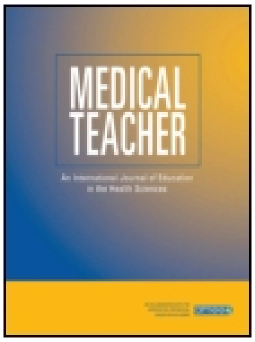

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: informahealthcare.com/journals/imte20

Progress testing in postgraduate medical education

M.G.K. Dijksterhuis, F. Scheele, L.W.T. Schuwirth, G.G.M. Essed, J.G. Nijhuis & D.D.M. Braat

To cite this article: M.G.K. Dijksterhuis, F. Scheele, L.W.T. Schuwirth, G.G.M. Essed, J.G. Nijhuis & D.D.M. Braat (2009) Progress testing in postgraduate medical education, Medical Teacher, 31:10, e464-e468, DOI: 10.3109/01421590902849545

To link to this article: https://doi.org/10.3109/01421590902849545

2009; 31: e464–e468 MEDICAL TEACHER

WEB PAPER

Progress testing in postgraduate medical education

M.G.K. DIJKSTERHUIS[1], F. SCHEELE[2], L.W.T. SCHUWIRTH[3], G.G.M. ESSED[3], J.G. NIJHUIS[4] & D.D.M. BRAAT[5]

[1]Ikaziaziekenhuis, Obstetrics and Gynaecology, Netherlands, [2]University of Maastricht, Netherlands, [3]University Hospital Maastricht, Netherlands, [4]Radboud University Nijmegen Medical Centre, Netherlands, [5]Sint Lucas Andreas Ziekenhuis, Obstetrics and Gynaecology, Netherlands

Abstract

Background: The role of knowledge in postgraduate medical education has often been discussed. However, recent insights from cognitive psychology and the study of deliberate practice recognize that expert problem solving requires a well-organized knowledge database. This implies that postgraduate assessment should include knowledge testing. Longitudinal assessment, like progress testing, seems a promising approach for postgraduate progress knowledge assessment.

Aims: To evaluate the validity and reliability of a national progress test in postgraduate Obstetrics and Gynaecology training.

Methods: Data of 10 years of postgraduate progress testing were analyzed on reliability with Cronbach's alpha and on construct validity using one-way ANOVA with a *post hoc* Scheffe test.

Results: Average reliability with true–false questions was 0.50, which is moderate at best. After the introduction of multiple-choice questions average reliability improved to 0.65. Construct validity or discriminative power could only be demonstrated with some certainty between training year 1 and training year 2 and higher training years.

Conclusion: Validity and reliability of the current progress test in postgraduate Obstetrics and Gynaecology training is unsatisfactory. Suggestions for improvement of both test construct and test content are provided.

Introduction

Knowledge and medical expertise

The role of knowledge in the development of medical expertise has been extensively debated during the last century, dating back to the beginning of the twentieth century when Flexner stressed the importance of sciences as the fundamental basis of medicine, while Osler took the opposite position pleading for a more practice-orientated teaching method (Osler 1903; Flexner 1910). The latter view has been embraced in the emergence of problem-based learning curricula in the early 1970s. Initially this educational philosophy tended to marginalize the role of knowledge, stating that students should first learn to solve problems and would gather the appropriate knowledge on the side. In these views emphasis shifted from memorizing to knowing how to use resources adequately (Patel et al. 1989).

However, later developments in cognitive psychology have persistently shown that a mere immersion in practice is not enough to develop expertise and that expert problem solving cannot take place without a well-organized knowledge database and thus requires expert knowledge. With that, the central role of knowledge in (medical) expertise has been re-emphasized (Schmidt et al. 1990; Norman 1991; Van de Wiel 1997; Ericsson 2004).

Practice points

- Knowledge testing should be part of postgraduate assessment in medical education.
- Theoretically, progress testing would be an interesting approach to knowledge assessment in postgraduate assessment in medical education.
- Reliability and construct validity of the current progress test in postgraduate Obstetrics and Gynaecology training are somewhat disappointing.

