws on the combined experience of the Dutch consortium, an individual medical school in the United Kingdom, and the bulk of the progress test literature to date. It embeds progress testing as a quality-controlled assessment tool for improving learning, teaching and the demonstration of educational standards. The paper describes strengths, highlights constraints and explores issues for improvement. These may assist in the establishment of potential or new progress testing in medical education programmes. They can also guide the evaluation and improvement of existing programmes.

Introduction

The introduction of problem-based learning (PBL) as a new educational philosophy in health sciences education began in the early 1970's in Canada at McMasters University and soon after at Maastricht Medical School in the Netherlands. This change brought the need for new methods to assess knowledge that were consistent with the PBL tenets of student-directedness, and deep and life-long learning, and which avoided the encouragement of rote and test-directed learning that were recognised to accompany traditional multiple-choice testing (van der Vleuten et al. 1996). This impetus resulted in the introduction of the progress test of applied medical knowledge in the late 1970s at both Maastricht University and the University of Missouri independently. Since then, it has been increasingly used in medical programs across the globe. A recent survey showed that this longitudinal, multiple choice question (MCQ) assessment tool has been introduced on all continents except Antarctica, involving such diverse regions as Southern Africa, Asia, several countries in Europe, the Middle East, North and South America, and in New Zealand and Australia (Freeman et al. 2010b).

For an assessment tool in medical education, the progress test offers some distinctive characteristics that set it apart from other types of assessment. It is usually administered to all students in the medical programme at the same time and at regular intervals (usually twice to four times yearly) throughout the entire academic programme. The test samples the complete knowledge domain expected of medical students on completion of their course, regardless of the year level of the student. The resultant scores provide longitudinal, repeated measures, curriculum-independent assessment of the objectives (in knowledge) of the entire medical programme. (van der Vleuten et al. 1996). These features enable the progress test to serve several important functions in medical programmes.

Considerable empirical evidence from medical schools in the Netherlands, Canada, United Kingdom and Ireland, as well postgraduate medical studies and schools in dentistry and psychology have shown that the longitudinal feature of the progress test provides a unique and demonstrable measurement of the growth and effectiveness of students' knowledge

Practice points

- Local, national and international progress testing is increasing worldwide.
- The inclusion of the progress test in the assessment regimes of medical and other health profession faculties offers several important advantages and benefits.
- A need for improved consistency and uniformity in progress testing can help improve the quality and defensibility of its data. This is suggested by evidence for significant error variance in multiple choice tests.
- Based on lengthy experience and empirical evidence from a review of the literature, a generic, systemic framework is presented in order to assist progress test practitioners to examine ways to improve consistency and uniformity.
- The strengths, constraints and issues of the parts of each component of the framework are examined and conclusions are drawn.

Correspondence: C.P.M. van der Vleuten, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel: +31 (0) 43 388 5726; fax: +31 (0) 388 5779; email: c.vandervleuten@maastrichtuniversity.nl

ISSN 0142–159X print/ISSN 1466–187X online/12/090683–15 © 2012 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.704437

acquisition throughout their course of study. (van der Vleuten et al. 1996; Boshuizen et al. 1997; Verhoeven et al. 2002b; van Diest et al. 2004; Dijksterhuis et al. 2009; Van der Veken et al. 2009; Bennett et al. 2010; Coombes et al. 2010; Finucane et al. 2010; Freeman & Ricketts 2010; Schaap et al. 2011)

As a result, this information can be consistently used for diagnostic, prognostic and remedial teaching and learning interventions. In the Netherlands, these interventions have been aided by the provision of a web-based results feedback system known as ProF (Muijtjens et al. 2010) in which students can inspect their knowledge level and growth (overall and in any subdomain) and compare it with the results of their peers.

Additionally, the longitudinal data can serve as a transparent quality assurance measure for programme reviews by providing an evaluation of the extent to which a school is meeting its curriculum objectives (van der Vleuten et al. 1996; Verhoeven et al. 2005; De Champlain et al. 2010). The test also provides more reliable data for high-stakes assessment decisions by using multiple measures of continuous learning rather than a one-shot method (Schuwirth 2007). Inter-university progress testing collaborations provide a means of improving the cost-effectiveness of assessments by sharing a larger pool of items, item writers, reviewers, and administrators. The collaborative approach adopted by the Dutch and other consortia has enabled the progress test to become a benchmarking instrument by which to measure the quality of educational outcomes in knowledge. The success of the progress test in these ways has led to the consideration of developing an international progress test (Verhoeven et al. 2005; Schauber & Nouns 2010).

The benefits for all stakeholders in a medical programme make the progress test an appealing tool to invest resources and time for inclusion in an assessment regime. This attractiveness is demonstrated by its increasingly widespread use in individual medical education institutions and inter-faculty consortia around the world, and by its use for national and international benchmarking practices. The progress test is currently used by national consortia in the United Kingdom (Swanson et al. 2010), The Netherlands (Schuwirth et al. 2010), in Germany (including Austria) (Nouns & Georg 2010), and in schools in Africa (Aarts et al. 2010), Saudi Arabia (Al Alwan et al. 2011), South East Asia (Mardiastuti & Werdhani 2011), the Caribbean, Australia, New Zealand, Sweden, Finland, UK, and the USA (Freeman et al. 2010b). The National Board of Medical Examiners in the USA also provides progress testing in various countries (De Champlain et al. 2010; International Foundations of Medicine 2011). The feasibility of an international approach to progress testing has been recently acknowledged (Finucane et al. 2010) and was first demonstrated by Albano et al (1996) who compared test scores across German, Dutch and Italian medical schools. An international consortium has been established in Canada (Finucane et al. 2010; International Partnership for Progress Testing 2011) involving faculties in Ireland, Australia, Canada, Portugal and the West Indies.

Despite its significance, its advantages for all stakeholders and its increasingly widespread use, evidence suggests that there is considerable variation in the content and application of the progress test (Ricketts et al. 2010). The blueprint and content sampling profiles can differ widely. Considerable divergence in test administration and composition can also be found, with testing repetitions varying between two to four tests per year, and the total number of items in a test differing between 100 and 250. There are also no accepted scoring and score calculation procedures, with differences evident in the inclusion of the 'don't know' option and formula scoring to prevent and correct for uninformed guessing, (McHarg et al. 2005) and using procedures, such as the cumulative deviation method for the analysis of results (Muijtjens et al. 2008; Schauber & Nouns 2010). Furthermore, there are differences in the purpose of the test as a summative or formative assessment, which can influence the student's test-taking attitude according to the outcome status of the test, thereby providing different results.

These variations are likely to have resulted in part from the differing availability of resources, institutional commitments and assessment programme designs, and are therefore not unexpected. They also reflect the widely acknowledged need for assessment practices to vary in order to accommodate and respect local conditions and issues (Prideaux & Gordon 2002; World Federation for Medical Examinaton 2003).

However, it is important to not confuse the accommodation of plurality in assessment approaches with the need for rigorous test uniformity and consistency that are prerequisites for achieving valid and reliable data. Indeed, the results of a recent study of assessments in the UK showed that unwanted outcomes can result from variations in assessment practices. McCrorie and Boursicot (2009) found in the clinical years of medical programmes across the UK that considerable variation in assessment processes made guarantees of minimum guidelines and formal quantitative comparisons of outcomes between medical schools questionable.

The need for improved consistency and uniformity in progress testing is also suggested by evidence that MCQ tests in medicine and the health sciences show considerable sensitivity for *"construct-irrelevant variance"* (Downing 2002). Questionable sampling procedures have been empirically found in which items were judged to reflect non-core medical knowledge (Koens et al. 2005). Frequent occurrences of flawed test items (Downing 2002; Jozefowicz et al. 2002; Downing 2005; Stagnaro-Green & Downing 2006; Tarrant et al. 2006; Tarrant & Ware 2008; Danish & Khan 2010), the use of imprecise terms (Holsgrove & Elzubeir 1998), item origin bias in progress test collaborations (Muijtjens et al. 2007), variation in test difficulty (van der Vleuten et al. 1996), and the influence of flawed test items in the outcome of high stakes examinations that lowered scores by up to 15% (Downing 2005; Tarrant & Ware 2008) have all been demonstrated.

It is to be expected that some variations in practice are inevitable and no assessment can be deemed perfect or completely free from error variance. However, achieving improved consistency and uniformity in progress test construction, content, administration, testing conditions, and scoring procedures in ways that are in line with the well-recognised testing guidelines of the American Educational Research Association (1999) are likely to help improve the quality and defensibility of progress test data.

