Medical Teacher, Vol. 27, No. 7, 2005, pp. 578–582

Progress testing with short answer questions

J. RADEMAKERS, TH. J. TEN CATE & P.R. BÄR
University Medical Center Utrecht, School of Medical Sciences, Center for Research and Development of Education, The Netherlands

SUMMARY *The concept of progress testing was developed in the 1970s. Significant features of progress tests are that the content is not linked to any specific course or unit, and that it reflects the final objectives of the curriculum as a whole. The questions are taken from a broad domain and cover a range of disciplines. Furthermore, the test is taken repeatedly over a period of time, to monitor students' progress. Known progress tests all use closed format questions. In 2002–2003 the University Medical Center Utrecht initiated a progress test with short answer questions. The test consists of 40 cases, each with a clinical and a basic science short-answer key feature question. This differs from other progress tests that use close format items, but also in the philosophy of mastery level testing and in the deliberate linking of basic science concepts to clinical case vignettes. The first four executions of the test show high internal consistencies (Cronbach's alpha 0.85 to 0.87) and satisfactory item parameters. The effort of marking answers is reasonable, the effort of writing case vignettes with short-answer items is less than writing MC-items if similar test reliabilities are to be achieved.*

Introduction

In 1999 a new six-year medical curriculum was introduced at the University Medical Center Utrecht. A key concept in this patient/problem-oriented curriculum is a constructivistic curriculum approach: knowledge should be acquired in a functional context, students should play an active role in the acquisition of knowledge and are largely responsible for their own learning process, and the teacher's role is to guide this process (Borleffs *et al.*, 2001). Assessment methods were chosen to reflect these educational principles. One of these methods, progress testing with short answer questions, will be described and discussed.

In The Netherlands, the general concept of progress testing has been developed and implemented in the Maastricht medical curriculum since 1977 (Van der Vleuten *et al.*, 1996). The concept was based on the idea that assessment should not interfere with desirable learning behaviour in a problem-based curriculum. Recent research has shown that active self-directed and open-discovery learning is indeed not hampered by progress testing, therewith confirming the intentions of the test (Verhoeven, 2003).

From the 1970s onwards, other progress tests have been developed, some of which led to publications in the medical education literature. At the University of Missouri Medical School in Kansas City (USA) the Quarterly Profile Examination-QPE was developed (Willoughby *et al.*, 1977; Arnold & Willoughby, 1990). Later, a similar test concept (Personal Profile Index-PPI) was introduced at McMaster University in Hamilton (Canada) (Blake *et al.*, 1995, 1996). Significant features of all progress tests are that: (1) the content is not linked to any specific course or unit; (2) the test reflects the final objectives of the curriculum as a whole (the questions are taken from a broad domain and cover a range of disciplines) and (3) the test is taken repeatedly over a period of time to monitor students' progress. These progress test examples all use closed format questions (yes/no/? or multiple choice). The Maastricht progress test consists of around 250 closed format questions (yes/no/?), whereas the QPE and the PPI have 400 and 180 multiple choice questions respectively. The use of closed format/multiple choice questions has several well known advantages (Norman *et al.*, 1991; Van der Vleuten *et al.*, 1991). Scoring can be computerized and is efficient and objective, therefore the method is appropriate for large groups of students and many questions can be asked in a relatively short period of time, while covering a broad knowledge domain. However, there are also disadvantages. The construction of high stakes multiple choice-questions is time-consuming, and many questions are needed to produce a reliable test. Other concerns regard the validity of the test: cueing effects present serious limitations, as well as the fact that recognition is tested, when sometimes recall would be desirable. Furthermore, it is relatively easy to construct closed format multiple choice-questions to test factual knowledge and specifically difficult to construct true–false items to assess clinical reasoning skills (Downing, 2002). Some medical schools have attempted to overcome these disadvantages by constructing a progress test with open-ended questions. For example, Dundee Medical School implemented a progress test with constructed response questions (CRQ's) in the undergraduate curriculum (Friedman-Ben David *et al.*, 2001). Also the University Medical Center Utrecht (The Netherlands) chose to develop a progress test with open-ended, short answer questions.

