

EVALUATION

## Quality Assurance in Test Construction: The Approach of a Multidisciplinary Central Test Committee

B. H. VERHOEVEN<sup>1</sup>, G. M. VERWIJNEN<sup>1</sup>,  
A. J. J. A. SCHERPBIER<sup>1</sup>, L. W. T. SCHUWIRTH<sup>2</sup> &  
C. P. M. VAN DER VLEUTEN<sup>2</sup>

<sup>1</sup>*Skillslab, Faculty of Medicine, Universiteit Maastricht; and*  
<sup>2</sup>*Department of Educational Development and Research,*  
*Faculty of Medicine, Universiteit Maastricht, The Netherlands*

**ABSTRACT** *What and how students learn is to a large extent directed and shaped by tests; and test results can have far-reaching consequences for students' progress. For both these reasons, it is vital that test construction be done with great care. In 1997 a report was published in our country based on site visits to all Dutch medical schools. This report recommended that central test review committees be set up to monitor test quality at all medical schools. The pivotal feature of the approach used by such committees is peer review. Peer review is widely used in quality assurance of research output, but it has only rarely been used to promote quality in health professions education. In this report we illustrate the workings of central test review committees, focusing on the approach used by the Progress Test Review Committee at the Maastricht Faculty of Medicine. This committee, which has been in operation since 1977, is responsible for the four examinations given to all medical students every year. Here, we highlight the steps and approaches that we have found necessary to achieve effective peer review and to produce tests of consistently high quality. These tests are designed to track the students' progress toward being able to answer a set of questions that a newly graduated MD should be able to answer correctly, as judged by an interdisciplinary group of faculty members from our school.*

The results of this study were previously published in Dutch, in the "Bulletin Medisch Onderwijs," a Dutch journal on medical education, in Volume 17(2), 1998, Pages 62–72. The revised version of that report presented here is done with the permission of the authors, editor, and publisher of the earlier report.

Address correspondence to: G. M. Verwijnen, Skillslab, Universiteit Maastricht, P.O. Box, 616, 6200 MD Maastricht, The Netherlands, Tel: (31)-43-3881771. Fax: (31)-43-3618612. E-mail: m.verwijnen@sk.unimaas.nl

**For Commentary, see pages 61–62.**

## Introduction

The central objective of cognitive testing is obtaining a reliable and valid assessment of students' intellectual achievements and competencies. In view of the impact of test results on students' progress it is vital to maximize the quality of the information provided by the tests so as to minimize the number of false-positive and false-negative decisions made about students. In addition, what and how students learn is in large part influenced by tests (Newble & Jaeger, 1983; Frederiksen, 1984). Thus, careful test construction is of the essence. The construction of good tests requires specific skills and experience, which are neither easy to acquire nor widely available (Cox, 1988; Ebel & Frisbie, 1991; Haladyna & Downing, 1989). Few teachers view test construction as an attractive task. Yet, most teaching institutions do have faculty members who are interested in this aspect of education and who are willing to acquire the necessary skills.

To ensure fair assessment, it is advisable to separate the roles of teacher and examiner as much as possible. At the Maastricht Faculty of Medicine these considerations prompted the decision to set up test review committees in 1977. These committees are responsible for quality assurance of both the test's design and the test's content. For every test there are committees composed of faculty who have shown a special interest in testing. The test committees are chaired by members of the Task Force on Assessment (TFA) of the Faculty of Medicine, which includes physicians and test experts. The TFA's objective is to develop, implement and investigate the assessment system. Over the years other Faculties of Maastricht University have followed this example by setting up similar test evaluation committees. There are also national committees on test assessment, such as the National Board of Medical Examiners in the USA, and the Institut für Medizinische und Pharmakologische Prüfungsfragen in Germany (Stokes, 1967; Kraemer *et al.*, 1976; Hubbard, 1978). A 1997 site visit report in the Netherlands recommended that all of this country's medical schools follow the Maastricht example and set up similar central test committees. Little has been published on the work of such test committees (Downing & Haladyna, 1997). To illustrate the workings of test committees, we describe the approach developed by the Maastricht Progress Test Review Committee (PTRC) and the process we follow in producing our progress tests.