This Guide describes an empirically-based, systemic framework for progress test practices and processes from which individual schools and consortia who have impending, new or

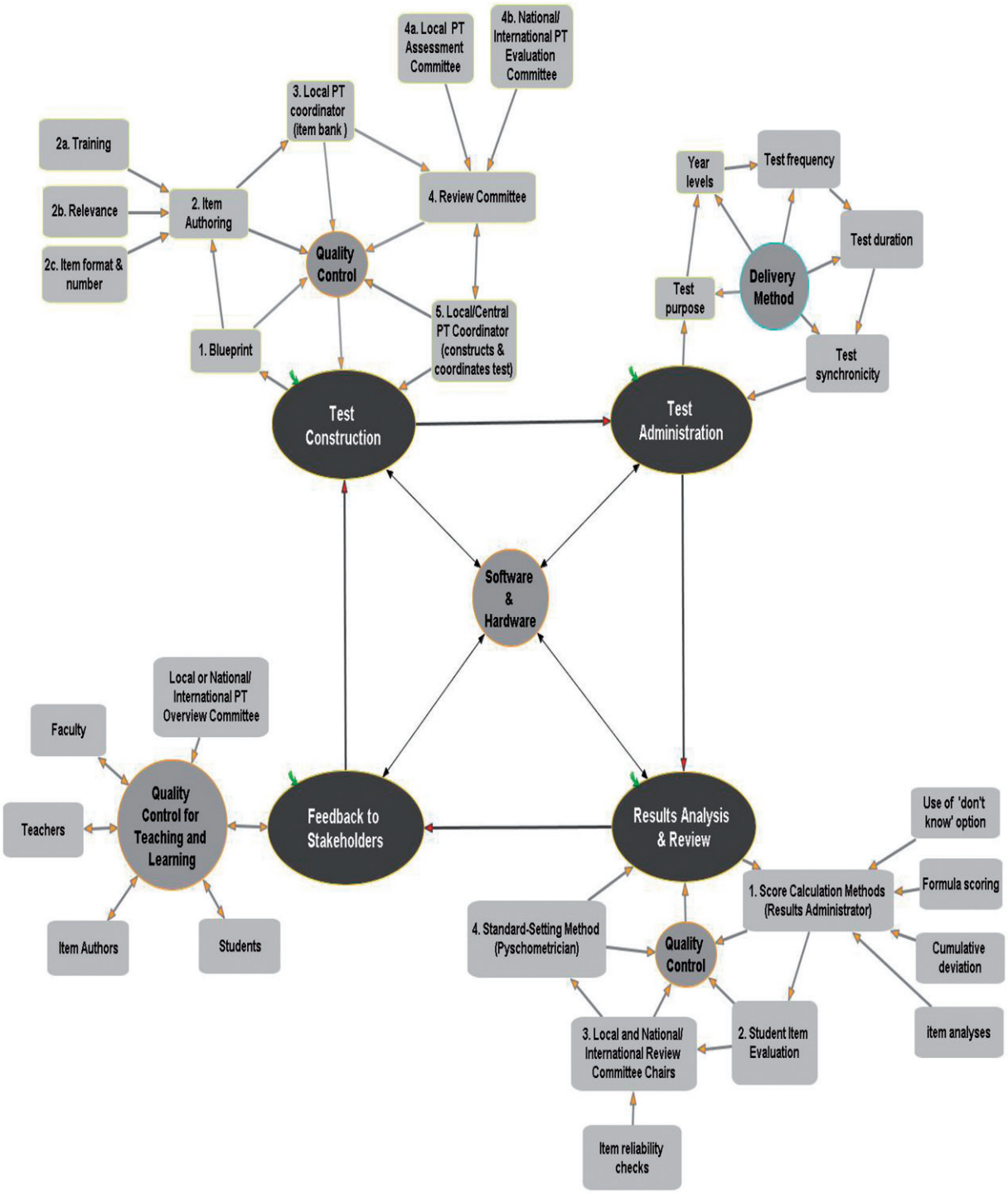

Figure 1. A generic systemic progress test framework.

existing progress testing can examine ways to improve consistency and uniformity. The framework was developed from the lengthy experience of progress testing in the Dutch consortium and Peninsula Medical School in the UK, and from examining the empirical literature on progress testing and multiple choice exams. The framework offers a systematic approach to identifying the strengths and basic requirements, constraints and issues in improving the validity, reliability and defensibility of progress test data. It is also hoped that the framework may provide a basis for the future development of consensually-determined principles of best practice.

Systemic progress test framework

Figure 1 displays the main components of a systemic progress test framework that has been generically adapted from the Dutch consortium system, Peninsula School and from a review

of the principles and processes described in the literature. This generic framework operates according to a systemic arrangement of several interacting parts that, as a whole, function cyclically to foster continuous quality control mechanisms.

There are four main phases of the framework, comprising test construction, test administration, results analysis and review, and feedback to stakeholders, with nested interactive parts pertaining to each, and quality control mechanisms operating as a central feature in three of the four components. Whether for an individual school, or a national or international progress test system, the review committee(s) and their associated local coordinator(s) (an additional central coordinator for consortia) play pivotal and recursive roles in the quality control procedures at the test construction, results analysis and review, and feedback phases.

Items for each test are drawn from the authors or item bank by the local or central coordinator according to a blueprint of content. They are then reviewed by a central committee and subsequently by the coordinator(s). Items are passed back and forth between committee and authors during this refinement process. A supra working group (the local or national/international progress test overview committee (in the Feedback to Stakeholders subsystem, see Figure 1) has the responsibility for ensuring the overall quality of the system, and has input through periodic test reviews and additional refinements of test construction and analyses processes.

In the Dutch consortium, unlike in most other institutions, students are also involved in quality checks by providing substantiated, post-test evaluative comments about the quality and accuracy of the test items which are then incorporated in the results analysis and review phase. This feedback has the advantage of helping to refine the item pool in the calculation of student scores and pass/fail standards. There are also learning advantages for students that come from their revision of test items and their required substantiated recommendations for changes. However, because this evaluation requires students to receive the test booklet and answers at post-test, new test items are required to be written for each test. The extra burden on resources and time this creates may mean that this component of the progress test system is not attractive for many Faculties.

Each part of the main systemic components of the framework is examined below in order to identify strengths, highlight constraints and describe areas for improvement which may assist in guiding the establishment or reviews of progress test systems. Also, in much the same manner as the WFME (2003) have accomplished with their international guidelines on assessment, the basic requirements of the system are identified where appropriate.

The components of the systemic progress test framework

Organisation

Although there are many individual schools worldwide that embrace the progress test, national and international inter-university consortia are becoming increasingly popular, in order to maximise their benchmarking, resource-sharing, and cost benefits. In regions where resources are limited, a consortium may be a particularly useful structural option. Experience from the Dutch and German consortia has suggested that a productive collaboration is likely to require a partnership agreement in regard to cost-sharing and funding arrangements, the use of data for research publications, as well as specific administrative, organisational and logistical arrangements (Schuwirth et al. 2010). An agreement that all partners pay fees to fund staff and infrastructure is also likely to be necessary (Nouns & Georg 2010).

Test construction

The test construction phase of the progress test system involves five main components, comprising the blueprint, item authoring, item bank, review committee and case administrator. Although these parts are numbered sequentially in Figure 1 to reflect the overall order of activities, in practice the maintenance of quality control often requires a frequent interaction and reciprocity between these elements.

Blueprint. The blueprint of knowledge classification is a basic and fundamental requirement on which the progress test relies for the valid and reliable construction of its content. The blueprint ensures adequate validity of and comparability for each test through representative and balanced sampling of the same content (Bridge et al. 2003). The Dutch consortium blueprint is described in Muijtjens and Wijnen (2010), and an example of another blueprint can be found in Swanson et al. (2010)

The blueprint contains the prescribed content for each test, usually according to a classification matrix of columns containing, for example, organ systems (respiratory, musculoskeletal etc) and skills (diagnosis, management etc), and rows containing disciplines (anatomy, surgery etc) or processes and tasks (mechanisms of disease, principles of therapeutics etc) (Coombes et al. 2010; Muijtjens & Wijnen 2010; Nouns & Georg 2010; Swanson et al. 2010). Each cell in this matrix contains the agreed frequency of items (questions) to be included in each test for that row x column combination. This weighting prescribes the importance or priority of the cells in terms of the end objectives of the educational programme in knowledge (Muijtjens & Wijnen 2010).

Some blueprints also specify the frequency of items for various levels of cognitive difficulty that the items test. It has been generally recognised that items need to be written so that they test the higher cognitive levels of knowledge application and problem-solving (Haladyna et al. 2002). Ware and Torstein (2009) have recently outlined five criteria for quality items that include at least 50% of items at a higher cognitive levels (application and reasoning). However, for the Dutch consortium the complexity of creating many tests has meant that it has been quite difficult to follow such a rigid formula. However, including guidelines for item frequencies according to cognitive level has been recognised as an area for further research and development.