The Utrecht Progress Test (UPT)

In 2002–2003, the Utrecht progress test was implemented in year 4 and 5 of the curriculum. The introduction of the test served several goals:

- Testing the minimum medical core knowledge that every medical student should have mastered by the time of graduation at a level of what is to be known by heart without preparation, with an emphasis on clinical reasoning. This goal reflects the summative use of the progress test and a quality assurance before students enter their final year clerkships;

Correspondence: Dr J.J.D.J.M. Rademakers, University Medical Center Utrecht, School of Medical Sciences, Center for Research and Development of Education, HB 4.01 PO Box 85500, 3508 GA Utrecht, The Netherlands. Email: j.j.d.j.m.rademakers@umcutrecht.nl

ISSN 0142–159X print/ISSN 1466–187X online/05/070578-5 © 2005 Taylor & Francis
DOI: 10.1080/01421590500062749

- Providing feedback with respect to the progress of the student and to possible gaps in their knowledge. This goal reflects the formative use of the test and can be applied to students individually and to cohorts of students for purposes of curriculum evaluation;
- Integrating basic science concepts with clinical cases. Vertical integration of basic science with clinical teaching has been widely advocated, but integration within test items has not been described yet;
- Consolidation of the medical core knowledge of the student, as a result of the frequent impetus to produce this knowledge;
- Introduction of an incentive for a longitudinal self-directed (vs. a short term test-directed) learning style.

The choice for open-ended, short answer questions was based on several studies. Ten Cate *et al.* (1996) and Ten Cate (1997) demonstrated that equal or higher test reliabilities can be achieved with fewer short answer questions compared to true/false questions, concluding that—even if the effort of marking the answers is taken into account—it can be more efficient to construct short answer tests if comparable reliabilities are to be attained. Fenderson *et al.* (1997) concluded that tests with open-ended questions are better able to discriminate well prepared students from marginal students, as they allow more possibility for differentiation in scores. The grading is done by the person who is primarily responsible for the assessment of students, i.e. the teacher, which connects teachers directly to the effect of their education on individual students. The UPT is based on the philosophy of mastery level testing (Gagné *et al.*, 1992). Since the test focuses on a basic knowledge level, the final standard for passing the test is not set at a relative cut-off score, as was done in Maastricht Progress Test for the first period of 20 years (Verhoeven *et al.*, 2002), nor is it set at 55% correct answers (which is a common absolute standard in The Netherlands), but at 80%. Ideally, one would want the students to fully 'master' this knowledge and hence know the answers to all the questions, but since question design and therefore the reliability of the test will only in theory be perfect, an 80% score is considered an adequate cut-off point.

Consistent with the educational principles of patient-centered and problem-oriented learning, basic science concepts are presented within a clinical context. Problem based learning produces deep levels of learning and understanding and an ability to apply basic science knowledge in the clinical setting (Newble & Entwhistle, 1986; Norman & Schmidt, 1996). Since assessment influences the approach to learning (Newble & Entwhistle, 1986), the UPT aims at reinforcing this learning style by linking the basic science questions to the clinical cases. The cases provide a real-life context for clinical reasoning and should lead to external face validity of the test.

Method

Test design and construction

Each version of the UPT consists of 40 clinical cases, each with one to three clinical items and one to three basic science items, all in a short-answer key feature format. For each version of the UPT, new cases and questions are written. The cases cover a broad range of patients and problems, and the National Dutch Blueprint (2001) and the USMLE (step 1) are being used as a blueprint framework. Furthermore, the test blueprint sets out the required number of questions for each of the ten participating clinical disciplines. The distribution of the basic science disciplines can vary, depending on the content of the case.

A rigorous procedure of question design assures the quality of each version of the test. Before a test, representatives of each clinical discipline are asked to write cases and clinical questions. In a standard form the clinical case (maximum 150 words), the question(s), the model answer and scoring indications are to be specified. Then each case and question is sent to a basic science representative, who adds basic science questions to the cases (see Figure 1).

A second step in the question design procedure is the assessment of the quality, level and content of the questions by a small panel of experts in the area of test construction and question writing. Comments are discussed in person with the authors of the questions, and a first general appraisal of the total level and content of the test is made.

