Medical Teacher, Vol. 26, No. 8, 2004, pp. 719–725

Taylor & Francis
healthsciences

Cross institutional collaboration in assessment: a case on progress testing

C.P.M. VAN DER VLEUTEN[1], L.W.T. SCHUWIRTH[1], A.M.M. MUIJTJENS[1], A.J.N.M. THOBEN[2], J. COHEN-SCHOTANUS[3] & C.P.A. VAN BOVEN[4]

[1]Department of Educational Development and Research, Maastricht, The Netherlands; [2]Department of Educational and Student Affairs, University Medical Centre Nijmegen, Nijmegen, The Netherlands; [3]Department for Educational Development and Quality Assurance, Institute for Medical Education, University of Groningen, Groningen, The Netherlands; [4]Leiden University Medical Centre, Leiden, The Netherlands

SUMMARY *The practice of assessment is governed by an interesting paradox. On the one hand good assessment requires substantial resources which may exceed the capacity of a single institution and we have reason to doubt the quality of our in-house examinations. On the other hand, our parsimonity with regard to our resources makes us reluctant to pool efforts and share our test material. This paper reports on an initiative to share test material across different medical schools. Three medical schools in The Netherlands have successfully set up a partnership for a specific testing method: progress testing. At present, these three schools collaboratively produce high-quality test items. The jointly produced progress tests are administered concurrently by these three schools and one other school, which buys the test. The steps taken in establishing this partnership are described and results are presented to illustrate the unique sort of information that is obtained by cross-institutional assessment. In addition, plans to improve test content and procedure and to expand the partnership are outlined. Eventually, the collaboration may even extend to other test formats. This article is intended to give evidence of the feasibility and exciting potential of between school collaboration in test development and test administration. Our experiences have demonstrated that such collaboration has excellent potential to combine economic benefit with educational advantages, which exceed what is achievable by individual schools.*

Introduction

Designing good test material is a resource intensive task. The amount of time and effort it takes to produce test items of acceptable quality is often underestimated. The quality of any test, especially its validity, depends probably more on a careful, systematic construction process than on any other aspect of the test. Although the technology for producing test material is widely available, particularly for written tests (Downing, 1997; Downing, 2002a; Case, 1998), most teaching staff are not familiar with it. Formal training in test construction is something only few teachers have had. It is not uncommon for examinations to be constructed at the very last moment. Quality control procedures, like pre- and post-administration review of test material, are rare in medical schools, whether this is due to lack of time, reluctance of us teachers to evaluate each other's work, or just plain ignorance of the possibility that one might use such a procedure. All in all, it should not come as a surprise that a recent empirical study of the quality of in-house examinations gave rise to disappointing conclusions (Jozefowicz *et al.*, 2002). Although more research in this area is needed, there are indications that the quality of measurement is indeed affected by the quality of test material (Downing, 2002b). Paradoxically, although we may think that the high quality of our teaching cannot but be crowned by excellent results of our students, the ultimate benchmark by which the quality of our students is measured is the quality of our examinations, and that quality leaves much to be desired.

Still another paradox is the fact that as individual teachers and schools, we jealously guard our test material, precisely because it is so time consuming and costly to produce. We prefer to keep the entire process in our own hands. We develop our own test material, we use our own system to file tests in our own (electronic) filing cabinets, we try to keep test items from the eyes of our students (probably in vain, considering the generally lively "black market" in test material among students) and we definitely refuse to share our tests with anyone else. This may well be one of the biggest wastes of resources in education.

Shared production of high quality test material would appear to be the logical solution to both paradoxes. It would enable more effort to be put into the development of high quality test material while at the same time spending by individual schools could be reduced. In addition to sharing production costs among several test developers and institutions and gains in the quality of test material, collaborative test production can be a stimulus to educational quality control, since it enables comparison between the participating institutions. In Europe in particular, this would be no small benefit, seeing that most European countries have no nationwide medical certification examinations. In summary, the sharing of test material would give a boost to the quality of examinations and education as well as to the efficient use of resources.

Correspondence: C.P.M. van der Vleuten, Department of Educational Development and Research, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: c.vandervleuten@educ.unimaas.nl

The paper resulted from collaboration between the medical schools of Maastricht, Nijmegen, Groningen and Leiden. The addresses of the schools can be obtained from the corresponding author.

