Benchmarking by cross-institutional comparison of student achievement in a progress test

Muijtjens, Arno M. M.; Schuwirth, Lambert W. T.; Schotanus, Janke; Thoben, Arnold J. N. M.; van der Vleuten, Cees P. M.; van, der

Published in:
Medical Education

DOI:
10.1111/j.1365-2923.2007.02896.x

Benchmarking by cross-institutional comparison of student achievement in a progress test

Arno M M Muijtjens,[1] Lambert W T Schuwirth,[1] Janke Cohen-Schotanus,[2] Arnold J N M Thoben[3] & Cees P M van der Vleuten[1]

OBJECTIVE To determine the effectiveness of single-point benchmarking and longitudinal benchmarking for inter-school educational evaluation.

METHODS We carried out a mixed, longitudinal, cross-sectional study using data from 24 annual measurement moments (4 tests × 6 year groups) over 4 years for 4 annual progress tests assessing the graduation-level knowledge of all students from 3 co-operating medical schools. Participants included undergraduate medical students (about 5000) from 3 medical schools. The main outcome measures involved between-school comparisons of progress test results based on different benchmarking methods.

RESULTS Variations in relative school performance across different tests and year groups indicate instability and low reliability of single-point benchmarking, which is subject to distortions as a result of school-test and year group-test interaction effects. Deviations of school means from the overall mean follow an irregular, noisy pattern obscuring systematic between-school differences. The longitudinal benchmarking method results in suppression of noise and revelation of systematic differences. The pattern of a school's cumulative deviations per year group gives a credible reflection of the relative performance of year groups.

CONCLUSIONS Even with highly comparable curricula, single-point benchmarking can result in distortion of the results of comparisons. If longitudinal data are available, the information contained in a school's cumulative deviations from the overall mean can be used. In such a case, the mean test score across schools is a useful benchmark for cross-institutional comparison.

KEYWORDS multicentre study [publication type]; *benchmarking; *educational, medical, undergraduate; *educational measurement; curriculum; programme evaluation; inter-institutional relations, schools, medical; Netherlands.

*Medical Education 2008: **42**: 82–88*
doi:10.1111/j.1365-2923.2007.02896.x

INTRODUCTION

Despite the pitfalls of inter-institutional collaboration, the expected benefits of joining forces seem to have an irresistible allure, as is shown by various recent initiatives. The International Database for Enhanced Assessments and Learning (IDEAL) consortium, set up in Hong Kong in 2001,[1] invites medical schools internationally to 'share materials and to enhance assessment of medical students'. The Universities Medical Assessment Partnership (UMAP) is a UK-based collaborative project, aimed at 'raising standards in written assessment in undergraduate medicine', which now has 12 partners after starting with 5 medical schools in 2003.[2]

The immediately obvious benefit of collaborative assessment is cost-effectiveness, for item sharing reduces test production costs without compromising quality. Quality improvement may be an added benefit as participants in a collaboration are able to devote the time and energy they have available to the production of fewer items. The quality of jointly

[1]Department of Educational Development and Research, Faculty of Medicine, Maastricht University, Maastricht, The Netherlands
[2]Institute for Medical Education, Centre for Innovation and Research of Medical Education, University Medical Centre Groningen, Groningen, The Netherlands
[3]Department of Educational and Student Affairs, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Correspondence: Arno M M Muijtjens, Department of Educational Development and Research, University of Maastricht, PO Box 616, 6200 MD Maastricht, The Netherlands. Tel: 00 31 43 388 5745; Fax: 00 31 43 388 5779; E-mail: a.muijtjens@educ.unimaas.nl

Overview

What is already known on this subject

Cooperation by medical schools on assessment lowers costs and fosters quality improvement.

What this study adds

Even with minimal between-school differences and uniform measurement moments and tests, comparisons based on a single test yield unstable (and thus unreliable) indications for systematic between-school differences.

Longitudinal test data allowing integration of information across measurement moments can reveal smooth, stable patterns which clearly reveal systematic between-school differences.

Suggestions for further research

Future research might extend the benchmarking procedure to examine all subscores of the progress test to yield more detailed information on school performance.

produced examinations may surpass that of in-house examinations.[3,4] The main advantage of cooperative assessment, however, is that it creates opportunities for comparing curricular effectiveness and identifying problem areas in schools.

The prospect of cross-institutional comparison has been a driving force for inter-faculty cooperation on progress testing in the Netherlands. This resonates with current ideas about benchmarking of curricula.[5–8] Sound benchmarking can yield information to underpin curriculum evaluation and is an essential component of efforts to raise educational standards across schools. Although benchmarking may seem straightforward at first sight, fair comparison can be thwarted by all sorts of hidden discrepancies.[9] Different admission strategies favour recruitment of different student populations, leading to confusion of student effects and curriculum effects. Varying attrition rates between schools may give rise to error similar to mortality in randomised controlled trials. Differences in student and teacher experience with benchmarking instruments can create unfair (dis)advantages. Still another potential confounder is test status, which can vary between schools from high-stakes summative tests to strictly formative tests. Finally, matters are complicated by psychometric sources of error: single test scores are generally not sufficiently reliable and group and cohort effects can limit the generalisability of findings. Obviously, benchmarking merits careful attention if we want to obtain reliable and usable information that is not distorted by insidious sources of bias.