After the revision of the questions the whole test is reviewed by a multidisciplinary 'Progress Test Committee', with a final authority with respect to the selection or rejection of questions, and the construction of the test as a whole. Main criteria of the 'Progress Test Committee' are the perceived difficulty level of the question ('Is it core knowledge?'), clarity and phrasing of the question and the model answer, and whether the content overlaps with other questions in the test.

After the test, the questions are separated. Item writers mark the answers on their own items, following their model answers. All scores are combined in a database, and the reliability of the test as a whole and the item-parameters are computed. On the basis of these statistics and the expert-opinion of committee members on the quality and content of the question, the Program Test Committee decides on rejection of questions if necessary (see Figure 2).

Regulations

The test is taken by students in their fourth and fifth year. Three times a year the test is executed. Students can sit the test as many times as they want, until the required cut off score of 80% is reached. As a second rule, students must reach a cumulative total score of 80% in at least two other separate exams together. In this way participation in at least three progress tests is stimulated without obligatory presence.

Students receive their scores on each separate question and on the test as a whole. The questions and model answers are made available through the Internet.

Results

Reliability and item quality

Analyses of the first four test executions show a high internal consistency (Cronbach's alpha 0.87, 0.87, 0.85, 0.86 respectively). Item parameters are usually satisfactory; five

Item code: V02-057	**Deadline for delivery:**
Clinical department: Obsterics/Gynecology	**Clinical item writer:** XX
Basic science department: Physiology	**Basic Science item writer:** YY
Dutch Blueprint code:	**Clinical problem:** edema
Subject of disease: Pre-ecclampsia	**The clinical item concerns:** **0 history** **x physical examination** **x diagnostic investigation** **0 therapeutic policy** **0 other**
Casus description (max. 150 words) You are a locum on Schiermonnikoog. A 32-year-old woman who is on a 2-week holiday comes to your practice with complaints of headache and blurred vision. She is 32 weeks pregnant (it is her first pregnancy). Her history shows that she also suffers from swollen legs, fluid retention in the fingers, and a marked gain of weight over the last weeks.	
Clinical item (max. 3 points) 1. What is the probable diagnosis? 2. What physical examination should you perform (mention three parts)? 3. What supplementary laboratory investigations can you perform as general practitioner to establish the diagnosis?	
Model answer 1. Pre-eclampsia. 2. Take blood pressure, measure height of fundus uteri, reflexes. 3. Test urine for protein.	
Scoring suggestions All answers 1 point. For 2: 1 point when at least two parts are mentioned, including taking blood pressure, otherwise 0 points.	
Basic science item (max. 3 points) 1. Name three factors that determine the degree of tissue perfusion. 2. Indicate, for each factor, whether it will be increased, decreased, or not changed.	
Model answer 1. (a) arterial blood pressure, (b) venous blood pressure, (c) colloid-osmotic pressure. 2. (a) increased, (b) not changed, (c) decreased.	
Scoring suggestions 1 point for every correct combination of factor + change.	

Figure 1. Form for preparing progress test items, with example.

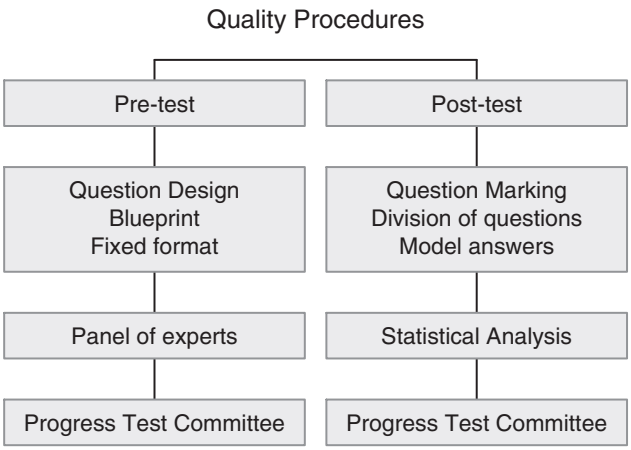

Figure 2. Quality procedures.

of 320 items were rejected after the test because of unsatisfactory item quality.

Time and costs

After the first two tests the authors of the questions were asked how much time the total process of writing a question, discussion with experts, revision and marking of the answers had taken them. In total, the average time was 4.7 hours per question: 2 hours for question design, 10 minutes discussion, 1 hour revision and 1.5 hour for marking the answers.