Choosing the right test format

Acknowledging the central role of knowledge in expertise, it appears at least remarkable that the assessment of knowledge is underexposed in postgraduate medical training, particularly outside the Anglo-Saxon countries. Furthermore, little is known about which test-format performs best in a postgraduate setting. Even in countries where postgraduate certifying exams are mandatory, the absence of data on validity and reliability of test methods in use is striking (Hutchinson et al. 2002). More recently introduced assessment methods like the 360-degree feedback (Joshi et al. 2004); the mini-CEX (Norcini et al. 1995) and portfolio (Mathers et al. 1999) seem to focus predominantly on practice performance. Knowledge

Correspondence: M.G.K. Dijksterhuis, Ikaziaziekenhuis, Obstetrics and Gynaecology, Montessoriweg 1, Rotterdam, 3083 AN, Netherlands. Tel: 0031-10-654267968; fax: 0031-10-2975017; email: m.dijksterhuis@wanadoo.nl

ISSN 0142–159X print/ISSN 1466–187X online/09/100464–5 © 2009 Informa Healthcare Ltd.
DOI: 10.3109/01421590902849545

assessment makes up only a limited part of these assessment methods and represents a small sample of the complete knowledge level of a trainee.

Under the assumption that knowledge testing should be part of postgraduate assessment, it is interesting to explore the most suitable test format for the particular circumstances that surround postgraduate training, as it differs from undergraduate training in various aspects. Firstly, it concerns usually only small groups of trainees per speciality per training centre. Secondly, it is characterized by highly individual learning pathways, which largely escape the control of the trainer and trainee and are dominated by the patient problems at hand. Moreover, rotations differ amongst training centres, despite end goals being quite similar for all trainees in the same specialty.

Traditionally, national modular tests and/or final examinations have been used to measure the knowledge level of trainees. However, there are indications that modular testing may lead to short-term exam-driven learning and could interfere with other training activities (Newble and Jaeger 1983; Farr 1987). Final examinations, on the other hand, may merely result in a one-point measurement, representing a ‘snapshot’ of a trainee’s knowledge level, not allowing any extrapolation to the maintained knowledge level over time (Van der Vleuten 2000).

The case for progress testing in postgraduate medical education

Longitudinal knowledge testing, aiming at measuring growth of knowledge over time, seems a more promising approach to overcome some of the specific problems of postgraduate knowledge assessment. In longitudinal testing, candidates repeatedly sit tests each of which tests the knowledge level that is expected at the end of training, regardless of the actual training year of the candidate. As each test is a sample out of the complete knowledge domain the test is supposed to be too comprehensive to study for. Instead, all study behaviour that results in the acquisition of relevant knowledge will be rewarded. This is believed to encourage more profound and deep learning. As each test tests the graduate knowledge level, longitudinal testing is training programme independent as long as the end terms are similar. This implies that longitudinal testing can be organized at a national level, overcoming problems associated with making tests for small groups. Additionally, this kind of flexible testing allows a great degree of freedom for the individual learning pathway. Last but not least, longitudinal testing is often praised for its formative possibilities, identifying strength and weaknesses in the knowledge level of a trainee, and providing a focus to guide further learning. The best-known example of longitudinal knowledge testing probably is progress testing (Arnold and Willoughby 1990; Blake et al. 1996; Van der Vleuten et al. 1996).

Aim of this study

Theoretically, it seems useful to apply progress testing in postgraduate medical training to assess the knowledge aspect of the developing expertise of a trainee. However, whether it is really viable remains to be evaluated. The purpose of this study is to evaluate two important factors that determine the utility of progress testing in postgraduate medicine, namely, reliability and validity (Van der Vleuten 1996).

Methods

Progress test

The postgraduate Obstetrics and Gynaecology progress test is a compulsory test taken at a yearly interval by all trainees in Obstetrics and Gynaecology in the Netherlands.