## The Progress Test

Wijnen introduced the concept of progress testing at this school in 1976 and since the 1977–78 academic year progress tests have been a fixture in the Maastricht Faculty of Medicine's assessment programme (Van der Vleuten *et al.*, 1996b). Four times each academic year all students (Years 1–6) take the progress examination simultaneously.

Progress testing is based on the idea that desirable learning behaviours, such as those involved in our school's self-directed problem-based learning, should not be hampered by formal assessments during day-to-day learning. The progress test occurs quarterly and is aimed at identifying increases in students' medical knowledge within the framework of the final objectives of undergraduate medical education. The progress test is composed of questions that a newly graduated MD should be able to answer correctly, as judged by our faculty. A new test, consisting of new and (adapted) previously used items, is constructed for each examination. The test results reflect how far students have progressed towards the final (cognitive) objectives of undergraduate medical education, and thus yield valuable information to guide individual students and evaluate the curriculum (Van der Vleuten *et al.*, 1996a, 1996b). For practical reasons (large numbers of both items and students) the test items are of the True/False/Question Mark format (Ebel & Frisbie, 1991). Progress tests contain some 250 items covering fifteen categories<sup>1</sup> derived from the International Classification of Diseases (ICD) (Van der Vleuten *et al.*, 1996b). The test blueprint sets out the required number of items for each category (Verwijnen *et al.*, 1982). The distribution of the items among the categories is based on morbidity data, the amount of space devoted to the various subjects in general medical textbooks, and a survey among the departments. There are no guidelines on how the items for each category are to be distributed across the different disciplines.

### **Item Bank and Item Production**

The departments devise the items, including a reference for each one, enabling students to read background materials after the test. All new items are entered into a central, partially computerized item bank, without being evaluated. Each is entered into one of the fifteen categories and given departmental and category labels. Some eight months before the date of the progress test, some 400–500 items are drawn from the item bank. The items are randomly selected by category in accordance with the distribution set out in the test blueprint. To ensure that spare items are available in case items are rejected, the number of items drawn per category exceeds the number required. The number of items that a particular department has contributed to the item bank partly determines how many of that department's items are included in the draw. That is, the more items, the greater the chance that a particular department's contributions are included among the items drawn. If a department fails to keep its share of items in the item bank at an adequate level, it runs the risk of being under-represented in the progress test. To prevent over-representation, the items of any one department may take up no more than 6% of the item bank at the time of the draw. If a department exceeds this percentage, surplus items are randomly selected and excluded from the final draw.

The items that are drawn are assessed by the PTRC and, after approval, included in the test. Unless items are rejected before or after the test, they remain in the item bank. Items can be re-used after a minimum of three years. As a result, the item bank is always changing. Over the years all departments together have contributed approximately 19,000 items. Since 1976 over 6000 items have been rejected, leaving some 13,000 items available for inclusion in tests. After each of the four annual draws, the departments receive a listing of the number of their items in the item bank by category. Departments are free to add new progress test items at any time.

### **The Progress Test Review Committee**

The PTRC is responsible for constructing the progress tests and for quality assurance. The committee is comprised of the chairperson, deputy chairperson, and six members. The chair and deputy chair are both physicians and members of the TFA. The members represent the pre-clinical, clinical, and behavioural sciences. The term of membership is four years and can be extended by one additional term. This relatively long term of membership helps ensure that full use can be made of the expertise that committee members acquire on the job. Administrative and logistic support is provided by one staff member, who is also responsible for the item bank and test administration.