Blueprint construction

The blueprint content is aligned to and reflects the end-of-programme learning objectives and competencies, usually according to those developed by national accreditation bodies, for example the CanMEDS in Canada (Frank 2005) Good Medical Practice in the UK (General Medical Council 2009), the Raamplan in the Netherlands (van Herwaarden et al. 2009), or those from the Accreditation Council for Graduate Medical Education in the USA (2012). It is important to note here that the blueprint is not directly aligned to a curriculum. Although the curriculum is naturally aligned with the end objectives, the progress test blueprint and curriculum are not directly related. This prevents having to change the blueprint with every curriculum reform.

Because there are a limited but representative number of items on a progress test, each row and column describing an aspect of the content domain is weighted in importance. Care is required to ensure that validity is not compromised by the under-representation of test content (Downing 2002). This can occur if the categories in the blueprint are too broad or ill-defined, insufficiently reflect the educational objectives, or if the distribution of question frequencies across categories are selected by less objective means.

Therefore, the blueprint construction requires two sequential decision processes – decisions about the content in the matrix and then decisions about the relative item weightings within each pertinent row x column cell of the matrix. Both these decisions are usually completed by expert consensus (Tombleson et al. 2000; McLaughlin et al. 2005; Coderre et al. 2009; Sales et al. 2010), for example by a Delphi procedure (Munro et al. 2000). The experts are most frequently asked to reach consensus on the weightings of the content based on the criteria of incidence and prevalence of clinical presentations, and these criteria may also be included in determining the item weightings in each cell.

There are two important factors to consider when selecting experts to seek consensus in constructing a blueprint - the experts' breadth of experience to ensure that it corresponds to the level of organisation that the blueprint will reflect (local, national, regional or global depending on the organisation), and the criteria they use to determine item weightings. The selection of experts with local or national experience to devise a blueprint for progress test organisations at these corresponding levels is an approach likely to yield acceptable ecological and content validity. An international blueprint will require the involvement of experts with international or regional practice experience to develop generic and culturally neutral classifications.

At Peninsula Medical School (UK), the content of the test material was blueprinted against a modified version of a Professional and Linguistic Assessments Board (PLAB) blueprint (Tombleson et al. 2000). The PLAB blueprint was derived from a study of the working activity of junior doctors in the UK (Freeman & Ricketts 2010). The Tombleson study used frequency of consultation rates and importance (not defined) as rated by specialists as its guiding criteria in selecting a list of *a priori* clinical problems appropriate for an OSCE test that had been compiled. The Dutch consortium ascertained the item frequencies for each cell of its discipline x organ system blueprint by determining the amount of written content in medical text books for each row or column and using expert consensus to translate these to the blueprint content. Other criteria have also been reported involving the distribution of patient age, gender, site of care, and diseases (Swanson et al. 2010).

Although there appear to be no studies that describe the construction of a blueprint for application across all year levels of an entire educational programme, a few studies have described the content selection or item weighting procedures for constructing a blueprint for the clinical years of a programme. (Munro et al. 2000; McLaughlin et al. 2005; Coderre et al. 2009; Sales et al. 2010) More research is needed to examine the efficacy of blueprint construction approaches.

Item authoring. Once the blueprint has been prepared, a range of expert and trained authors are required to prepare each item of the progress test according to the parameters of the blueprint and the content gaps in the item bank. Experience from the Dutch consortium has shown that there are several challenges in maintaining a flow of well-written items. Faculty and staff require motivation to frequently contribute questions on a continuing basis. It is often a challenging and demanding task to produce items that are not directly related to a particular unit or course but instead measure the ill-defined end objectives of the programme, and which should be accompanied by a literature reference. As a result, test items often have to be rejected because they are too detailed or not sufficiently relevant (van der Vleuten et al. 2004). Regular training sessions, updates, author-specific item analysis feedback that shows strengths and areas for improvement, and peer reviews all of which are undertaken in a rigorous and supportive environment (Malau-Aduli & Zimitat 2011) can help to improve item writer motivation and the quality of items.

There are several important aspects of the item authoring component of the framework that experience has shown need to be addressed, namely, author training, item relevancy, guessing and item format and number.

Training

Writing high quality MCQ items is a complex and time consuming task. A useful taxonomy of item writing rules is available (Haladyna & Downing 1989a,b), and advice to help rectify them has been well documented (Case & Swanson 2002; Haladyna et al. 2002). Frequent occurrences of item writing flaws have been empirically found (Downing 2002; Stagnaro-Green & Downing 2006). This means that item writers require ongoing and periodic training to not only ensure continual improvement in writing quality items but also to help reduce the time and cost of prolonged quality control checks resulting from the production of flawed items. Evidence has demonstrated the efficacy of both peer review and structured training in improving item writing quality (Josefowicz et al. 2002; Wallach et al. 2006; Malau-Aduli & Zimitat 2011; Naeem et al. 2011). Author training is therefore a

vital quality control measure that is a basic requirement of a quality progress test system.

Relevance

The relevance of items for testing a new graduate's knowledge is a potential source of error that can compromise progress test validity. Each item needs to be relevant according to the core knowledge required at graduation to meet end of curriculum objectives (Koens et al. 2005). Downing (2002) has maintained that the inclusion of trivial questions that have a low degree of importance for future learning or clinical care have too often appeared in medical education tests of achievement. This may have resulted from unclear criteria for the definition of relevance to guide item writers.

Past experience from the Dutch consortium has found that it can be quite difficult to obtain a consensus agreement among item writers on the precise definition of relevance, and that the concept can be mistakenly confused with item difficulty. As a result, in more recent times consensual expert agreement among Dutch consortium partners has been used to develop five criteria of relevance in order to encourage more consistency and accuracy in its interpretation among item authors when constructing and reviewing items. It was agreed that the items should test knowledge that is specific to the specialty of medicine, test ready knowledge (knowledge required as a prerequisite to function in a practical situation), be important knowledge which is required for the successful practice of medicine, have a practical relevance for the successful handling of high-prevalence or high-risk medical situations, and the knowledge should form the basis of one or more important concepts of the curriculum (Schuwirth 2011). Further research is required to clarify the impact of relevance on the student progress test outcomes, and to examine its meaning in regard to the knowledge required of a new graduate.

Item format, guessing and item number

Format

Several multiple choice item formats have been described in the literature, including the one-best-answer, alternative-choice, true-false, multiple true-false, extended matching and complex multiple choice (Case & Swanson 2002; Haladyna et al. 2002). Each of these involves a different option format and number. It has been argued that true-false questions are not very reliable (Zimmerman & Williams 2003), and in comparison, the single best answer option provides more reliable scores and a lower guessing probability (Muijtjens & Wijnen 2010). Rademakers et al. (2005) found that with a limited number of students and questions, a short answer progress test was also reliable and feasible.

In the Dutch consortium test the number of options selected for each item varies between two and five, with most having three or four options while the progress test at Peninsula has consistently used 5-option items. The variation in option numbers has the advantage of constructing as many alternatives as are relevant for a particular question, and of not being forced to include less appropriate alternatives that are easily recognised as being incorrect, even by students lacking knowledge of the item. However, Peninsula have found that consistently using a 5-option test provides a constant subtracted mark of –0.25 for an incorrect answer thereby removing the need to regularly alter the rubric in programmes that mark automatically.

Guessing and the use of 'don't know' option

The use of the 'don't know' option among the alternatives of an item in combination with a penalty for guessing has been used variably across medical schools, and as a result, can contribute to significant discrepancies in outcomes between institutions. It is included in the progress tests of the Dutch, German and Canadian consortia's and at Peninsula to reduce the frequency of guessing as well as to reduce the influence of guessing on the score. However, in the UK and some USA consortia tests it is not included. It has also received mixed research support. From their calculations, Zimmerman and Williams (2003) found that error from guessing could be larger than for other sources of variance for many MCQs. Muijtjens et al (1999) have argued that because the progress test measures achievement of the curriculum end-objectives repeatedly and longitudinally throughout the curriculum with all students in the programme, *"a don't-know option is inevitable because it expects students not to have covered all the objectives assessed in the test"* (p. 268). The results of a recent study by Wade et al. (2011) provide empirical support for the use of methods to control for guessing. They examined students' perceptions of the progress test and found that students believed that luck and guessing were stronger contributors to their success on the progress test than their knowledge. That belief was significantly stronger in a school that did not use a penalty for guessing.

Furthermore, it has been argued that omitting the 'don't know' option and using number correct scoring gives equal reward for uncertainty as for confident knowledge (Burton 2002). Also, the use of the don't know option provides students with a measure of the limits of their knowledge by showing them what they don't know, while discouraging the practice of guessing in a way that emulates the requirements of real medical practice (Burton 2002).