ISSN 0142–159X print/ISSN 1466–187X online/00/000719–7 © 2004 Taylor & Francis Ltd
DOI: 10.1080/01421590400016464

In this paper we describe how such a strategy was adopted by several medical schools who decided to enter into a partnership for the production of an assessment tool: the progress test. Three of the eight medical schools in The Netherlands decided to collaboratively develop items for progress tests and administer the tests simultaneously. Currently, a fourth school buys the test and takes part in the synchronous test administration in anticipation of full collaboration in the production process. We will describe the progress test and its suitability for cross-institutional collaboration, the steps taken on the road to full collaboration, some samples of test results by way of illustration, and the plans for the future.

The progress test

A progress test is a comprehensive knowledge test which assesses the end objectives of undergraduate medical training as far as knowledge is concerned across all medical disciplines. With progress testing, the entire student population of a medical school sits the same paper and pencil test. Annually, four progress tests are administered to all students. A single progress test consists of approximately 250 test items of the true/false/don't know format. The percentages of basic science items (28%), clinical science items (55%) and behavioural science items (17%) in each test are fixed and based on a two-dimensional blueprint. The columns of the blueprint represent (organ) systems and the rows represent disciplines. Students' test scores are expressed as the percentage of correct minus incorrect answers (formula scoring). In a six-year curriculum, like those of all medical schools in The Netherlands, progress testing entails assessment of every student's knowledge on at least 24 measurement moments, that is four tests per year. Students keep the test booklets and are given the answer key, which means that test items are publicly available. Detailed computerized performance feedback is given to staff and students and tests are stored in an item bank. Each test is newly constructed. Although items from the item bank can be re-used, they have to go through the review procedure again, which more often than not leads to modification. Progress test results are used for decisions about student promotion and a different pass/fail standard is set for each of the 24 measurement moments. Students receive the qualification distinction, satisfactory or unsatisfactory. The qualifications obtained by a student in one academic year are combined to yield a pass/fail decision for that year based on predefined combinations of qualification patterns, with later performance being weighted more heavily.

Progress testing was originally developed for a problem-based learning (PBL) context (van der Vleuten, 1996). In the PBL curriculum of Maastricht Medical School, The Netherlands, it has been used since 1977. It is intended as a disincentive to test-directed revision. This is achieved by two features. First, there is no direct connection between the test and any specific course unit and second, the test samples all the end objectives of the undergraduate curriculum. As a consequence, it is impossible to prepare for one particular test, while at the same time all learning activities are rewarded. Furthermore, knowledge about all topics – also those learned in the past – is continuously being re-evaluated. Short-term learning strategies like cramming and rote memorization are therefore ineffective, while deep learning strategies – such as focusing on functional long-term knowledge and self-directed learning – are reinforced. In a nutshell, progress testing enables knowledge assessment through objectively formatted questions without the customary adverse effects on the learning of students. Research has shown that progress testing does indeed reinforce a deeper learning style (Til, 1997; Verhoeven, 2003) does not alter existing patterns of study behaviour after its introduction (Blake *et al.*, 1996), and offers several other proven educational benefits besides (van der Vleuten, 1996).

The use of progress testing is not restricted to PBL programmes, nor to a single institution. In any medical curriculum knowledge is an important part of competence (Norman, 1991) and thus requires assessment, regardless of the instructional format used. The blueprint of the progress test reflects the entire domain of medical knowledge and is not related to the curriculum of one particular institution. This makes the progress test eminently suitable for knowledge assessment in other than PBL medical schools (Verhoeven *et al.*, 1998), and even across national boundaries (Albano *et al.*, 1996).

Cross-institutional collaboration

Despite the clear educational advantages of progress testing, low production costs are not among its strong points. Annually, a staggering number of at least a thousand test items must be written, vetted and approved. When a proper review process is in place, the overall production time of one item may easily exceed one hour. In order to obtain one thousand approved questions, many more items must be written. Re-using questions reduces production time only marginally, because the rigorous review process requires those questions to be subjected to the same review process as new ones, which often necessitates rewriting items. The prospect of combining high test quality with acceptable costs was a strong driving force behind cross-institutional collaboration.