### **Approach Used by the PTRC**

The PTRC follows a tight production schedule, with times and dates of all its activities being set well in advance. All teachers who are involved in item production and quality control are given the production dates at the start of each academic year. The production cycle of one progress test takes 30 weeks (Table 1). The PTRC meets twice weekly for three to four hours. Good planning is crucial for co-ordinating the production schedules of the four tests needed each year.

### **Pre-test Review**

The 400–500 items that are drawn from the item bank for each progress test are scrutinized by the PTRC. Each committee member is responsible for the items pertaining to one to four categories. Initially, committee members individually review the items in their assigned categories. Then, they discuss and check the results of these initial reviews in plenary sessions of the committee.

A first round of six meetings is reserved for these review sessions. The committee meets in a room with a large number of books available for checking the items' content validity. The review takes account of whether the item is

**Table 1.** Production cycle of one progress test

Week	Activity
1, 2	Entry into item bank, new items are processed
3	Item bank update and item draw
4, 5	Assessment by individual members of the PTRC
6-8	First round of six PTRC meetings (two per week)
9	Prepare, collect and send correspondence to the departments
10-12	Consultations with the departments
13-15	Second round of six PTRC meetings (two per week)
16, 17	Items with adaptations are processed/prepare item selections by PTRC members
17	Check of adaptations and selection of test items (all members of the PTRC)
18	Prepare draft test
19	Check draft test and report (chairperson and deputy chairperson of the PTRC)
20	Prepare final test
21-25	Test is printed/logistic preparations for test
26	Test (Wednesday 9:00 a.m. to 1:00 p.m.)
27	Process results and students' comments/prepare post-test PTRC meeting
28	PTRC meeting/consultations with departments/calculate final scores
29	Report to students and committees/prepare item report
30	Item report to departments

formulated clearly and unambiguously, whether the content is correct, and whether the information assessed is adequately and unequivocally documented. If an item fails to meet any of these requirements, specific editorial changes are suggested and the contact person of the department that supplied the item is consulted, either in writing or in person. An item can only be adapted or removed after consultation with the department concerned. The departments are responsible for the quality of item content. A second round of six plenary committee meetings is planned after the initial consultations with the departments. At these sessions, previously questioned items may be approved or removed, or more consultations may be initiated. Since many items require more than one round of consultations, and since the committee is working on multiple progress tests simultaneously, the PTRC is usually working on a thousand items at a time.

A final meeting of the PTRC is held to compose the draft progress test. Every committee member makes a final draft for his or her own categories. During this stage, members check whether all adaptations have been processed accurately, and a selection is made from the available, approved items. How many items a department eventually contributes to the test depends both on the department's initial number of items in the item bank and its co-operation with the PTRC in test production. A department that fails to respond adequately to the PTRC's comments runs the risk that fewer of its test items will be available for inclusion

in the progress test. Every member of the PTRC selects a small surplus of items for the final draft, in case items need to be removed in the process of constructing the definitive test.

The final draft of the complete progress test is again checked and corrected by the chairperson and the deputy chairperson. Originally, this was intended as a final check for catching typing errors and overlapping items. Experience has taught us that even at this stage serious problems are detected. Thus, this final round now includes a check for incorrect wording, references, and content-related errors. The results of this check are reported in writing to the PTRC members and discussed in a plenary committee meeting. Subsequently, key distribution and cluster distribution are determined. By removing items the committee aims at an even distribution of keys (50% 'true', 50% 'false') and the desired distribution among clusters (40% pre-clinical science/40% clinical science/20% behavioural science). The chairperson performs a final check of the accuracy of all adaptations and the test is then ready for printing.

The large majority (over 80%) of the items drawn from the item bank (both new items that have never been assessed and items from previous tests) are not suitable for inclusion in the progress test without adaptation. In almost all cases (97%) adaptation concerns the way the item is worded. Over half (56%) of the adaptations also involve content-related changes.