However, some have argued that including the 'don't know' option introduces measurement error by discriminating against those students whose personalities encourage them to adopt a risk-averse response set. However, mathematical analysis has suggested that this effect is small compared with the measurement error of guessing that it is designed to prevent (Espinoza & Gardeazabal 2010). A further disadvantage may apply to the early years of the programme for whom the achievement of a very high frequency of don't-knows, although to be expected, may have the unwanted negative result of lowering confidence in their progressive development of knowledge and therefore their motivation.

From this outline of the arguments and empirical evidence for and against the inclusion of the 'don't know' option, it would appear that it may be more efficacious to include the

option. However, further research is required to provide evidence to support this conclusion.

Item number

There is considerable variability in the number of items selected for the test. The Dutch and German consortia select 200, Canada, 180, and schools in the United Kingdom 120 or 125 items. This variation may be a function of the test frequency per year. However, some research has suggested that tests that are too short can reflect under-represented content, thereby challenging the content validity of the test and support for the legitimacy of its inferences (Downing 2002). Consistent with this, a fairly large number of items have been recommended (Langer & Swanson, 2010). An analysis of the interaction effects on reliability between item number and test frequency is presented below (see section "Year level, test frequency and test size" below).

Item bank. All progress test systems require an item bank, and an administrator to securely house, classify, check, and review incoming items and existing stocks in accordance with the blueprint. For consortia, the Dutch experience has shown that a central item banking system and a central coordinator is necessary to maintain an adequate and secure quality control system among its members. The administrator also helps to keep the bank up to date by notifying the local committee chair(s), and, where necessary, item authors of the types and frequencies of items required. Agreement is required in regard to the blackout or retirement rule, for example one to five years, to avoid students encountering an item they may be familiar to them from a previous test.

Dedicated IT hardware and software resources, support and coordination, whether locally or online, are required for a secure and well-functioning item bank for both the item writer and administrator. An advantage for a consortium is that these arrangements become more cost-effective.

The size of the item bank will be influenced by the frequency of the test in the academic year, the number of items in the test, the reusability policy, and whether the students are given the test booklet and answers at post-test. These factors will also determine the frequencies of new items that are required for the bank. As an example, the Dutch consortium has a policy of item reuse after the elapse of three years. This means that with 200 items required for each test four times per year, the item bank needs to contain at least 2400 items ($3 \times 4 \times 200$).

Review committee and local/central coordinator(s). An iterative process of item checking and review is an important quality control feature of the progress system. This involves checks that each item is up-to-date with the current literature, is relevant, consistent with the blueprint, and that each is free of specific item writing flaws (Haladyna et al. 2002). In the progress test framework described here, these checks are overseen during the test construction phase by a local review committee, and for consortia, also by a national or international review committee. The Dutch consortium comprises four local and one national review committee, each consisting of approximately six members who have backgrounds in the basic, clinical and behavioural sciences.

Each of these committees has a designated chairperson and associated coordinator to organise the item authoring, undertake quality checks, construct a preliminary set of items for a test from the item bank for the committee to review, complete editorial checks of the final set of test items, and construct the test booklet.

For these arrangements to work well, an individual institution or consortium might establish a central test organisation to coordinate and safeguard the quality of item preparation, test construction and administration (Muijtjens & Wijnen 2010). For the Dutch consortium, working conferences and workshops were provided to train the local committees to help establish a production cycle of high quality items that were acceptable to all partners (van der Vleuten et al. 2004).

Test administration

There are several features of the test administration component of the progress test framework that have significant bearing on the outcome of the test results. These involve the purpose of the test, the year levels included in the test, the test delivery method, and the test frequency and duration. Figure 1 shows the test delivery method as a central organising feature of the test administration subsystem. The choice of whether the delivery method is computer or paper-based will dictate the guidelines for the other components of the system.

Test purpose. The purpose of the progress test as a formative or summative test is variably used across institutions. In the Dutch consortium and at Peninsula it is summative whereby students are required to pass the test, based on the aggregated results of all instances of the test (4) in a year, in order to progress to the next year level. However, in the German and Canadian consortia, a formative approach is adopted in which the main focus is on the results providing feedback for the students' learning. The choice of a formative test is often influenced by the presence of external assessment, such as a national licensing exam, or by internal grading policies. The Dutch consortium selected a summative approach in order to encourage students to respond to the items as a high stakes assessment and thereby to stimulate deep and continuous learning. This approach also includes a strong formative focus through the provision of extensive feedback to students and other stakeholders (see section on Feedback to Stakeholders). Experience shows that whether a combined or singular purpose is chosen, the quality and detail of the test feedback should not be affected.

Test synchronicity

Synchronised testing, whether paper- or computer-based in which test administrations occur on the same day at the same time for all student cohorts in an institution, or with all participating institutions in a consortium, is required in order to benchmark scores among the consortium partners. The Dutch consortium uses paper-based synchronised testing and has found the benchmarking benefits to outweigh the

disadvantages. Synchronised testing produces logistical and resource pressures which may be alleviated by computer-based testing to some degree. However, other difficulties also arise with computer testing. This arrangement usually requires students to be tested in batches because of limited space or computer access. Under these circumstances, several versions of the same test are required to prevent information being exchanged between batches of students. This necessitates either a significant increase in the number of items required in the item bank or a sufficiently large item bank from which to draw items. The Dutch experience has shown that the volume of item production can become problematic in the item construction process. The versions of the test also need to be of the same difficulty and require stringent psychometric calibration which may be a daunting task for a consortium with limited resources.

Synchronised testing is less important with systems that choose a formative approach, such as the German consortium (which includes Austrian institutions) where members administer the test within one to two weeks of each other. In this context, the motivation for students to cheat, although not removed, is greatly diminished because the formative nature of the test means there is nothing to gain for them by cheating, and only interferes with the potential formative gains for the individual student.

Some schools administer different test forms to students at the same time (Swanson et al. 2010). Variations in difficulty have been recognised as an issue for this delivery form. The use of the equated scores method, which adjusts for differences in difficulty between different test forms has been discussed to address this drawback (Langer & Swanson, 2010).

Computerised adaptive testing, in which students are administered test questions that are matched to their performance level, may also reduce this pressure further and remove the need for synchronicity. However, other constraints arise with this approach. For example, all items need to be pre-tested and calibrated which can place considerable pressure on item authoring requirements.

Year level, test frequency and test size. Although there is some variation, it seems that most schools around the world include all year levels of their programme in the progress test (Freeman et al. 2010b). The Dutch and Peninsula Schools test students at all levels of the programme (years one to six and one to five, respectively).

However, the frequency of progress testing within the academic year, as with the number of items selected for the test (see subsection above "Item Number") varies considerably, usually between twice (German consortium, (Schauber & Nouns 2010)) and four times (for example, the Dutch consortium and schools in the UK).

Although there are no fixed guidelines for choosing the frequency and number of items in the progress test, and factors, such as cost and the availability of resources are influential considerations, the examination of test reliability is a useful and important guide in helping to determine the selection of test size and frequency in a progress test system. To this end, a recent generalizability analysis examining the combined effect on reliability of item number per test and test frequency was undertaken using the total scores (correction scoring) in the four progress tests of the academic year 2010/2011 obtained by Maastricht University students in each of the six years of the programme.

The design of the analysis involved three sources of variation indicated as p (*persons,* that is, students), m *(measurement occasions)*, and $i{:}m$ *(items* nested within *measurement occasions)*. The term nested indicates that per *measurement occasion* a different set of *item*s is used. *Person* is the subject of measurement and, hence, is associated with the variance of interest. *Measurement occasion* and *items within measurement occasions* represent two facets and the corresponding variances contribute to measurement error. The total variance V_{tot} is defined as:

$$V_{tot} = V_p + V_m + V_{i:m} + V_{pm} + V_{pi:m}$$

where $V_p, V_m, V_{i:m}, V_{pm}$, and $V_{pi:m}$, represent the variance components of the main effects of *persons*, *measurement occasions*, and *items within measurement occasions*, and the interaction effects of *persons* and *measurement occasions*, and *person*s and *items within measurement occasions*, respectively. The generalisability coefficient G is defined as:

$$G = \frac{V_p}{V_p + V_{pm}/N_m + V_{pi:m}/(N_m \times N_{i:m})}$$

where N_m and $N_{i:m}$ are the numbers of *measurement occasions*, and *items within measurement occasion*, respectively. In the first step of the generalizability analysis, the G study, the above variance components are estimated on the basis of the available data. In the second step, the D study, these variance estimations are used to predict the generalizability G by substituting varying hypothetical values for N_m and $N_{i:m}$intheexpression above.

Table 1 presents the relative contribution of each of the five variance components to the total variance. The number of students per year level is indicated in the lower part of the table. The estimation method allows only complete cases (students with data from four measurement occasions for the same year level) to enter the analysis; the corresponding proportion of included cases is indicated in the last row. The error term variance $V_{pi:m}$, that is, the interaction effect of persons and items, is by far the largest (72%–78%), followed by $V_{i:m}$, the item main effect variance (20%–27%), V_p, the person variance (0.7%–1.5%)V_m, the main effect of measurement occasion (0.2%–1.1%), and finally V_{pm}, the interaction effect of person and measurement occasion (0.08%–0.24%).