The universities of Nijmegen and Groningen, who joined with Maastricht in a test producing partnership, had a more important, educational, motivation for seeking collaboration. They were in the process of curricular innovation aimed at increased interdisciplinary integration and problem and skill orientation. A non-course dependent knowledge test as an overarching continuous assessment format would allow teachers to design course-dependent assessment formats with less focus on knowledge and more opportunities to include exercises that better reflect the learning tasks in the course, like simulations, hands-on performance tasks, case reports, etc. This would obviate the criticisms about relevance and acceptability which are generally levelled at objectively formatted questions by both students and teachers. Another important reason for joining the collaborative effort was the prospect of quality control of educational outcomes.

We will describe the steps taken in establishing the partnership between the universities of Maastricht, Nijmegen and Groningen, the current state of affairs and the plans for the future. It must be noted that the University of Maastricht has used progress testing since 1977 and had in place a well-established test production line.

This circumstance afforded the medical schools ample time for a gradual process towards full collaboration.

Preparatory activities, informing staff and staff development

The collaboration started with the medical schools of Maastricht and Nijmegen. In Nijmegen a curriculum revision was initiated in 1995. It involved the introduction of horizontally and vertically integrated course units, small group work based on assignments as well as lectures and practicals, and clinical skills training in a clinical skills training unit (Holdrinet, 2002). The curriculum planners were keen to introduce progress testing as part of the innovation, but they were also aware that staff were not entirely convinced of the need to do so. Two key activities were undertaken: staff meetings in which curriculum planners introduced the concept of progress testing and two pilot test administrations of a Maastricht progress test in 1994 and 1995. The results were used in staff development sessions. The pilots also gave insight into the administrative consequences. Staff were already burdened by the curriculum renewal, therefore the Nijmegen curriculum planners decided not to contribute test items immediately, but to start by buying tests from Maastricht. Since September 1995, the schools of Maastricht and Nijmegen have administered progress tests simultaneously and used the same standards for pass/fail decisions. In the meantime curriculum planners at Groningen Medical School had implemented a PBL-like curriculum and were developing their own progress test. Soon they discovered that setting up and maintaining a production line was no easy task and that a joint effort was likely to be the best way forward. They also started out by buying the progress test, beginning in September 2000. Both Nijmegen and Groningen had difficulty setting up a centralised and co-ordinated item production process, due to the costs involved and – perhaps more importantly – the unfamiliarity of staff with a central test organization. For Maastricht Medical School a prerequisite for engaging in co-production was the preservation of item quality. Achieving this was the next step in the collaboration.

Setting up a review process and production line

Progress test construction at Maastricht follows a tight production schedule with a rigorous review procedure and two quality control cycles, one before and one after test administration (Verhoeven *et al.*, 1999). The items contributed by the departments are submitted to the progress test review committee. The eight committee members have backgrounds in basic, clinical and behavioural sciences. In the first quality control cycle wording, content and relevance of the submitted items are scrutinized. All test items must be accompanied by a literature reference. There is frequent communication between reviewers and item authors. The approved items are then pooled to compose the test and test administration follows. After the test students are expressly invited to submit any critique of test items, preferably supported by evidence from the literature. Student comments together with psychometric information from the test results are reviewed by the progress test review committee. The committee decides whether any items should be dropped. After this second quality control cycle, the final test scores are calculated and detailed feedback is given to students, item authors and departments. All information is stored in the item bank. Note that students play an important part in the quality control cycle. Pre-test review takes six four-hour meetings of the review committee. On average more than 80% of the questions are altered, with 56% of the changes concerning content (Verhoeven *et al.*, 1999). Post-test review involves one meeting which lasts one hour. On average, 2% of questions are rejected, primarily as a result of student comments. Membership of the review committee is a teaching role that is credited with 8 hours per week. This may seem excessive, but committee members need much time to prepare for committee meetings by reading items, suggesting revisions and consulting textbooks to check item content.

The initial idea was to set up local progress test review committees in the partner schools modelled on the Maastricht review committee. In addition, the chairs of the local committees would constitute an umbrella committee, which would decide on the final composition of the test. Setting up a rigorous review process and an item production line with comparable production capacity in each of the partner schools was the most challenging operation of the collaboration. Committees were appointed and working conferences and workshops were held both in Maastricht and in Groningen and Nijmegen to train the new local review committees. Faculty staff had to be motivated to contribute questions. It was by no means easy for staff to start writing questions for a test not connected to any particular course but instead related to the – ill-defined – end objectives of undergraduate medical education. Many test items had to be rejected because they were too detailed or irrelevant. In Nijmegen production started in September 1999 and in Groningen in September 2000. Parity in production (about 300 questions per institution per year) was not achieved until September 2003. This illustrates how hard it is to start from scratch a centralized test production process including a quality control procedure in a medical school. Nevertheless, the production cycle is becoming an established feature in the different schools and high-quality items are generated that are acceptable to all partners.