Instrument

The test consists of 150 questions divided over the sub domains – obstetric perinatology, gynaecology, reproductive medicine, oncology and health and society issues, according to a pre-set blue-print. Until 2004, true–false (TF) items were used. From 2005 onwards, single-best-option multiple-choice questions (MCQ) are being used. This change was made because of growing dissatisfaction with the previous format. Watertight TF-items are more difficult to construct than MCQs. More importantly repeatedly low correlations between true-keyed and false-keyed items were found. As these two subtests can be seen as one of the possible split half tests, correlations roughly equal to the test reliability would be expected. Apparently, the answer key constitutes an unwanted source of error and this can be seen as an inherent draw back of the TF-format.

Test questions reflect the whole domain of Obstetrics and Gynaecology based on the end terms as defined by the Dutch Society of Obstetrics and Gynaecology (NVOG). Questions are mainly drawn from the work floor, with an emphasis on knowledge that is applicable in daily gynaecologic practice. The correct answers are lined with literature references.

An extensive quality control, consisting of three pre-test review cycles and a post-test item analysis by a review committee, surrounds each question. This committee consists of five gynaecologists with various backgrounds, two trainees and one expert on medical education. Questions are reviewed on phrasing, content and relevance.

Results are calculated as a correct-minus-incorrect score, and a relative norm is used. Scores lower than the training year mean minus two standard deviations are considered unsatisfactory. The results, subdivided over the above-mentioned sub-domains, are solely disclosed to the trainee and his/her tutor. Until now, test results had no summative consequences, but were supposed to be used in a formative way during in-training assessments.

Data collection

Anonymized data of 10 years of progress testing are recorded and archived at the Department of Educational Development and Research of the University of Maastricht, the Netherlands.

Table 1. Number of exam-candidates (*n*), mean correct-minus-incorrect scores (mean) and standard deviations (sd) per training year per year of testing.

	1999			2000			2001			2002			2003		
Training year	*n*	mean	sd	*n*	mean	sd	*n*	mean	sd	*n*	mean	sd	*n*	mean	sd
1	42	31	8.8	39	22	9.0	42	22	8.6	58	17	9.1	49	23	9.4
2	31	36	7.7	48	28	9.4	39	26	7.9	46	18	7.9	58	30	8.3
3	28	40	8.2	28	29	8.2	40	29	8.9	32	23	9.7	45	33	8.9
4	29	44	9.0	26	31	9.0	27	31	8.7	42	26	8.3	32	36	7.7
5	24	46	7.7	26	36	11.4	29	31	7.7	33	27	9.6	38	36	9.2
6	34	47	8.9	27	36	7.4	29	34	9.2	23	27	8.7	15	39	3.9
	2004			2005			2006			2007			2008		
	n	mean	Sd	*n*	mean	sd	*n*	mean	sd	*n*	mean	sd	*n*	mean	sd
1	42	21	7.5	47	32	7.6	37	34	10.2	49	16	9.3	49	17	9.9
2	50	29	10.0	48	37	8.0	47	41	8.6	37	17	8.6	50	26	9.9
3	60	32	9.1	47	45	9.4	48	44	9.1	45	19	9.7	40	30	11.5
4	45	34	9.8	53	47	8.3	47	47	8.8	43	22	10.3	48	33	9.1
5	30	39	8.8	41	46	7.1	52	51	8.4	41	25	9.5	41	34	8.1
6	33	40	7.9	21	47	7.2	22	54	7.1	46	26	8.3	40	36	8.9

Statistical analysis

Reliability. Per test and per training year reliability was calculated using Cronbach's alpha.

Construct validity. Mean scores per training year per year of testing were calculated. As an indication for construct validity the growth of knowledge per training year per year was used. This was tested for significance with one-way ANOVA and Scheffe's *post hoc* tests.

Results

Descriptives

From 1999 to 2008, 10 successive progress tests have been organized and a total of 2358 test results were available for analysis.

Mean correct-minus-incorrect scores per training year per year of testing including standard deviations are shown in Table 1.