### **Post-test Review**

After the test is administered, the items are reviewed again. Despite the careful pre-test review procedure, we typically discover that both the departments and the PTRC have missed some problematic items. Students can alert the PTRC to problem items. To this end, students are given the opportunity to participate in the assessment process, as has been recommended by others (Bandaranayake & Cox, 1988; Prince & Visser, 1997). They can offer comments until one week after the test. To avoid readability problems, only typewritten comments are acceptable. When a student includes references in his or her comments, copies of the referenced texts must be enclosed. On average, 4% of the students comment on some 20% of the items in each test.

As a further step in identifying problematic items, item analyses are done on the test results (Ebel & Frisbie, 1991). The chairperson and deputy chairperson judge every item, taking account of the item analyses and students' comments. Students' comments are evaluated by checking the references provided. If students' arguments are valid, the item is put on the agenda of a plenary PTRC meeting. The agenda for this meeting also includes all items with an unusual answering pattern, such as: (1) items for which there has been no increase in correct answers among all students; (2) items for which there has been a striking increase or decrease in correct answers by a particular class; (3) items that are answered incorrectly by over 30% of sixth year students; or (4) items that are not answered by more than 50% of sixth year students.

Ten days after the test these items are discussed in a final plenary PTRC meeting. Committee members receive the items together with each item's history and the chairperson's and vice chairperson's recommendations as to how each item should be dealt with. Discussions can result in decisions to withdraw items or changing their keys. If the PTRC considers either such action, the department that contributed the item is consulted. The department and the committee must reach agreement on whether to leave the item as it is, change the key, or withdraw the item. No item is removed or key changed without permission from the department that contributed the item. After these consultations with the departments, the definitive test scores are calculated. The items that have been withdrawn are not considered in calculating the scores. An average of 5.5% of items are withdrawn, but this percentage varies substantially across tests, with percentages ranging from 0.4% to 11.5%. The number of key changes average 0.7%, ranging from 0% to 2%. After the key changes have been incorporated and the eliminated items removed, there always remain items that relatively few students have answered or that relatively many students have answered incorrectly. At the end of the sixth year (test 24) an average of 15% of items (about 34) are left unanswered by more than 50% of sixth year students. There are considerable differences, however, among progress tests. The number of items that are answered incorrectly by more than 30% of sixth year students remains fairly stable at around 16% (about 36 items).

## **Reporting**

The results are reported to the Certifying Committee, which formally establishes the results and announces them to the students. The students receive a form listing both their own individual scores and the mean scores of their class (Figure 1).

The scores are given by category and by discipline and are expressed as percentages of the highest possible scores. The form lists both the correct/incorrect/? scores and the correct-minus-incorrect scores. In addition to the individual scores a ' + ' or ' - ' indicates whether the student has scored high or low compared to the full class. At the bottom of this list the overall score (correct-minus-incorrect) and the qualification (pass/fail) are entered. The form also states which items were withdrawn and which keys were changed.

The results are also reported to the Education Committee, the TFA, and all departments. All bodies involved receive a list of the mean scores by class for the entire test and by categories, departments and departmental clusters that contributed the items. The report includes the so-called growth curves and the results on the progress tests of the current academic year (Figure 2).

In addition, every department receives an item report for all of their items included in the test. For each item the definitive text, the answering profile, supplementary item-analysis parameters, any student comments, and history