In 1999, the University of Leiden, also in the wake of a curricular revision, expressed an interest in joining the collaboration. Considering the effort needed to build up test production with the three existing partners and the fact that Leiden was in the middle of a curriculum innovation, it was deemed prudent to postpone full participation in test production. Leiden started by buying three tests annually (skipping the first – September – progress test) and decided recently to continue doing so for another three to four years, although full participation is still envisaged for the not too distant future.

The collaborative efforts have been successful. Since September 2000 four medical schools in The Netherlands have annually administered to all of their students simultaneously four (in Leiden three) progress tests, which are produced collaboratively by three of the schools.

Analysing and comparing performance

Every three months a large data set comes available, which can be used for all sorts of purposes. To illustrate the type

of data, we will present two examples, one at the level of the total test and one at item level.

Figure 1 shows the average scores (percentage correct minus incorrect) of the four medical schools (Maastricht, Nijmegen, Groningen and Leiden) for the May 2003 progress test.

Note that for three schools the results of all six year groups are presented, whereas for Leiden, years five and six are missing. These groups do not sit the progress test, because they follow the 'old' curriculum. When the new curriculum is fully implemented, in two years time, the tests will be administered to all students. For reasons of clarity, the results are presented as line plots. It should be noted, however, that the data are quasi-longitudinal, with consecutive average year group scores originating from different cohorts of students. The graph shows a steady growth of knowledge for all medical schools, except for Maastricht, which appears to show a slight dip in year six. While the Maastricht and Groningen score patterns are rather similar looking, Nijmegen shows a somewhat smaller growth rate in years three through six. This may be partly attributable to the fairly large influx in years 3 and 4 of foreign students and students with a biomedical first degree, for whom the progress test is new. Leiden's growth rate resembles that of Maastricht and Groningen, although the level of knowledge is lower.

Figure 2 provides an illustration of performance on a single item for three medical schools. It shows interesting differences between the schools. It would be logical to assume that students learn about a direct inguinal hernia in the clinical years (year 5 and 6). However, the Groningen data show an increase in correct responses in third year with a further increase during the clerkship years. Nijmegen students start scoring in year 4 and score highest in year 6. An explanation may be that in Nijmegen acute abdominal pain is addressed in the emergency medicine unit early in year 4. The surgical clerkship starts at the end of year 5 and continues in year 6. Abdominal complaints are also addressed in other clerkships in that year. In Maastricht we see a different and rather unexpected pattern: an increase in year 5 (the surgical clerkship is in year 5) followed by a lower correct score and a substantially higher incorrect score in year 6. It would appear that some students' knowledge about this relevant (and straightforward) clinical phenomenon has disappeared or that students have become confused. This sort of information is highly interesting for discussions of ''what's going on'' in the curriculum and these results are fed back to the departments concerned.

Caution should be exercised in interpreting these cross-institutional data, because different medical schools may have different ways of assigning students to year groups. Methods are being developed that will make comparisons less vulnerable to local administrative differences.

Plans for the future

A major item on the agenda is how to use the synergy from the collaboration to further professionalize and optimize measurement information. There are a number of plans, which include:

- Changing the item format from true/false items to multiple choice questions: True/false items have some intrinsic disadvantages and there is agreement that simple single best answer or matching items offer better alternatives (Case, 1998).
- Increasing the number of context-based or vignette-based questions: There are still too many progress test questions that focus too much on isolated, factual knowledge. We are striving to include more contextually rich questions, which trigger reasoning and the application of knowledge. This will require intensive training of staff, since all (modest) efforts to produce such questions have met with limited success and the ''natural'' tendency of most authors is to produce recall-type test items.
- Statistical equating of tests across time and institutions: Variation in test difficulty is the biggest impediment to comparing performance across tests. This means that equating strategies are needed. We are also developing a statistical model that incorporates the growth aspect