The same mean correct-minus-incorrect scores per year of testing are shown in Figure 1. We used linear graphs as this summarizes the data in a comprehensible way as differences between training years can be easily visualized. Besides, this figure demonstrates clearly that the level of difficulty of the test fluctuates per year of testing. However, shown data are transversal and not longitudinal (meaning that per year of progress testing mean scores per training year were calculated, each colour indicating a different year of progress testing).

Reliability

Table 2 shows the reliabilities of the test expressed in Cronbach's alpha per training year and per year of testing.

The alphas represent an estimation of the internal consistency reliability, which explains the variance per test per training year. The lowest value was found for training year 6 in 2003 ($\alpha = 0.2$), highest value was reached in training year 1 in 2008 ($\alpha = 0.81$). Average reliability approximated 0.50 with TF items. This improved to a reliability of 0.65 after the introduction of single-best-option MCQ, which is moderate. This led to the decision to set the pass–fail score to the training year mean minus two standard deviations. With these reliabilities, this is more than two standard errors of measurement below the mean. Therefore, an

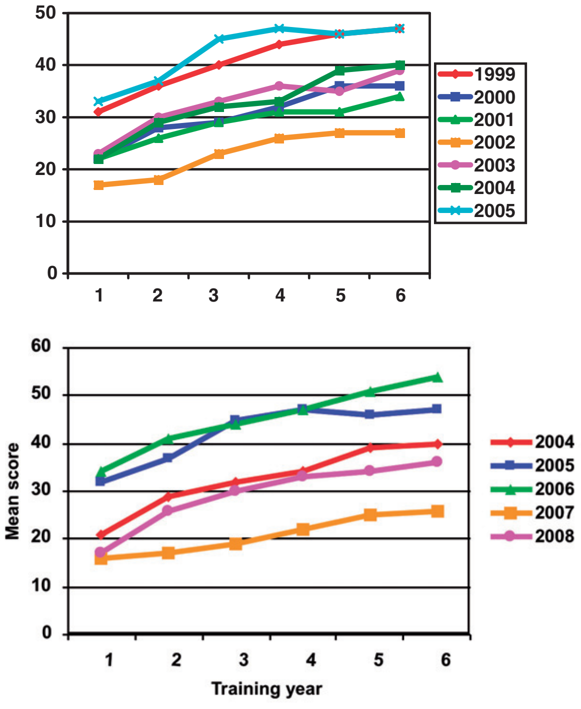

Figure 1. Mean correct-minus-incorrect scores per training year per year of testing (transversal).

Table 2. Cronbach's alpha for each training year per year of testing.

Training year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
1	0.50	0.52	0.44	0.56	0.59	0.36	0.66	0.79	0.73	0.81
2	0.33	0.53	0.34	0.30	0.47	0.63	0.69	0.74	0.48	0.66
3	0.45	0.39	0.48	0.56	0.52	0.58	0.77	0.75	0.65	0.64
4	0.55	0.43	0.42	0.39	0.40	0.61	0.72	0.74	0.65	0.52
5	0.41	0.70	0.23	0.54	0.53	0.54	0.60	0.74	0.62	0.56
6	0.51	0.34	0.54	0.41	0.20	0.42	0.62	0.66	0.56	0.61

'unsatisfactory' decision can be made with an accuracy of $p < 0.05$.

Construct validity

Significant differences where calculated per training year per year of testing by using a one-way ANOVA. We found significant differences between the training years for every year of testing. However, *post hoc* analysis using Scheffe's test has shown this to be mainly due to differences in mean scores between training year 1, training year 2 and higher training years.

Discussion

Despite the fact that an intensive quality control process supports our progress test, only moderate reliabilities are reached per test administration. Even though for formative assessment, reliabilities of 0.70–0.79 would be acceptable (Downing 2004) and this standard is still not met in our study. The fact that the change to MCQ instead of TF questions has raised reliabilities is, however, encouraging. Furthermore, construct validity or discriminative power can only be demonstrated with some certainty for training year 1 and training year 2. This implies that the current postgraduate progress test is not suitable for summative purposes in higher training years.