RESULTS BY CATEGORY (Percentages)													
Category	no. of items	individual				class (n=215)							
		Correct	Incorrect	?	C-I	Correct	SD	Incor.	SD	?	SD	C-I	SD
1 Respiratory	28	50++	7-	43	43++	37	12	23	10	41	16	14	16
2 Hematol and lymph	14	36	14	50	21	43	15	14	12	43	20	28	19
3 Muscle and skeleton	18	44+	22++	33--	22	35	13	11	10	54	17	23	15
4 Mental health	17	53-	29++	18	24--	62	14	13	9	25	14	49	19
5 Reproduction	16	13-	31+	56+	-19--	41	15	20	12	38	18	21	20
6 Cardiovascular	28	71++	14	14-	57++	50	13	16	9	34	16	34	15
7 Hormone and metabol	16	19-	13	69+	6	28	17	16	11	56	21	11	19
8 Skin and conn. tissue	10	50+	0-	50	50++	31	19	13	14	56	24	17	22
9 Social factors illness/health	14	64	14	21	50	58	14	16	11	26	16	43	18
10 Gastrointestinal	23	13-	4-	83++	9	24	13	14	11	62	20	9	13
11 Renal and urinary	18	11-	17	72+	-6-	23	15	16	10	62	19	7	16
12 Nervous and sensory	21	33	14-	52	19	35	13	21	11	44	18	15	16
13 Other	11	9-	18	73++	-9-	39	17	12	12	49	19	27	22
14 Research and methodology	9	56+	22	22-	33	44	16	15	14	41	22	29	21
15 Indiv. factors illness/health	6	50	50++	0-	0-	57	20	22	15	21	20	35	29
Total	249	39	16	45	22	39	9	16	7	44	14	23	8
RESULTS BY DEPARTMENT (Percentages)													
Department	no. of items	individual				class (n=215)							
		Correct	Incorrect	?	C-I	Correct	SD	Incor.	SD	?	SD	C-I	SD
1 Anatomy/Embryology	10	30-	20	50+	10-	45	16	19	13	37	19	26	22
2 Biochemistry	15	33	7-	60	27+	31	18	16	11	53	22	15	20
3 Biophysics	0	-	-	-	-	-	-	-	-	-	-	-	-
4 Pharmacology	10	50+	20	30-	30	37	18	18	13	45	20	18	24
5 Physiology	20	55	5--	40+	50++	48	14	23	12	30	16	25	20
6 Genetics/Cell biology	10	0-	30+	70++	-30--	49	18	20	13	31	21	29	23
7 Immunology	9	22-	11+	67+	11--	42	20	4	8	54	23	39	21
8 Medical microbiology	18	6-	11	83++	-6-	25	13	16	12	59	20	9	15
9 Pathology	9	33	11	56	22	30	17	13	12	57	20	17	21
Total preclinical sciences	101	30-	13-	57+	17	38	11	17	7	45	15	21	10
10 General surgery	10	60++	10	30-	50++	34	17	14	13	52	23	19	20
11 Cardiology	12	67++	8-	25	58++	49	16	16	11	34	19	33	20
12 Dermatology	9	44+	0-	56	44++	30	20	12	15	58	25	19	24
13 Obstetrics/Gynecology	10	20-	50++	30-	-30--	34	17	24	15	42	21	9	24
14 Family medicine	10	40	20	40	20-	46	17	15	11	40	19	31	21
15 Internal medicine	14	14-	7-	79+	7	24	15	18	13	58	21	6	18
16 Pediatrics	9	33	22	44-	11	26	17	17	14	57	24	8	19
17 Ear, Nose, Throat	5	80++	0-	20-	80++	39	22	18	16	43	25	21	30
18 Neurology	5	20	20	60	0	30	23	23	20	47	27	8	33
19 Ophthalmology	4	25	25+	50	0	22	22	14	17	63	26	8	29
20 Orthopedics	3	0	33++	67-	-33--	8	21	4	11	88	29	4	17
21 Pulmonology	9	33	11	56	22	38	16	13	12	49	20	24	21
22 Radiology	3	67++	33	0-	33+	32	27	43	27	25	26	-11	48
23 Rehabilitation	1	0	0	100+	0	15	36	12	32	73	44	3	51
24 Urology	2	0-	0	100+	0-	29	33	3	12	68	35	26	36
Total clinical sciences	106	38	16	46	22+	33	10	17	8	50	16	17	9
25 Health care economics	2	50	50	0-	0	35	30	42	37	23	30	-7	60
26 Epidemiology	4	50	0-	50+	50+	48	23	18	20	34	25	30	36
27 Public health law	7	71	14	14	57	73	15	9	11	18	15	64	22
28 Medical ethics	5	80++	20	0-	60+	52	22	13	15	35	24	40	30
29 Medical psychology	10	70++	30++	0-	40	51	14	16	13	33	18	35	20
30 Medical sociology	5	40-	60++	0-	-20--	70	20	15	17	14	17	55	33
31 Psychiatry	9	56	11	33	44	61	16	11	11	29	17	50	21
Total behavioral sciences	42	62	24++	14-	38	58	10	15	7	27	13	43	12
Total	249	39	16	45	22	39	9	16	7	44	14	23	8