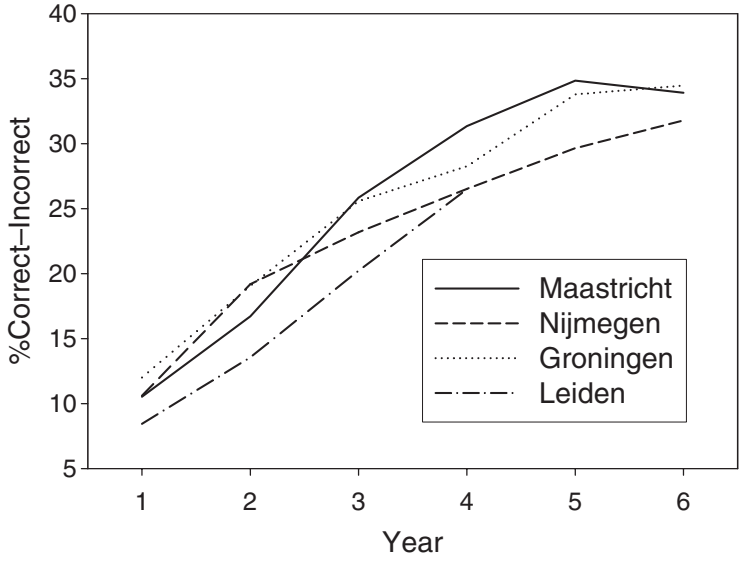

Figure 1. Average total test scores (percentage correct minus incorrect) on the May 2003 Progress Test of the students of each of the four schools.

Question: A direct inguinal hernia is a protrusion of the peritoneum due to a weakening of the transversalis fascia. The localisation of this hernia, in relation to the epigastric vessels, is:

- medial

Answer key: Correct

Last test: March 2003

Item statistics:

Year of Training	University of Maastricht			University of Groningen			University of Nijmegen		
	% Correct	% Incorrect	% Do not know	% Correct	% Incorrect	% Do not know	% Correct	% Incorrect	% Do not know
1	1	0	99	2	0	98	1	1	98
2	3	4	94	2	1	96	3	1	95
3	8	6	86	47	13	40	12	12	76
4	13	11	76	40	9	51	34	13	53
5	53	19	28	55	13	32	30	19	52
6	38	26	36	64	16	20	54	18	28

Figure 2. Sample question and item scores across three medical schools.

of progress testing. Equating procedures easily go beyond the expertize and possibilities of individual medical schools. Hopefully, collaboration will generate the means for developing a more professional strategy in this respect.

- Dynamic web-based score and item feedback from the central database: The web provides an ideal instrument for providing feedback to students as well as to the faculty organization (item writers, departments, committees), particularly when users are geographically spread across the nation. We are working on ''static'' web-based feedback, that is software that generates web-based overviews of users' performance after each test. The next step will be ''dynamic'' feedback, which allows users to generate a performance profile across or within tests. The statistical model that is being developed should also make it possible to predict future performance from past performance. These measures could have a substantial impact on the formative value of progress testing, thereby maximizing the benefits of this continuous cross-sectional and longitudinal assessment method.
- Computerized test administration: The long-term aim is to put an end to the need to rent colossal test administration facilities (sport accommodations, theatres, etc) to assess all students concurrently. This would also improve the acceptability of the test format to students. We realize that this is a long term goal since it will not only require sufficient technical facilities, but also a much larger item bank (and thus production capacity) to allow for multiple test forms to be administered consecutively. The synergy of collaboration should eventually make this possible.

A second item on the agenda is extending the collaboration to the other Dutch medical schools and to the schools in Flanders (the Flemish language resembles the Dutch language sufficiently to use the test without translation). The University of Leiden will take part in test production in the near future. One Flemish school has experimented with pilot administrations and is contemplating joining the collaboration. In the future more schools will be invited to participate. Experiences so far have taught that each additional partner means several years of investment before any return can be expected. Actually, expanding the partnership is the idealistic component of the drive towards collaboration.