Table 2 shows the corresponding values for the generalizability coefficient G calculated for the indicated combinations of values for test frequency (N_m),andtestsize($N_{i:m}$). As expected, the general pattern shows that test reliability increases with increasing frequency and test size. The results in Table 2 also indicate that, given a fixed total amount of items available for testing in an academic year, reliability becomes more favourable with an increase in the frequency rather than the test size. For example, the reliability coefficients for Year 1 in Table 2 show that two tests of 200 items produced a reliability of 0.70, while four tests of 100 items achieved 0.74. This is not surprising when inspecting the equation for G: when N_m is increased while keeping the total amount of items

Table 1. Percentage of total variance for each variance component in each year level for Maastricht University students Years 1-6 in the academic year 2010/11.

	Year level					
	1	2	3	4	5	6
Variance component	% of total variance					
Vp	0.73	0.97	0.92	1.47	1.30	1.26
Vm	1.07	0.70	0.65	0.44	0.47	0.15
Vi:m	20.06	23.73	23.26	23.78	25.91	26.61
Vpm	0.24	0.19	0.13	0.11	0.09	0.08
Vpi:m	77.90	74.42	75.04	74.20	72.24	71.89
Number of Students	252	239	268	133	177	112
% included	71	76	85	42	55	33

Table 2. G coefficients for test size (number of items) by test frequency for Maastricht University students Years 1-6 in the academic year 2010/11.

		Year 1								Year 2							
		Test size								Test size							
		25	50	75	100	150	200			25	50	75	100	150	200		
Test Frequency	1	0.18	0.29	0.36	0.42	0.49	0.54	Test Frequency	1	0.23	0.37	0.45	0.51	0.59	0.63		
	2	0.30	0.45	0.53	0.59	0.66	0.70		2	0.38	0.54	0.62	0.67	0.74	0.78		
	3	0.40	0.55	0.63	0.68	0.74	0.78		3	0.48	0.63	0.71	0.76	0.81	0.84		
	4	0.47	0.62	0.70	0.74	0.79	0.82		4	0.55	0.70	0.77	0.81	0.85	0.87		
		Year 3								Year 4							
		Test size								Test size							
		25	50	75	100	150	200			25	50	75	100	150	200		
Test Frequency	1	0.23	0.36	0.45	0.51	0.59	0.64	Test Frequency	1	0.32	0.48	0.57	0.63	0.71	0.76		
	2	0.37	0.53	0.62	0.68	0.74	0.78		2	0.49	0.65	0.73	0.78	0.83	0.86		
	3	0.47	0.63	0.71	0.76	0.81	0.84		3	0.59	0.74	0.80	0.84	0.88	0.90		
	4	0.54	0.69	0.77	0.81	0.85	0.88		4	0.66	0.79	0.84	0.87	0.91	0.93		
		Year 5								Year 6							
		Test size								Test size							
		25	50	75	100	150	200			25	50	75	100	150	200		
Test Frequency	1	0.30	0.46	0.55	0.62	0.70	0.74	Test Frequency	1	0.30	0.45	0.55	0.61	0.69	0.74		
	2	0.47	0.63	0.71	0.76	0.82	0.85		2	0.46	0.62	0.71	0.76	0.82	0.85		
	3	0.57	0.72	0.79	0.83	0.87	0.90		3	0.56	0.71	0.78	0.82	0.87	0.89		
	4	0.64	0.77	0.83	0.87	0.90	0.92		4	0.63	0.77	0.83	0.86	0.90	0.92		

$N_m \times N_{i:m}$ constant, the error-term $V_{pi:m}/(N_m \times N_{i:m})$ does not change, but the error-term V_{pm}/N_m decreases, allowing the reliability to increase. So, for the sake of reliability these results suggest that it is better to have more test occasions and a smaller test size than the other way around. Of course there are practical considerations of cost and resource availability that prevent Maastricht University from following this principle to its utmost consequence.

The upper left panel of Table 2 shows that for a reliability level of 0.80 in Year 1, four test occasions per academic year and 200 items per test is required which corresponds to the frequency and test size currently used by the Dutch consortium. This frequency and test size also produces reliabilities greater than 0.90 for the higher years (four to six). At first glance this might seem unnecessarily high. However, providing reliable feedback to students, particularly from sub-domain scores, has been a highly valued purpose of the Dutch consortium progress test system. These sub-domains (for example, respiratory system, blood and lymph system, digestive system) are generally represented in the test with less than

25 items each. An inspection of the reliability coefficients in Table 2 across all years for tests containing 25 items demonstrates that in order for sub-domain scores to reach an acceptable level of reliability, a frequency of four occasions per year is not overdone. Furthermore, experience has shown in the Dutch consortium that frequent measurement of all year levels during the academic year of a programme helps maximise the reliable tracking and monitoring of students' developmental and longitudinal progression.

The results of Wade et al.'s (2011) study have provided further empirical support for testing students four times per year. Their results showed a positive impact on the perceived value of the test by the students. This positive perception is also likely to encourage deep approaches to learning.

Test duration. The duration allowed to complete each test varies worldwide between 2.5 hours (UK) to five hours (NBME, USA). The duration of the test will significantly depend on the number of items in the test and the reading time required for each item. The Dutch consortium has found that it is important to find a balance between discouraging students from guessing and being consistent with the underlying principle that the progress test is not a speeded test. Experience has shown that a useful time permitted for answering a single item is approximately 75–85 seconds. Although the number of items chosen for a test will be prescribed by the blueprint, to achieve this recommendation, it is important to ensure that the test does not consist of too many items that take a lengthy time to read.

Result analysis and review

The third phase of the progress test system involves the analysis and review of the test results. There are several important features of this part of the framework that require careful consideration and expertise in order to produce reliable and defensible measures of students' progress. These involve selection of the calculation methods for the scores and standards, the skills of the local or national review committee, and the inclusion of students' item evaluations.

Score calculation method. The total score a student achieves on an MCQ test is significantly influenced by the way in which it is calculated. The main two score calculation methods used are scores based on the total score, either number correct or correction scoring. Some consortia calculate scores according to the total number correct (Swanson et al. 2010). However, others have argued that this method does not account for guessing to the extent that if no penalty is imposed and students have no knowledge of the subject matter being tested, they will receive an average score of 100/A, where A is the number of choices per question (Scharf & Baldwin 2007).

Correction or formula scoring has been used to control for the measurement error arising from guessing that is not taken into account with the number-correct scoring. In order to dissuade students from engaging in this form of error variance, the Dutch consortium applies a penalty for incorrect scores whereby fractional points for incorrect responses, depending on the number of options for the question, are subtracted from the correct score.

When guessing, the chances of achieving a correct answer are smaller than for choosing an incorrect response. For example, the chance of guessing a correct answer to a four-option question is 25% and for an incorrect answer, 75%. Therefore, in the pursuit of fairness, the size of the deducted fractional score for an incorrect answer is calculated by dividing one by the number of incorrect options. This means that for an incorrect item, a score of one is deducted from the total score for an item with two options (that is, $1 \div 1$), -0.5 ($1 \div 2$) for three-option items, -0.33 ($1 \div 3$) for four-option items and -0.25 ($1 \div 4$) for five-option items. This method relies on the inclusion of the 'don't know' option in all test items, and responses to this option are given a score of 0.

There have been persuasive arguments and evidence provided for formula scoring. It is particularly consistent with the underlying philosophy of progress testing. The 'don't know' option provides direct information for students and teachers in determining gaps in knowledge in order to promote learning. It's use also reinforces the view that the admission of ignorance is preferable to guessing (Tweed & Wilkinson 2009), and is appropriate for longitudinal tests such as the progress test in which many items are too difficult for the abilities of students in the lower years (McHarg et al. 2005). Formula scoring has been used with good effect in individual institutions (Freeman et al. 2010a), and as a method of inter-institutional comparison (Schauber & Nouns 2010) together with the cumulative deviation method (Muijtjens et al. 2008; Schauber & Nouns 2010) to reveal stable between-school differences. Error variance resulting from guessing has also been found to be larger than other sources of error variance (Zimmerman & Williams 2003), and formula scoring has been shown to be more reliable than number-right scoring (Muijtjens et al. 1999; Alnabhan 2002).

However, the use of formula scoring in score calculation methods has been a debated issue for some time. For instance, it has been argued that it may add error variance because the correct minus incorrect score includes irrelevant measures related to test-taking attitude (Downing 2003). Also, in some quarters it has been interpreted that applying a penalty to an item results in the questionable practice of removing a mark already gained from another item.

