Progress testing in postgraduate medical education

Dijksterhuis MGK, Scheele F, Schuwirth LWT, Essed GGM, Nijhuis JG, & Braat DDM. Progress testing in postgraduate medical education. Medical Teacher 2009;31(10):e464-e468
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 20th 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. The latter view has been embraced in the emergence of problem-based learning curricula in the early 1970s. Initially this educational philosophy tended to marginalise 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 memorising to knowing how to use resources adequately.

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-organised knowledge database and thus requires expert knowledge. With that, the central role of knowledge in (medical) expertise has been re-emphasised.

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. More recently introduced assessment methods like the 360-degree feedback; the mini-CEX and portfolio seem to focus predominantly on practice performance. Knowledge 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 characterised 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. 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.
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 organised 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 strengths 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.\(^\text{15-17}\)

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.\(^\text{18}\)

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 blueprint. Until 2004 true-false items (TF) were used. From 2005 single-best-option multiple-choice questions (MCQ) are 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 3 pre-test review cycles and a post-test item analysis by a review committee, surrounds each question. This committee consists of 8 gynaecologists with various backgrounds, 2 trainees and 1 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 2 standard deviations are considered unsatisfactory. The results, subdivided over the above mentioned sub domains, are solely disclosed to the trainee and his/her educational supervisor. Until now, test results have no summative consequences, but are supposed to be used in a formative way during in-training assessments.

**Data collection**
Anonymised 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.

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

<table>
<thead>
<tr>
<th>TRAINING YEAR</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>mean</td>
<td>sd</td>
<td>n</td>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td>31</td>
<td>8.8</td>
<td>39</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>40</td>
<td>8.2</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>44</td>
<td>9.0</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>46</td>
<td>7.7</td>
<td>26</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>47</td>
<td>8.9</td>
<td>27</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>50</td>
<td>6.5</td>
<td>47</td>
<td>47</td>
</tr>
</tbody>
</table>

The same mean correct-minus-incorrect scores per year of testing are shown in figure 1. We used linear graphs as this summarises the data in a comprehensible way as differences between training years can be easily visualised. 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).

Figure 1  Mean correct-minus-incorrect scores per training year per year of testing (transversal)
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 (α = 0.2), highest value was reached in training year 1 in 2008 (α = 0.81). Average reliability approximated 0.50 with true-false items. This improved to a reliability of 0.65 after the introduction of single-best-option multiple-choice-questions, which is moderate. This led to the decision to set the pass-fail score to the training year mean minus 2 standard deviations. With these reliabilities, this is more than 2 standard errors of measurement below the mean. Therefore, an ‘unsatisfactory’ decision can be made with an accuracy of p < 0.05.

<table>
<thead>
<tr>
<th>Training Year</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.50</td>
<td>.52</td>
<td>.44</td>
<td>.56</td>
<td>.59</td>
<td>.36</td>
<td>.66</td>
<td>.79</td>
<td>.73</td>
<td>.81</td>
</tr>
<tr>
<td>2</td>
<td>.33</td>
<td>.53</td>
<td>.34</td>
<td>.30</td>
<td>.47</td>
<td>.63</td>
<td>.69</td>
<td>.74</td>
<td>.48</td>
<td>.66</td>
</tr>
<tr>
<td>3</td>
<td>.45</td>
<td>.39</td>
<td>.48</td>
<td>.56</td>
<td>.52</td>
<td>.58</td>
<td>.77</td>
<td>.75</td>
<td>.65</td>
<td>.64</td>
</tr>
<tr>
<td>4</td>
<td>.55</td>
<td>.43</td>
<td>.42</td>
<td>.39</td>
<td>.40</td>
<td>.61</td>
<td>.72</td>
<td>.74</td>
<td>.65</td>
<td>.52</td>
</tr>
<tr>
<td>5</td>
<td>.41</td>
<td>.70</td>
<td>.52</td>
<td>.54</td>
<td>.53</td>
<td>.54</td>
<td>.60</td>
<td>.74</td>
<td>.62</td>
<td>.56</td>
</tr>
<tr>
<td>6</td>
<td>.51</td>
<td>.34</td>
<td>.54</td>
<td>.41</td>
<td>.20</td>
<td>.42</td>
<td>.62</td>
<td>.66</td>
<td>.56</td>
<td>.61</td>
</tr>
</tbody>
</table>

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 this standard is 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. However, this seems highly unlikely, as item analyses 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\textsuperscript{18,20}.

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\textsuperscript{4}. 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 specialised knowledge is acquired that is not assessed by present testing methods\textsuperscript{21}. 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\textsuperscript{12}. 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\textsuperscript{18,20,23-25}.

So, what could be done to improve both reliability and validity? Firstly, the inclusion of zero variance items should be avoided. Creating an item bank of effective questions is highly recommended\textsuperscript{18,19}. 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\textsuperscript{18,20}. 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\textsuperscript{22,26}. 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\textsuperscript{27}.

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 specialisations or countries.

**Conclusion**

Validity and reliability of the current progress test in postgraduate obstetrics and gynaecology training is unsatisfactory. Suggestions for improvement are provided.
References

1. Flexner A. Medical education in the United States and Canada. From the Carnegie Foundation for the Advancement of Teaching 1910.
2. Osler W. On the need of a radical reform in our methods of teaching senior students. The Medical News 1903;82:49–53.