However, why do we find a moderate reliability and validity despite all our efforts? A possible explanation for the moderate reliabilities is that the progress test contains too many irrelevant (zero variance) items, and this is a test construction problem (Downing and Haladyna 2004). However, this seems highly unlikely, as item analysis after each test did not show many zero variance items.

Another, more convincing possibility, is inadequate sample size: perhaps the sub domains questioned are too diverse, containing too few questions to adequately sample the complete domain, resulting in loss of measurable signal (Van der Vleuten 1996; Downing and Haladyna 2004).

With only moderate reliabilities it is not surprising that validity is moderate as well. However, apart from flaws in test construction, this may also represent a problem with test content. Failure to demonstrate knowledge growth above a certain experience level has been reported previously and is also known as the intermediate effect (Schmidt et al. 1990). So far, it has not been possible to provide a satisfactory explanation for this phenomenon, but it appears likely that with growing expertise a form of specialized knowledge is acquired that is not assessed by present testing methods (Frederiksen 1984). Current written tools of assessment are mostly measuring the capacity to solve well-defined problems by the application of rules and principles, while the essence of expertise in the professions lies in the capacity to solve ill-defined problems, that is, reasoning in contexts of uncertainty (Charlin and Van der Vleuten 2004). Using MCQ to test knowledge entails that test items have to concentrate on proven facts and areas of controversy need to be avoided. For us it resulted in having to discard the majority of test questions that could have been extracted from daily clinical practice, as the answers were potentially ambiguous. So, even though a huge effort was made to warrant authenticity and to design test questions that reflect functional knowledge needed in daily practice, this may have been less successful than anticipated (Crohnbach 1983; Ebel 1983; Van der Vleuten 1996; Downing and Haladyna 2004; Van der Vleuten and Schuwirth 2005).

So, what could be done to improve both reliability and validity?

Firstly, the inclusion of zero items should be avoided. Creating an item bank of effective questions is highly recommended (Van der Vleuten 1996; Downing 2004). Secondly, the sample size should be increased. This can be achieved by increasing testing time, either by increasing the number of questions or, alternatively, by increasing the frequency of postgraduate progress testing (Van der Vleuten 1996; Downing and Haladyna 2004). Lastly, to overcome both the intermediate effect and the problems with test content, every effort should be made to construct test questions that reflect medical expertise. Item format appears to be relatively important in this context and extended matching questions as well as the script concordance test have been suggested to be superior in measuring medical expertise (Beullens et al. 2005; Charlin and Van der Vleuten 2004). Another alternative is to incorporate more short case-based questions as they have been shown to be better in measuring problem-solving ability than factual knowledge questions (Schuwirth et al. 2001).

The strength of this study is that it is one of the first evaluations of construct validity and reliability of progress testing in postgraduate medical education. Furthermore group sizes and sampling period appear reasonable. However, the study is somewhat limited by the fact that it only involves Dutch postgraduate trainees in Obstetrics and Gynaecology and as a consequence results may not be generally acceptable to other specializations or other countries.

Conclusion

Validity and reliability of the current progress test in postgraduate Obstetrics and Gynaecology training is unsatisfactory. Suggestions for improvement are provided.

Contributions

Jan Nijhuis, Gerard Essed and Lambert Schuwirth initiated postgraduate progress testing and collected and managed test results. Fedde Scheele, Lambert Schuwirth, Gerard Essed and Didi Braat contributed to the study design. Marja Dijksterhuis performed the statistical analysis under supervision of Lambert Schuwirth and Fedde Scheele. All authors helped to prepare the final report.

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

Notes on contributors

MISS MGK DIJKSTERHUIS (Marja). Consultant gynaecologist. Currently undertaking PhD research on assessment during postgraduate medical training. Member of the progress test examination committee.

PROF Dr F SCHEELE (Fedde). Consultant gynaecologist and teaching professor, focus on clinical medical education. Chairman of the progress test examination committee.

PROF Dr LWT SCHUWIRTH (Lambert). Research and development in the field of innovative assessment forms. Expertise: assessment of medical competence and performance; testdevelopment; higher-order cognitive skills; quality assurance of assessment; educational research in matters of assessment. Member of the progress test examination committee.