The score calculation method is not an easy choice and one that is likely to be influenced by the practicalities of cost and availability of resources. Indeed the issue may be alleviated, if or when computer-adapted testing is used for progress testing in which questions are tailored to the ability of the student (Roex & Degryse 2004). With this approach guessing becomes less of an issue and the inclusion of the 'don't know' option becomes superfluous. At this stage there are no published studies that report the use of computer adaptive testing for a medical progress test that might assist in determining the efficacy of this approach.

Standard-setting method. The selection of a standard-setting method to determine pass/fail cut scores and other grades is the final step in the results analysis and review process. It is

usually the case that the higher the stakes in the consortium, the stronger the requirements become for standard-setting. Various approaches have been used for progress testing to determine the cut scores (Verhoeven et al. 1999, 2002a; Ricketts et al. 2009; Ricketts & Moyeed 2011) and there is a vast literature describing these (Bandaranayake (2008) for a useful overview of commonly used methods, and Downing et al. (2006) for an overview of setting absolute standards).

The merits between norm-referenced and criterion-referenced methods, two of the most commonly used, are controversial. Each has its advantages and disadvantages, and each result in varying drains on resources. Muijtjens et al. (1998) found that because of the variation in progress test difficulty, using a fixed, absolute cut off score was more precarious than norm-referenced scores.

The Dutch and German consortia have relied on norm referencing calculations to determine cut scores (Muijtjens et al. 2008; Schauber & Nouns 2010). Although a more rigorous and commonly used method may be more preferable and defensible, such as the Angoff process

(Muijtjens et al. 1998; Verhoeven et al. 1999; Basu et al. 2004) which uses the agreement between expert judgements of a minimally competent performance to determine standards, the higher costs involved in such a procedure for each of the tests per year prevents the use of this method. An interesting standard-setting variant that may be an alternative to norm referencing is one recently demonstrated by Ricketts et al. (2009). They have shown the usefulness of triangulating standard-setting data across a number of internal sources involving student test results and an external source of data from newly qualified doctors.

Student item evaluation. In the Dutch progress test system, students are given the test booklet and correct answers without explanation at the completion of each test to take home so that they can provide critical, substantiated feedback (Muijtjens et al. 2010). This provides two valuable advantages. It offers an important quality control mechanism by aiding the removal of flawed items during the post-test review analysis of the test results before the final calculation of the standards. It also encourages students' deeper learning by encouraging them to review, confirm or correct knowledge.

A variant of this practice has been described by Kerfoot et al. (2011) in which student reviews of the questions and answers were cycled over spaced intervals of two and six weeks to improve long-term retention rates. Their results showed that longer-term retention of core knowledge was more than doubled by this method.

Although there are demonstrated learning benefits of student post-test reviews of the progress test content, it is not a common practice among institutions, mainly because of the disadvantage that the test items do not remain secret, thereby reducing the utility of the item bank and placing extra demands on resources from the increased need for new items to be written for each subsequent test iteration.

Local and national/international chair review committee. An important principle to enhance content validity and reliability of progress testing is the provision of expert, post-test quality control reviews of test items. This is consistent with the recommendation that *"the ability of a course director to demonstrate this review process, including the recommendations of the experts and the actions taken on those recommendations, is a key factor in assuring content validity"* (Bridge et al. 2003, p. 415). In the framework presented in this Guide, this quality control mechanism is reflected in the system of post-test reviews by the progress test review committee, or in the case of a consortium, by the chairs of the local and national committees. The post-test review committee(s) can review all items, and by consensus decide which items will be included in or withdrawn from the final analyses to determine the pass/fail standard. They can also identify questions that have not performed well and feed them back to the item review committee for change or rejection from the bank. This information can also be passed to the item author as part of the feedback procedures of the framework.

An adaption of the review committee membership can be to include doctors in the review panel who regularly work with newly qualified doctors, thereby more closely matching the progress test principle of measuring the knowledge level required at the end point of a curriculum, that is, of a newly qualified doctor.

Feedback to stakeholders

An underlying principle of the progress test is its utility in providing developmental and longitudinal feedback to students in order to aid deeper learning. The test results can also offer valuable quality control information for item authors, teachers, faculty, and the progress test overview committee.

Students. The progress test system of repeated, reciprocal cycles of knowledge testing, feedback to students, and consequent student-directed learning can help to enhance learning (Norman et al. 2010; Ricketts et al. 2010; Kerfoot et al. 2011). These findings are consistent with research showing that repeated testing that encourages retrieval practice can promote learning, retention and transfer of knowledge (Roediger & Karpicke 2006; Carpenter et al. 2008; Larsen et al. 2009; Butler 2010; Roediger & Butler 2010).

Detailed feedback of student results in the Dutch consortium is provided through graphical, query-based online information from the ProF system (Muijtjens et al. 2010). Figure 2 shows an example of the ProF feedback of total scores using correction scoring (dark line) across 24 consecutive measurement moments (four per year across the 6-year program) for a student in Year 6 of the Maastricht University program compared with all peers in the same year in the Dutch consortium (white line).

This web-based tool gives teachers and students displays of scores and patterns of knowledge growth according to various parameters, including the content of specific sub-domains (such as the respiratory system or the discipline of anatomy), cumulative scores of average knowledge growth, group comparisons with peers at each or across year levels, as well as benchmarks against national outcomes (Muijtjens & Wijnen

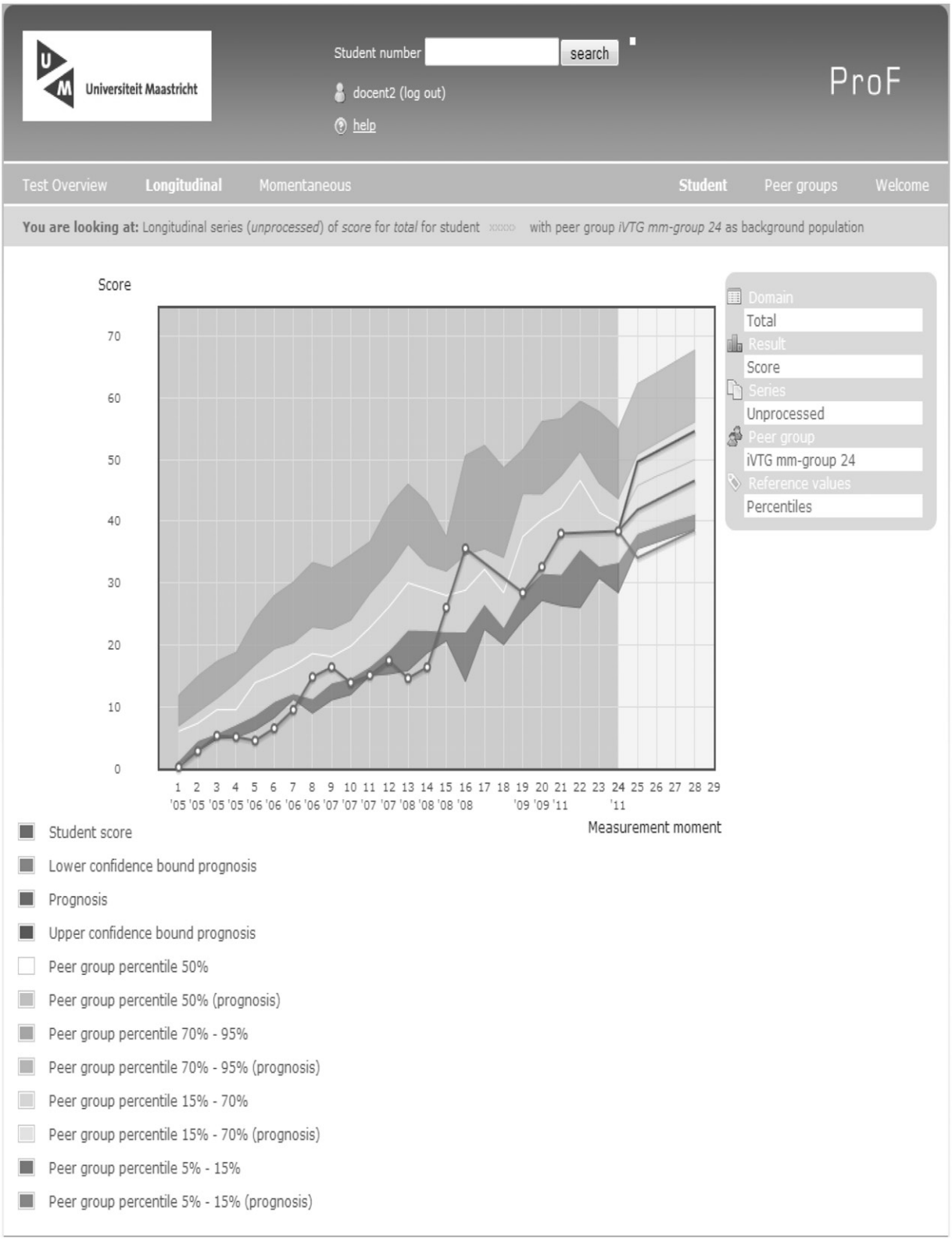

Figure 2. Longitudinal view of total scores on 24 consecutive measurement moments for a Maastricht University student compared with all peers from universities in the Dutch consortium.