Test results

The results of the first two test occasions showed mean scores of 44.7% and 47.6% of the maximum score, and a gradual increase in scores on following test occasions.

Table 1. Test scores.

Year	2002–2003		2003–2004	
Test occasion	1	2	3	4
Highest score in %	66.7	67.1	70.8	72.2
$N=$	164	165	352	382

However, this increase was much less than we had expected, and not one student managed to come near the 80% criterion on a single test occasion (Table 1). This was a great concern for both students and staff, since it meant that a whole year group wasn't going to pass the test and couldn't move on to the next year.

To determine the cause of these low scores, a separate study on the level of the questions by two separate panels (residents and experienced physicians) was done. The results from this study showed that more than one third of the test questions were too difficult and asked for details that cannot be regarded as core knowledge (Koens *et al.*, in press). As a result of this study, the instructions to the authors and the reviewing of the questions have become even more strict with regard to the difficulty level. Furthermore the decision was made to set a new pass/fail standard using a combination method (Cohen-Schotanus *et al.*), so that the difficulty level of the test was taken into account. The absolute maximum score (100%) was replaced by a relative maximum (average score of the top 10% of fifth year students). Students now pass the test when they reach 80% of this (relative) maximum score. The main advantage of this method is that it corrects for alternating difficulty levels per test occasion.

Discussion and conclusion

Our conclusion is that it is feasible to develop and implement a reliable progress test with short answer questions. The time and costs of question design and marking the answers are reasonable, given a limited number of questions and students. Computerizing the test process in the future might simplify the logistical process even further.

Open-ended questions hold the image that they are not reliable, or at least less reliable than closed format/multiple choice-questions (Schuwirth & Van der Vleuten, 2003). This may be caused by the fact that tests with open-ended questions are often not statistically analysed. Of course the marking of the answers has a subjective element and can be potentially biased. The high reliability on four subsequent UPT occasions proves that it is possible to have satisfactory reliability with open-ended questions. The UPT consists of short answer questions (instead of essays) which leave little room for variance in the answer. Error as a result of guessing is eliminated. Model answers and clear instructions for answer marking probably reduce errors of subjectivity.

Central in the concept of the UPT is the philosophy of mastery level testing. To be able to determine whether students master this core knowledge, it is essential that the questions have an adequate level. A thorough assessment of the questions is therefore essential in the quality procedure of the test. In the first four tests, students had lower scores than we would have been expected. A separate study showed that more than one third of the test questions were too difficult and could not be regarded as core knowledge (Koens *et al.*, in press). As a result of this, the instructions to the authors and the reviewing of the questions have become even more strict, and the pass/fail standard was adjusted to compensate for alternating difficulty levels per test.

A reason frequently given for the use of multiple choice questions in testing is that the effort of marking answers to open ended questions is time-consuming and therefore costly. We found the effort of marking answers to be very reasonable, although some departments consider marking answers not a teacher's core task. However, we estimate the effort of writing case vignettes with short-answer items less than writing multiple choice-items if similar test reliabilities are to be achieved. Starting from UPT3, two year groups took the test. With this increase of participants up to 400, the time for marking the answers has increased as well, but is still within reasonable limits. Further expansion of the number of participants (e.g. adding other cohorts) would generate organizational problems.

Practice points

- Progress tests with structured short answer questions can have high internal consistencies.
- Time and costs of question design and marking the answers are reasonable.
- Thorough assessment of the questions is essential in the quality procedure.

Notes on contributors

Rademakers J.J.D.J.M., is associate professor at UMC Utrecht, School of Medical Sciences, Center for Research and Development of Education.

Ten Cate Th.J, is professor of medical education and associate dean of education at UMC Utrecht, School of Medical Sciences, Center for Research and Development of Education.

Bär PR, is a full professor of Biomedical Sciences and chairman of the Utrecht Progress Test Committee at UMC Utrecht, School of Medical Sciences.

References

Arnold, L. & Willoughby, T.L. (1990) The quarterly profile examination, *Academic Medicine*, 65, pp. 515–516.