PROF Dr GGM ESSED (Gerard): Consultant Gynaecologist. Focus on methodology of clinical teaching. Initiated progress testing during Postgraduate Obstetrics and Gynaecology training in 1999. Training programme director of postgraduate Obstetrics and Gynaecology training, University of Maastricht the Netherlands.

PROF Dr JG NIJHUIS (Jan): Consultant Gynaecologist. Expertise: perinatology; prenatal diagnosis; fetal behaviour. Initiated progress testing during Postgraduate Obstetrics and Gynaecology training in 1999.

PROF DR DDM BRAAT (Didi): Consultant gynaecologist. Expertise: IVF; medical ethics; multiple birth pregnancies; reproduction; fertility. Training programme director of postgraduate Obstetrics and Gynaecology training, University of Nijmegen the Netherlands.

References

Arnold L, Willoughby TL. 1990. The quarterly profile examination. Acad Med 65:515–516.

Beullens J, Struyf E, Van Damme B. 2005. Do extended matching multiple-choice questions measure clinical reasoning? Med Educ 39:410–417.

Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. 1996. Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. Acad Med 71:1002–1007.

Charlin B, Van der Vleuten C. 2004. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. Eval Health Prof 27:304–319.

Crohnbach LJ. 1983. What price simplicity. Educational Measurement: Issues and Practice 2:11–12.

Downing SM. 2004. Reliability: On the reproducibility of assessment data. Med Educ 38:1006–1012.

Downing SM, Haladyna TM. 2004. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ 38:327–333.

Ebel RL. 1983. The practical validation of tests of ability. Educational Measurement: Issues and Practice 2:7–10.

Ericsson KA. 2004. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 79:S70–S81.

Farr MJ. 1987. The long-term retention of knowledge and skills. A cognitive and instructional perspective. New York: Springer-Verlag.

Flexner A. 1910. Medical education in the United States and Canada. From the Carnegie Foundation for the Advancement of Teaching.

Frederiksen N. 1984. The real test bias, influences of testing on teaching and learning. Am Psychol 39:193–202.

Hutchinson L, Aitken P, Hayes T. 2002. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ 36:73–91.

Joshi R, Ling FW, Jaeger J. 2004. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. Acad Med 79:458–463.

Mathers NJ, Challis MC, Howe AC, Field NJ. 1999. Portfolios in continuing medical education–effective and efficient? Med Educ 33:521–530.

Newble DI, Jaeger K. 1983. The effect of assessments and examinations on the learning of medical students. Med Educ 17:165–171.

Norcini JJ, Blank LL, Arnold GK, Kimball HR. 1995. The mini-CEX (clinical evaluation exercise): A preliminary investigation. Ann Intern Med 123:795–799.

Norman GR. 1991. What should be assessed? In: Boud DAFG, editor. The challenge of problem based learning. London: Kogan Page. pp 254–259.

Osler W. 1903. On the need of a radical reform in our methods of teaching senior students. The Medical News 82:49–53.

Patel VL, Evans AE, Groen GJ. 1989. Biomedical knowledge and clinical reasoning. In: Patel V, Evans L, DA, editors. Cognitive science in medicine: Biomedical modeling. Cambridge, MA: MIT Press. pp 53–112.

Schmidt HG, Norman GR, Boshuizen HP. 1990. A cognitive perspective on medical expertise: Theory and implication. Acad Med 65:611–621.

Schuwirth LW, Verheggen MM, Van der Vleuten CP, Boshuizen HP, Dinant GJ. 2001. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ 35:348–356.

Van de Wiel MWJ. 1997. Knowledge encapsulation: Studies on the development of medical expertise. Department of Educational Development and Research. University of Maastricht.

Van der Vleuten CP, Schuwirth LW. 2005. Assessing professional competence: From methods to programmes. Med Educ 39:309–317.

Van der Vleuten CPM. 1996a. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1:41–67.

Van der Vleuten CPM. 2000. Validity of final examinations in undergraduate medical training. BMJ 321:1217–1219.

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