2010). Similar methods have been reported in the UK (Coombes et al. 2010).

Although this rich source of feedback is readily available for students, more research is needed to determine the impact of students' usage of feedback on their progress test scores and other curriculum assessments. Recent educational moves have been made by Maastricht University, a partner of the Dutch consortium, to provide more scaffolding for students through

the mandatory exploration and reflection on their scores with their mentor to encourage them to more effectively examine their progress scores in order to aid deeper learning.

Item author, teacher and faculty and overview committee. Progress test scores are also an important source of information for item authors, teachers in the programme, faculty and the overview committee who has responsibility for the overall functioning of the progress test system in order to foster quality control mechanisms. For example, providing the item authors with reliability scores for the specific items they constructed over several tests can be useful information in assisting them to determine their strengths and help recognise and address their weaknesses. This has been an indentified improvement required in the Dutch consortium as part of improving the quality control mechanisms in its system. Feedback is also useful to teachers to assist with diagnostic, prognostic or remedial interventions, for curriculum development and quality assurance for faculty, and for guiding improvements in the progress test system for the overview committee.

Conclusions

Over its 25 year history, the Dutch consortium has proceeded through several transformations of various aspects of the framework as improved methods and approaches have been developed or researched. This has helped to maintain a robust and systemic framework within which to maintain improved quality control mechanisms. As a new school, Peninsula was able to draw on best evidence in its 10 year history of progress testing (Freeman & Ricketts 2010) resulting in a relatively stable format. However changes have occurred, such as increasing feedback methods and adaptations of standard setting formulae to reflect the moving environment of medical education.

This explication of the various component parts of the generic systemic framework provides a means for an evidence-based evaluation of existing progress test arrangements. It can be used to not only identify strengths and areas for improvement for impending, start-up, new, or developed progress test systems, but also to help guide a strategic progress test plan for effective change.

The examination and implementation of the many and varied parts of the systemic framework presented here to provide a quality-controlled, repetitive and longitudinal progress test will be influenced by and proportional to the scope of the progress test system, as well as curriculum demands and the internal and external assessment and grading policies and rules within which the system must operate.

The present analysis of the framework also shows that its implementation and maintenance requires an institutional commitment and the availability of resources to ensure it promotes satisfactory levels of test validity and reliability. The basic requirements of a quality system show that a blueprint requires development and updating by reviewers as curricula mature, and considerable efforts are required to provide ongoing item writing training in order to produce interdisciplinary, contextualised and relevant items (Vantini & Benini 2008). The analysis and review of results requires several quality control checks, feedback to stakeholders requires analysis and monitoring, and a commitment to software and hardware funding and support are important features of a successful operation.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

WILLIAM (BILL) WRIGLEY, MA (Clin Psych), PhD, is Project Leader and a member of the assessment task force in the School of Health Professions Education at Maastricht University. He has responsibility for implementing the progress test system in projects involving universities in Saudi Arabia, and works in the Netherlands progress test consortium. He has also worked in assessment, teacher training, and PBL tutoring in medical education at universities in the Netherlands and Australia.

CEES PM VAN DER VLEUTEN, PhD, is Professor of Education, Chair of the Department of Educational Development and Research, Scientific Director of the School of Health Professions Education (SHE) at Maastricht University, Maastricht, The Netherlands. He holds honorary appointments in the University of Copenhagen (Denmark), King Saud University (Riyadh) and Radboud University (Nijmegen).

ADRIAN FREEMAN MMedSci FRCGP, is Associate Professor & Director of Assessments at Peninsula Medical School, UK. Plymouth University

ARNO MM MUIJTJENS, MSc, PhD, is Associate Professor at the Department of Educational Development and Research, Faculty of Health, Medicine, and Life sciences, Maastricht University. As a Statistician – Research Methodologist, he contributes to the research programme of the School of Health Professions Education, and as a member of the assessment task force he is involved with projects regarding research and development of progress testing.

References

Aarts R, Steidel K, Manuel BAF, Driessen EW. 2010. Progress testing in resource-poor countries: A case from Mozambique. Med Teach 32:461–463.

Accreditation Council for Graduate Medical Education. 2012. Retrieved 24 January, 2012. Available from http://www.acgme.org/acWebsite/home/home.asp.

Al Alwan, I, Al-Moamary, M, Al-Attas, N, Al Kushi, A, ALBanyan, E, Zamakhshary, M, Al Kadri, HMF, Tamim, H, Magzoub, M, Hajeer, A et al., 2011. The progress test as a diagnostic tool for a new PBL curriculum. Educ Health. 24:493. Epub.

Albano MG, Cavallo F, Hoogenboom R, Magni F, Majoor G, Manenti F, Schuwirth L, Stiegler I, van der Vleuten C. 1996. An international comparison of knowledge levels of medical students: The Maastricht progress test. Med Educ 30:239–245.

Alnabhan M. 2002. An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric properties of a test. Soc Behavior Personality 30:645–252.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 1999. Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bandaranayake RJ. 2008. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach 30:836–845.

Basu S, Roberts C, Newble DI, Snaith M. 2004. Competence in the musculoskeletal system: Assessing the progression of knowledge through an undergraduate medical course. Med Educ 38:1253–1260.

Bennett J, Freeman A, Coombes L, Kay L, Ricketts C. 2010. Adaptation of medical progress testing to a dental setting. Med Teach 32:500–502.

Boshuizen HPA, van der Vleuten CPM, Schmidt H, Machiels-Bongaerts M. 1997. Measuring knowledge and clinical reasoning skills in a problem-based curriculum. Med Educ 31:115–121.

Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. 2003. Measurement practices: Methods for developing content-valid student examinations. Med Teach 25:414–421.

Burton RF. 2002. Misinformation, partial knowledge and guessing in true/false tests. Med Educ 36:805–811.

Butler AC. 2010. Repeated testing produces superior transfer of learning relative to repeated studying. J Experiment Psychol: Learning, Memory Cognit 36: 1118–1133.

Carpenter SK, Pashler H, Wixted JT, Vul E. 2008. The effects of tests on learning and forgetting. Memory Cognit 36:438–448.

Case SM, Swanson DB. 2002. Constructing written test questions for the basic and clinical sciences. Philadelphia, PA: National Boad of Medical Examiners.

Coderre S, Woloschuk W, McLaughlin K. 2009. Twelve tips for blueprinting. Med Teach 31:322–324.

Coombes L, Ricketts C, Freeman A, Stratford J. 2010. Beyond assessment: Feedback for individuals and institutions based on the progress test. Med Teach 32:486–490.

Danish KF, Khan RA. 2010. Role of effective feedback in multiple choice questons (MCQs) designing for faculty development. J Rawalpindi Med Coll 14:98–100.

De Champlain A, Cuddy MM, Scoles PV, Brown M, Swanson DB, Holtzman K, Butler A. 2010. Progress testing in clinical science education: Results of a pilot project between the National Board of Medical Examiners and a US medical School. Med Teach 32:503–508.

Dijksterhuis MGK, Scheele F, Schuwirth LWT, Essed GGM, Nijhuis JG. 2009. Progress testing in postgraduate medical education. Med Teach 31:e464–e468.

Downing SM. 2002. Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Adv Health Sci Educ 7:235–241.

Downing SM. 2003. Guessing on selected-response examinations. Med Educ 37:670–671.

Downing SM. 2005. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ 10:133–143.

Downing SM, Tekian A, Yudkowsky R. 2006. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med 18:50–57.

Espinoza MP, Gardeazabal J. 2010. Optimal correction for guessing in multiple-choice tests. J Math Psychol 54:415–425.

Finucane P, Flannery D, Keane D, Norman G. 2010. Cross-institutional progress testing: Feasibility and value to a new medical school. Med Educ 44:184–186.

Frank, JR (Ed.) 2005. The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better Care., Ottawa, The Royal College of Physicians and Surgeons of Canada.

Freeman A, Nicholls A, Ricketts C, Coombes L. 2010a. Can we share quesitons? Performance of questions from different question banks in a single medical school. Med Teach 32:464–466.

Freeman A, Ricketts C. 2010. Choosing and designing knowledge assessments: Experience at a new medical school. Med Teach 32:578–581.

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

General Medical Council. 2009. Good medical practice. Retrieved 24 January 2012.Available from http://www.gmc-uk.org/static/documents/content/GMP_0910.pdf

Haladyna TM, Downing SM. 1989a. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ 2:37–50.