Blake, J.M., Norman, G.R., Kinsey, E. & Smith, M. (1995) Report card from McMaster: student evaluation at a problem-based medical school, *Lancet*, 345, pp. 899–902.

Blake, J.M., Norman, G.R., Keane, D.R., *et al.* (1996) Introducing progress testing in McMaster university's problem-based medical curriculum: psychometric properties and effect on learning, *Academic Medicine*, 71, pp. 1002–1007.

Borleffs, J.C.C., Daal, M.M. & Mulder, H. (Eds.) (2001) CRU'99. Supplement Tijdschrift Medisch Onderwijs, *Dutch Journal of Medical Education*, 20, pp. S1–S65.

Cohen-Schotanus, J., van der Vleuten, C.P.M., Bender, W. (1996) A better pass/fail standard at tests; taking the best students as point of reference (in Dutch). In: Th. J. Ten Cate, J.H. Dikker, E.H. Houtkoop, *et al.*, *Proceedings of the Dutch Association for Medical Education*. (Houten, Bohn Stafleu Van Loghum, pp. 83–87.

Fenderson, B.A., Damjanov, I., Robeson, M.R., *et al.* (1997) The virtues of extended matching and uncued tests as alternatives to multiple choice questions, *Human Pathology*, 28, pp. 526–532.

Friedman-Ben David, M., Hunter, I. & Harden, R. (2001) Introduction of a progress test as a tool for defining core undergraduate curriculum, *Educacion Medica*, 4, pp. 97–99.

Gagné, R.M., Briggs, L.J. & Wager, W.W. (1992) *Principles of Instructional Design* (Fort Worth, Harcourt Brace Jovanovich).

Koens, F., Rademakers, J.J.D.J.M. & Ten Cate, Th.J. Can postgraduate medical trainees and medical specialists evaluate the required level of core biomedical knowledge for graduation? *Medical Education*, in press.

Newble, D.I. & Entwhistle, N.J. (1986) Learning styles and approaches: implications for medical education, *Medical Education*, 20, pp. 162–175.

Norman, G.R., van der Vleuten, C.P.M. & de Graaff, E. (1991) Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability, *Medical Education*, 25, pp. 119–126.

Norman, G.R. & Schmidt, H.G. (1996) The psychological basis of problem-based learning. A review of the evidence, *Academic Medicine*, 67, pp. 557–565.

Schuwirth, L.W.T. & Van der Vleuten, C.P.M. (2003) ABC of learning and teaching in medicine: written assessment, *BMJ*, 326, pp.643–645.

Ten Cate, Th.J., Gouma, D.J. & Taminiau, A. (1996) Open and closed format test items in one test: exploration of causes of reliability differences I: experience with four test executions [in Dutch], in: Th.J. Ten Cate, J.H. Dikker, E.H. Houtkoop, *et al.*, *Proceedings of the the Dutch Association for Medical Education*, Houten, Bohn Stafleu Van Loghum, pp. 67–74.

Ten Cate, Th.J. (1997) Comparing reliabilities of true/false and short-answer questions in written problem solving tests. in: A. Scherpbier, *et al.* (Ed). *Advances in Medical Education*, pp.193–196. (Dordrecht, Kluwer Academic Publishers).

Van der Vleuten, C.P.M., Verwijnen, G.M. & Wijnen, W.H.F.W. (1996) Fifteen years of experience with progress testing in a problem-based learning curriculum, *Medical Teacher*, 18, pp.103–109.

Van der Vleuten, C.P.M., Norman, G.R. & de Graaff, E. (1991) Pitfalls in the pursuit of objectivity: issues of reliability, *Medical Education*, 25, pp. 110–118.

Verhoeven, B.H., Verwijnen, G.M., Muijtjens, A.M.M., *et al.* (2002) Panel expertise for an Angoff standard setting procedure in progress testing; item writers compared to recently graduated students, *Medical Education*, 36, pp. 860–867.

Verhoeven, B.H. (2003) *Progress Testing; the Utility of an Assessment Concept.* Dissertation Maastricht University.

Willoughby, T.L., Dimond, E.G. & Smull, N.W. (1977). Correlation of quarterly profile examination and national board of medical examiner scores, *Educational and Psychological Measurement*, 37, pp. 445–449.