Discussion

Cross-institutional collaboration on progress testing has demonstrated the added value of the sharing of test material. Through the collaboration we have achieved greater efficiency in test production. In Maastricht the review committee could be reduced by two members. The new partners had to invest in an item production and review process, but this investment was only a fraction of what would have been needed had they set up their own separate progress testing system. In addition to savings on resources, a number of educational benefits have been accomplished. First, it is very likely that the quality of the test material is superior to that of most local in-house examinations. Second, the introduction of a quality control process has contributed to the professional development of staff involved in test development and item writing. Hopefully, their new expertize will also have a favourable impact on other in-house examinations. Third, performance can now be monitored across institutions. With the progress test as an assessment format, our ambitions stretch further than the mere sharing of test material. The synchronous and continuous administration of the same tests to all students in all four institutions four (and three) times per academic year yields valuable information to evaluate and improve medical education. Finally, the synergy of the current and hopefully still widening collaboration should give an impetus to further professionalization of the assessment procedure. An individual medical school will never be able to achieve the professional quality of, for example, certification agencies and institutions (Prideaux & Gordon 2002). Although we do not pretend that we are likely to ever match

their quality, we firmly believe that we will be able to achieve a sizeable improvement on what would ever be attainable by individual medical schools.

Unlike for example in Germany or the United States, there are no national examinations in undergraduate (or postgraduate) medical education in The Netherlands. The described collaboration in progress testing can be said to have taken us partway towards some form of national licensure examination. Moreover, this approach might avoid some of the adverse effects of national examinations, particularly their huge impact on curricula. We have used a bottom-up approach, with schools retaining ownership of the entire process. This bottom-up process is probably facilitated by the comparability in quality of the eight medical schools in The Netherlands. The schools are all state-funded and admit (through a national grade-weighted lottery system) students from a highly selective secondary school system with national examinations. This situation probably lessens the need for national licensure. Our motivation is therefore not primarily to achieve a sort of national licensure system, but rather to do a better job on assessment, which includes the use of relevant curricular feedback to improve the educational quality of our schools. At the same time, the fact that we are able to guarantee the quality of our graduates to society is a valuable add-on of that process.

Having learned from our experiences with joint progress testing, we now envisage future collaboration on other assessment forms. For example, all medical schools in The Netherlands use some form of OSCE assessment (van der Vleuten *et al.*, 1995) and several schools use a computerized system of case-based testing (Schuwirth, 1998). A recent initiative has led to nationwide standardized assessment of professional behaviour in the Dutch medical schools (Anonymous, 2002). To put it briefly, there are manifold and excellent opportunities for pooling efforts in those areas. With our current experience, we know not only that collaboration costs time and effort, but also that with patience and perseverance success is within our grasp.

We were very pleased to hear of recent initiatives to share item banks, both internationally (Prideaux & Gorden 2002; http://www.hkwebmed.org) and within the UK (http://orgs.man.ac.uk/projects/ucam/). We endorse such initiatives wholeheartedly. The time has come for us to start sharing our testing materials on a larger scale, so that we can stimulate the professionalization of assessment, improvement of the quality of assessments, and put a stop to the wastage of educational capital.

Practice points

- In-house examinations often lack quality due to unprofessional development of the test material.
- Constructing good test material requires substantial resources.
- Sharing test material across schools and institutions provides an economic benefit that would allow a more quality-driven process of test construction.
- Sharing achievement results across schools and institutions may provide an impulse for monitoring educational quality.

Notes on contributors

Cees van der Vleuten is professor of education and chair of the Department of Educational Development and Research, Faculty of Medicine, University of Maastricht, The Netherlands.

Lambert Schuwirth is assistant professor at the Department of Educational Development and Research, Faculty of Medicine, University of Maastricht, The Netherlands.

Arno Muijtjens is assistant professor at the Department of Educational Development and Research, Faculty of Medicine, University of Maastricht, The Netherlands.

Arnold Thoben is academic counsellor, student advisor and secretary of the Committee on Progress Testing of the Department of Educational and Student Affairs, University Medical Centre Nijmegen, Nijmegen, The Netherlands.

Janke Cohen-Schotanus is head of the Department for Educational Development and Quality Assurance, Institute for Medical Education, Faculty of Medical Sciences, University of Groningen, Groningen, The Netherlands.

Cees van Boven is emeritus professor of Medical Microbiology, Leiden University Medical Centre, Leiden, The Netherlands.

References

Albano, M.G., Cavallo, F., Hoogenboom, R., Magni, F., Majoor, G., Manenti, F., Schuwirth, L., Stiegler, I. & van der Vleuten, C.P.M. (1996) An international comparison of knowledge levels of medical students: the Maastricht progress test, *Medical Education*, 30, pp. 239–245.