Haladyna TM, Downing SM. 1989b. Validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ 2:51–78.

Haladyna TM, Downing SM, Rodriquez MC. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 15:309–334.

Holsgrove G, Elzubeir M. 1998. Imprecise terms in UK medical multiple-choice questions: What examiners think they mean. Med Educ 32:343–350.

International Foundations of Medicine 2011. Retrieved 20 July 2011. Available from http://www.nbme.org/Schools/iFoM/index.html

International Partnership for Progress Testing 2011. Retrieved 18 July 2011. Available from http://ipptx.org/

Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson DB, Glew RH. 2002. The quality of in-house medical school examinations. Acad Med 77:156–161.

Kerfoot BP, Shaffer K, McMahon GT, Baker H, Kirdar J, Kanter S, Corbett EC, Berkow R, Krupat E, Armstrong EG. 2011. Online "spaced education progress-testing" of students to confront two upcoming challenges to medical schools. Acad Med 86:300–306.

Koens F, Rademakers JDJ, Ten Cate TJ. 2005. Validation of core medica knowledge by postgraduates and specialists. Med Educ 39:911–917.

Larsen DP, Butler AC, Roediger HL. 2009. Repeated testing improves lngterm retention relative to repeated study: A randomised controlled trial. Medical Education :43**,** 1174–1181.

Malau-Aduli BS, Zimitat C. 2011. Peer review improves the quality of MCQ examinations. Assess Eval Higher Educ, DOI:10.1080/02602938.2011.586991.

Mardiastuti HW, Werdhani RA. 2011. Grade point average, progress test, and try outs's test as tools for curriculum evaluation and graduates' performance prediciton at the national baord examination. J Med Med Sci 2:1302–1305.

McCrorie P, Boursicot AM. 2009. Variations. in medical school graduating examinations in the United Kingdom: Are clinical competence standards comparable? Med Teach 31:223–229.

McHarg J, Bradley P, Chamberlain S, Ricketts C, Searle J, McLachlan JC. 2005. Assessment of progress tests. Med Educ 39:221–227.

McLaughlin K, Lemaire J, Coderre S. 2005. Creating a reliable and valid blueprint for the internal medicine clerkship evaluation. Med Teach 27:544–547.

Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, van der Vleuten CPM. 1998. Relative or absolute standards in assessing medical knowledge using progress tests. Adv Health Sci Educ 3:81–87.

Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, van der Vleuten CPM. 2007. Origin bias of test items compromises the validity and fairness of curriculum comparisons. Med Educ 41:1217–1223.

Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, van der Vleuten CPM. 2008. Differences in knowledge development exposed by multi-curricular progress test data. Adv Health Sci Educ 13:593–605.

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

Muijtjens AMM, van Mameren H, Hoogenboom RJI, Evers JLH, van der Vleuten CPM. 1999. The effect of a 'don't know' option on test scores: Number-right and formula scoring compared. Med Educ 33:267–275.

Muijtjens AMM, Wijnen W. #2010. Progress testing.. In: Van Berkel H, Scherpbier A, Hillen H, Van Der Vleuten C, editors. Lessons from problem-based learning. Oxford: Oxford University Press.

Munro N, Rughani A, Foukles J, Wilson A, Neighbour R. 2000. Assessing validity in written tests of general practice – Exploration by factor analysis of candidate response patterns to Paper 1 of the MRCGP examination. Med Educ 34:35–41.

Naeem N, van der Vleuten C, Alfaris EA. 2011. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ , DOI 10.1007/s10459-011-9315-2.

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.

Nouns ZM, Georg W. 2010. Progress testing in German speaking countries. Med Teach 32:467–470.

Prideaux D, Gordon J. 2002. Can global co-operation enhance quality in medical education? Some lessons from an international assessment consortium in medical education. Med Educ 36:404–405.

Rademakers J, Ten Cate TJ, Bar PR. 2005. Progress testing with short answer questions. Med Teach 27:578–582.

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

Ricketts C, Freeman AC, Coombes LR. 2009. Standard setting for progress tests: Combining external and internal standards. Med Educ 43:589–593.

Ricketts C, Moyeed R. 2011. Improving progress test score estimation using Bayesian statistics. Med Educ 45:570–577.

Roediger HL, Butler AC. 2010. The critical role of retrieval practice in long-term retention. Trends Cognit Sci 15:20–27.

Roediger HL, Karpicke JD. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. Psychol Sci 17:249–255.

Roex A, Degryse J. 2004. A computerized adaptive knowledge test as an assessment tool in general practice: A pilot study. Med Teach 26:178–183.

Sales D, Sturrock A, Boursicot K, Dacre J. 2010. Blueprinting for clinical performance deficiencies – Lessons and principles from the General Medical Council's fitness to practise procedures. Med Teach 32:e111–e114.

Schaap L, Schmidt H, Verkoeijen PJL. 2011. Assessing knowledge growth in a psychology curriculum: Which students improve most? Assess Eval Higher Educ :1–13.

Scharf EM, Baldwin LP. 2007. Assessing multiple choice question (MCQ) tests – A mathematical perspective. Active Learn Higher Educ 8:31–47.

Schauber S, Nouns ZB. 2010. Using the cumulative deviation method for cross-institutional benchmarking in the Berlin progress test. Med Teach 32:471–475.

Schuwirth L. 2007. The need for national licencing examinations. Med Educ 41:1022–1023.

Schuwirth L. 2011. Personal communication.

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

Stagnaro-Green AS, Downing SM. 2006. Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. Med Teach 28:566–568.

Swanson DB, Holtzman KZ, Butler A, Langer MM, Nelson MV, Chow JWM, Fuller R, Patterson JA, Boohan M, Committee M-SPT. 2010. Collaboration across the pond: The multi-school progress testing project. Med Teach 32:480–485.

Tarrant M, Knierim A, Hayes SK, Ware J. 2006. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Today 26:662–671.

Tarrant M, Ware J. 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 42:198–206.

Tombleson P, Fox RA, Dacre JA. 2000. Defining the content for the objective structured clinical examination component of the Professional and Linguistic Assessment Board examination: Development of a blueprint. Med Educ 34:566–572.

Tweed M, Wilkinson T. 2009. A randomized controlled trial comparing instructions regarding unsafe response options in a MCQ examination. Med Teach 31:51–54.

Van der Veken J, Valcke M, De Maeseneer J, Schuwirth L, Derese A. 2009. Impact on knowledge acquisition of the transition from a conventional to an integrated contextual medical curriculum. Med Educ 43:704–713.

van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM, Thoben AJNM, Cohen-Schotanus J, van Boven CPA. 2004. Cross institutional collaboration in assessment: A case on progress testing. Med Teach 26:719–725.

van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. 1996. Fifteen years of experience with progress testing in a problem-based learning curriculum. Med Teach 18:103–109.

van Diest R, van Dalen J, Bak M, Schruers K, van der Vleuten C, Muijtjens AMM, Scherpbier A. 2004. Growth of knowledge in psychiatry and behavioural sciences in a probelm-based learning curriculum. Med Educ 38:1295–1301.

van Herwaarden, CLA, Laan, RFJM, Leunissen, RRM 2009. Raamplan artsopleiding 2009. In: Centra, Nfvum (Ed.

Vantini I, Benini L. 2008. Models of learning, training and progress evaluation of medical students. Clin Chim Acta 393:13–16.

Verhoeven BH, Snellen-Balendong HAM, Hay IT, Boon JM, Van Der Linde MJ, Blitz-Lindeque JJ, Hoogenboom RJI, Verwijnen GM, Wijnen WHFW, Scherpbier AJJA, et al. 2005. The. versatility of progress testing assessed in an international context: A start fro benchmarking global standardization? Med Teach 27:514–520.

Verhoeven BH, Van der Steeg AFW, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, van der Vleuten CPM. 1999. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ 33:832–837.

Verhoeven BH, Verwijnen GM, Muijtjens AMM, Scherpbier AJJA, van der Vleuten CPM. 2002a. Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. Med Educ 36:860–867.

Verhoeven BH, Verwijnen GM, Scherpbier AJJA, van der Vleuten CPM. 2002b. Growth of medical knowledge. Med Educ 36:711–717.

Wade L, Harrison C, Hollands J, Mattick K, Ricketts C, Wass L. 2011. Student perceptions of the prgores test in two settings and the implications for test deployment. Adv Health Sci Educ 01 November 2011 ed., Springer.

Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. 2006. Use of a committee review process to improve the quality of course examinations. Adv Health Sci Educ 11:61–68.

Ware J, Torstein VK. 2009. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach 31:238–243.

World Federation for Medical Examinaton. 2003. Basic medical education. Copenhagen: WFME Global Standards for Quality Improvement.

Zimmerman DW, Williams RH. 2003. A new look at the influence of guessing on the reliability of multiple-choice tests. Appl Psychol Meas 27:357–371.