- / - / + / ++ / low, high compared with group

The cutoff score for this test is: 15.85%

Your CORRECT - INCORRECT score is: 22.49% (Absolute: 56)

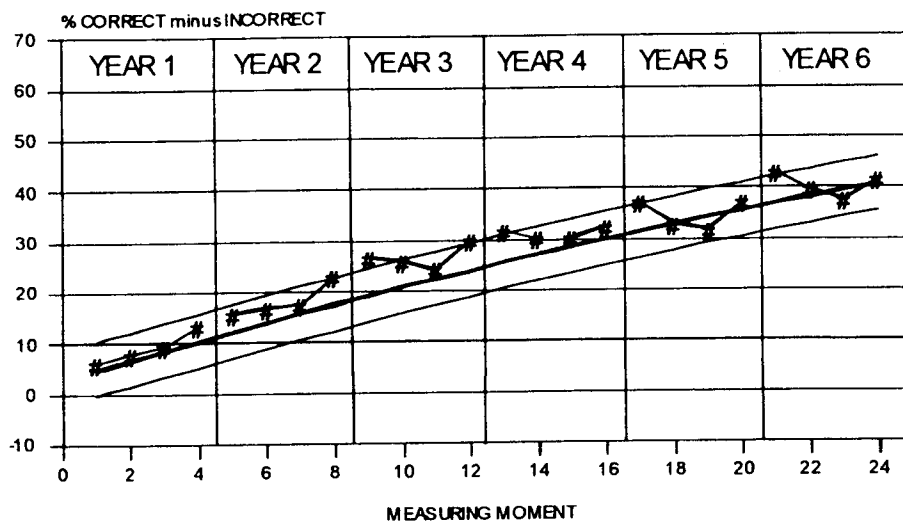
Eliminated items: 38,68,69,133,152,155, 242.

Your result: PASS

Changes of key: 12,116,215, 243.

Figure 1. Results form sent to students. The upper half shows the individual score and the class score by category, the lower half shows the scores by department.

(previous versions, results of previous tests, any changes in the key, withdrawn after the test, etc.) are reported. Feedback includes PTRC comments. The latter comprise a brief interpretation of the item analyses and possible explanations of identified problems. The item report is available to students (Figure 3).



**Figure 2.** Growth curve of results on the four annual progress tests for the period September 1985 to May 1995. The curve represents the 'line of best fit' for the mean scores of ten classes (ten times 24 data points). The dotted lines represent the 95% confidence intervals. The asterisks indicate the mean scores of the six classes on four consecutive tests.

## Appeal Procedure

After the test results have been published, any student who disagrees with his or her result can appeal to the College of Appeal for the Examinations. If a student appeals, the chairperson of the PTRC provides the Certifying Committee with written comments and advice regarding the appeal. If necessary, consultation with experts is sought. It is a statutory requirement that these comments and advice are discussed with the student to determine if an amicable settlement can be reached. If no agreement is reached, the appeal is heard at a public session of the College of Appeal for the Examinations, and expert witnesses may be called. Since 1990, appeal cases have occurred fairly regularly. Before that only one appeal case occurred, presumably because until March 1990 the possibility of appeal was not mentioned in the official test results form. Until then, the option was only included in the study guide. In seven years (1990–1996), 121 items from 19 out of the total of 28 tests have been appealed. This is less than 2% of the total number of items included in tests during these seven years. The majority of appeals were resolved by amicable settlement. On 27 of 121 items the student and the Examination Committee failed to reach agreement and the College of Appeal for the Examinations had to decide. In most cases students appeal because they stand to gain if their test result is changed. During the past seven years, 59 students have appealed, most of whom needed a slightly improved score to achieve a better qualification.