Anonymous (2002) Professioneel Gedrag (Professional Behaviour), Report of the Project Team Consilium Abeundi (Utrecht, Association of Dutch Universities (Vsnu)).

Blake, J.M., Norman, G.R., Keane, D.R., Barber Mueller, C., Cunnington, J. & Didyk, N. (1996) *Academic Medicine*, 71, pp. 1002–1007.

Case, S.M. (1997) Assessment truths that we hold as self-evident and their implications, in: Scherpbier, A.J.J.A., Van Der Vleuten, C.P.M., Rethans, J.J. & Van Der Steeg, A.F.W. (Eds) *Advances in Medical Education*, pp. 2–6 (Dordrecht, Kluwer Academic Publishers).

Case, S.M. & Swanson, D.B. (1998) *Constructing Written Test Questions for the Basic and Clinical Sciences* (Philadelphia, National Board of Medical Examiners).

Downing, S.M. (2002a) Assessment of knowledge with written test forms, in: Norman, G.R., Van Der Vleuten, C.P.M. & Newble, D.I. (Eds) *International Handbook of Research in Medical Education*, Pp. 647–672 (Dordrecht, Kluwer Academic Publishers).

Downing, S.M. (2002b) Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77, pp. S103–S104.

Downing, S.M. & Haladyna, T.M. (1997) Test item development: validity evidence from quality assurance procedures, *Applied Measurement in Education*, 10, pp. 61–82.

Holdrinet, R.S.G. (2002) Realisering Van Een Vernieuwd Nijmeegs curriculum (Realising a renewed Nijmegen curriculum), in: *Umc St Radboud* (Nijmegen, University of Nijmegen).

Jozefowicz, R.F., Koeppen, B.M., Case, S.M., Galbraith, R., Swanson, D.B. & Glew, R.H. (2002) The quality of in-house medical school examinations, *Academic Medicine*, 77, pp. 156–161.

Norman, G. R. (1991), What should be assessed, in: Boud, D. & Feletti, G. (Eds) *The Challenge of Problem-Based Learning*, pp. 254–259 (New York, St. Martin's Press).

Prideaux, D. & Gordon, J. (2002) Can global co-operation enhance quality in medical education? Some lessons from an international assessment consortium, *Medical Education*, 36 (5), pp. 404–405.

Schuwirth, L.W.T. (1998) *Computerized Case-Based Testing: An Approach to the Assessment of Medical Problem Solving*. Doctoral Dissertation, University of Maastricht, Maastricht.

TIL, C., VAN DER VLEUTEN, C.P.M. & VAN BERKEL, H.J.M. (1997) Problem-based learning behavior: the impact of differences in problem-based learning style and activity on students' achievement. Chicago: Annual Meeting of The American Educational Research Association, Eric No. Tm026783 (Ed409333).

VAN DER VLEUTEN, C.P.M., SCHERPBIER, A.J.J.A. & VAN LUIJK, S.J. (1995) Use of Osces in The Netherlands, in: Rothman, A.I. & Cohen, R. (Eds) *Proceedings of the Sixth Ottawa Conference on Medical Education*, pp. 320–321 (Toronto, University of Toronto Bookstore Publishing).

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. 103–110.

VERHOEVEN, B.H., VAN TIL, C.T., VERWIJNEN, G.M., SCHERPBIER, A.J.J.A. & VAN DER VLEUTEN, C.P.M. (2003) The consequential validity of the progress test: an investigation into the relationship between test results and problem-based learning behaviour in: Verhoeven, B.H. (Ed.) Progress Testing. The Utility of an Assessment Concept, (PhD Thesis, Groningen, Stichting Drukkerij G. Regenboog).

VERHOEVEN, B.H., VERWIJNEN, G.M., SCHERPBIER, A.J.J.A., SCHUWIRTH, L.W.T. & VAN DER VLEUTEN, C.P.M. (1999) Quality assurance in test construction. the approach of a multidisciplinary central test committee, *Education for Health*, 12(1), pp. 49–60.

VERHOEVEN, B.H., VERWIJNEN, G.M., SCHERPBIER, A.J.J.A., HOLDRINET, R.S.G., OESEBURG, B., BULTE, J.A. & VAN DER VLEUTEN C.P.M. (1998) An analysis of progress test results of Pbl and Non-Pbl students. *Medical Teacher*, 20(4), 310–316.