PROGRESS TEST FACULTY OF MEDICINE-MAASTRICHT UNIVERSITEIT: December 1994 TASK FORCE ON ASSESSMENT

PHYS0134/06-00195  
 105- Systolic coronary circulation rate differs from diastolic coronary circulation rate.  
 The diastolic circulation rate is higher.  
 KEY: True

ITEM ANALYSIS:

YEAR:	1	2	3	4	5	6
%-CORRECT:	11	58	58	64	71	79
%-INCORRECT:	8	21	28	20	20	15
%-?:	82	21	14	16	9	6
RIT Cml:	0.089	0.001	0.118	0.104	0.093	0.216
DI Cml:	-0.037	-0.026	0.048	0.133	0.026	0.160

STUDENT COMMENTS:  
 #93152  
 Coronary circulation rate varies considerably during the heart cycle. The left coronary artery is compressed by the high intraventricular pressure during systole (esp. the inner capillaries and veins). Circulation rate here is highest during diastole. The right coronary artery is less affected by the intraventricular pressure, because it is located lower compared with the left coronary artery, i.e. coronary circulation rate follows aortic pressure and is therefore highest during systole. In 50% of people the right coronary artery is dominant (in 20% the left artery is dominant, in 30% there is no dominance) - i.e. the right coronary artery serves both the right-hand side and part of the left-hand side. This implies that in 50% of people the coronary system consists for the greater part of the right coronary artery, where systolic circulation rate is higher than diastolic circulation rate.  
 Lit.: Bernards en Bouman, Fysiologie van de mens, 1988, blz. 361.

#XXXX8  
 I think this is a questionable item because coronary circulation cannot be viewed as one single entity. Early during diastole the circulation rate is indeed higher in the left coronary artery whereas the systolic circulation rate is very low. HOWEVER, the right coronary artery follows aortic pressure, i.e. the maximum circulation rate occurs at the end of the systole.  
 (Bernards & Bouman 1988, pg. 361)

COMMENTS OF THE PROGRESS TEST REVIEW COMMITTEE (PTRC):  
 After consultation with the contact person from the department it was decided to WITHDRAW the item.  
 Motivation: The students' comments are correct.  
 Do you think the item should be revised or should it be removed from the progress test item bank?

History:  
 Original version (used in PT-DEC87: response comparable):  
 Diastolic circulation rate is higher than systolic circulation rate [true]

Figure 3. Example of an item report.

## Discussion

It is the task of test review committees to assess, independently of the item's author(s), whether test items are valid indicators of needed knowledge. By identifying potential error sources the committee strives to minimize the risk of false-positive or false-negative decisions and to ensure that the test does measure what needs to be measured. It appears to be worthwhile to have a multidisciplinary group scrutinize test items to identify problems relating to both the form and content of items. The peer review approach adopted by the Maastricht PTRC has evolved over the past 20 years and it is continually being adapted to the prevailing circumstances, demands, and ideas. A recent development is the publication of a national document, which sets out the final objectives of undergraduate medical education in the Netherlands (Metz *et al.*, 1994). Currently, proposals are being considered to tailor test items to this document. We also plan to improve efficiency by re-computerizing the procedure.

Our experiences have taught us that the described approach to test review leads to the timely detection and solution of problems. Most items drawn from

the item bank, whether new or previously used, are not immediately suited for inclusion in the progress tests. Despite the intensive review procedure, student comments and item analyses identify an average of 5.5% of items that should be withdrawn from the test. This percentage has remained stable for 20 years. Test reliability, mean class scores, and percentage of items that are not answered or are answered incorrectly have also remained stable (Swanson, 1988; Van der Vleuten *et al.*, 1996b).

We have no proof that the approach described in this article produces the highest possible quality test items. What this approach does provide, however, is a clear, thorough and open assessment procedure, which signifies to all involved that the test is taken seriously. Given that test results have far-reaching consequences for students' progress, students are entitled to careful quality assurance of tests. Peer review is commonly used in evaluating research, but its application in education is rare. Based on our experiences in test review, we believe that peer review could also be valuable in educational quality assurance.

### Acknowledgement

The authors thank Ms. Mereke L. B. Gorsira for translating the Dutch manuscript.

### Note

1. The categories embrace the main human organ systems and some general content categories. See Figure 1.

### References

- BANDARANAYAKE, R.C. & COX, K.R. (1988). Writing multiple choice questions. In: K.R. COX & C.E. EWAN (Eds), *The medical teacher* (pp. 152–156). London: Churchill Livingstone.
- COX, K.R. (1988). How to construct a fair multiple choice question paper. In: K.R. COX & C.E. EWAN (Eds), *The medical teacher* (pp. 157–160). London: Churchill Livingstone.
- DOWNING, S.M. & HALADYNA, T.M. (1997). Test item development: validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10, 61–82.
- EBEL, R.L. & FRISBIE, D.A. (1991). *Essentials of educational measurement*. New Jersey: Englewood Cliffs.
- FREDERIKSEN, N. (1984). The real test bias. Influences of testing on teaching and learning. *American Psychologist*, 39, 193–202.
- HALADYNA, T.M. & DOWNING, S.M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2, 37–50.
- HUBBARD, J.P. (1978). *Measuring medical education*. Philadelphia: Lea & Febiger.
- KRAEMER, H.J., DUPPRÉ, H.J., BOELCKE, G., MICAELIS, J. & VOIGTMANN, K. (1976). *Institut*

- für medizinische und pharmazeutische Prüfungsfragen. Aufgaben, Entwicklung, Analysen IMPP.* Mainz: Verlag Druckhaus Schmidt & Bödige.
- METZ, J.C.M., PELS RIJCKEN-VAN ERP TAALMAN KIP, E.H. & VAN DEN BRAND-VALKENBURG, B.W.M. (1994). *Blueprint 1994: training of doctors in The Netherlands. Objectives of undergraduate medical education.* Nijmegen: University Publication Office.
- NEWBLE, D.I. & JAEGER, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education*, 17, 165–171.
- PRINCE, C.J.A.H. & VISSER, K. (1997). The student as quality controller. In: A.J.J.A. SCHERPBIER, C.P.M. VAN DER VLEUTEN, J.J. RETHANS & A.F.W. VAN DER STEEG (Eds), *Advances in medical education* (pp. 15–18). Dordrecht: Kluwer Academic Publishers.
- STOKES, J.F. (1967). Examining in the United States: the national board of medical examiners. *British Journal of Medical Education*, 1, 320–329.
- SWANSON, D.B. (1988). *Review of the assessment system used by the University of Limburg Medical School* (EPG-publ. no. 88–22). Maastricht: University of Limburg.
- VAN DER VLEUTEN, C.P.M., SCHERPBIER, A.J.J.A., WIJNEN, W.H.F.W. & SNELLEN, H.A.M. (1996a). Flexibility in learning: a case report on problem-based learning. *International Higher Education*, 17–24.
- VAN DER VLEUTEN, C.P.M., VERWIJNEN, G.M. & WIJNEN, W.H.F.W. (1996b). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher*, 18, 103–109.
- VERWIJNEN, M., IMBOS, T., SNELLEN, H., STALENHOF, B., POLLEMANS, M., VAN LUYK, S., SPROOTEN, M., VAN LEEUWEN, Y. & VAN DER VLEUTEN, C. (1982). The evaluation system at the Medical School of Maastricht. *Assessment and Evaluation in Higher Education*, 7, 225–244.