Fortunately, the cross-institutional cooperation on progress testing by 3 Dutch medical schools presented an excellent authentic experimental set-up for benchmark testing.[3] The student population of the schools is homogeneous, because admission to all Dutch medical schools is determined by a national lottery procedure.[10] All schools adhere to the same nationally agreed statutory objectives of undergraduate medical education.[11] The cooperation on progress testing is based on these shared final objectives, joint item production, comparable test status and simultaneous assessment of all students at the 3 schools.

The current study investigates the effectiveness of single-point benchmarking and longitudinal benchmarking for inter-school educational evaluation. For this purpose, the effects of 2 benchmarking methods are examined: single-point benchmarking, involving comparison of single test results between schools, and longitudinal benchmarking, based on the cumulative deviation of test results from overall mean scores across schools. At first sight, single-point benchmarking appears simple and easy to use, especially when the schools to be compared are very homogenous. Appearances can be deceptive, however.

METHODS

Context

Since September 1999, the medical schools of the Universities of Groningen, Maastricht and Nijmegen have jointly constructed and administered progress tests for all their undergraduate medical students. Tests comprise 250 true/false questions representing a sample from the domain of relevant and functional knowledge that recent graduates are expected to have acquired by the end of the 6-year curriculum. A blueprint ensures stratification of the sample by discipline and disease or complaint categories. All of the schools' students sit 4 progress tests per year simultaneously. Four annual tests (totalling 1000 items) provide longitudinal information charting the growth of students' medical knowledge over the curriculum. Test dates and times, scoring and standards are identical and the examination regulations testify to comparable test status across schools. Test scores of approximately 5000 students are obtained per test and stored and analysed in 1 database.

We tested the single-point and cumulative deviation benchmarking methods by using them for comparative analysis of the progress test scores of the 3 cooperating schools in the academic years 2001–02 through 2004–05.

All 3 schools (in random order designated as A, B and C) offer a 6-year undergraduate medical curriculum. The schools share the same end objectives, but their curricula differ. Revision of the problem-based curriculum of school A in 2001 resulted in greater integration of the curriculum which, until then, had been divided into 4 mainly theory-oriented years and 2 clinical years. In the revised curriculum patient contacts were introduced earlier and a focus on the basic sciences continued during the clinical clerkships.

School B's curriculum is student-oriented. Years 1–4 are devoted to pre-clinical training characterised by multidisciplinary modules, small-group work, structured self-study tasks and core textbooks. Clerkships start as early as Year 3 and Years 5 and 6 are exclusively devoted to clinical training. This curriculum was stable from 1995 until 2005.

School C's curriculum is patient-oriented. In Years 1–4 a week starts with an interactive plenary presentation of a real patient. Four study tasks related to this patient are tackled in tutorial groups of 10 students per group. At the end of the week, the results are presented. Remaining questions are collected and dealt with in a wrap-up session. The learning process is supported by additional lectures and practicals. Years 5 and 6 consist of clinical clerkships. Starting in 2003 the curriculum was revised by the introduction of additional educational activities focused on the development of general skills and competencies for the medical profession.

Instrumentation

Progress test items are produced by teaching faculty at the 3 participating schools. Items are first reviewed by the schools' local test review committees and then by 1 central, cross-institutional, test review committee (chair, deputy chair and 6 members representing the pre-clinical, clinical and behavioural disciplines).

Reflecting the national final objectives, test content is curriculum-independent. As test content reflects knowledge at graduation level and all student year groups sit the same tests, junior students are not expected to know all the answers, which is why a "Don't know' option is provided. This necessitates a negative marking procedure in which incorrect answers are penalised by subtracting marks. A 'Don't know' answer yields no marks. This so-called *formula scoring* is used to correct for random guessing.[12,13] Test scores are expressed as percentage scores.

Subjects and data

We analysed the percentages of correct minus incorrect scores of all students from the 3 cooperating medical schools (1600, 1600 and 2100 students, respectively) on 16 progress tests in the academic years 2001–02 through 2004–05.

Analysis

Four tests and 6 student year groups represent 24 measurement moments. Moments 1–4 yield the results of Year 1 students in 1 academic year; measurement moments 5–8 yield those of Year 2 students, etc. Measurement moment 24 is the last progress test for students in Year 6. The results of the 3 medical schools can be compared by considering between-school differences in mean test scores (percentage of correct minus incorrect answers) for each measurement moment (Fig. 1a).

Single-point benchmarking

Single-point benchmarking involves calculating the mean score of a year group on 1 test across schools and subtracting it from individual schools' mean scores to obtain the *deviations from the overall mean score*, indicating each school's position relative to the average score of the peer institutes. Positive and negative deviation scores indicate superior and inferior performances, respectively (Fig. 1b).

Longitudinal benchmarking

The longitudinal benchmarking method is more complicated. After calculation of a school's deviation scores according to the single-point benchmarking method, average deviation scores for each measurement moment are calculated by summing the deviation scores for the preceding measurements and dividing them by the number of measurements up to that point. Thus, the *cumulative deviation score* for measurement moment 12 is the sum of the deviation

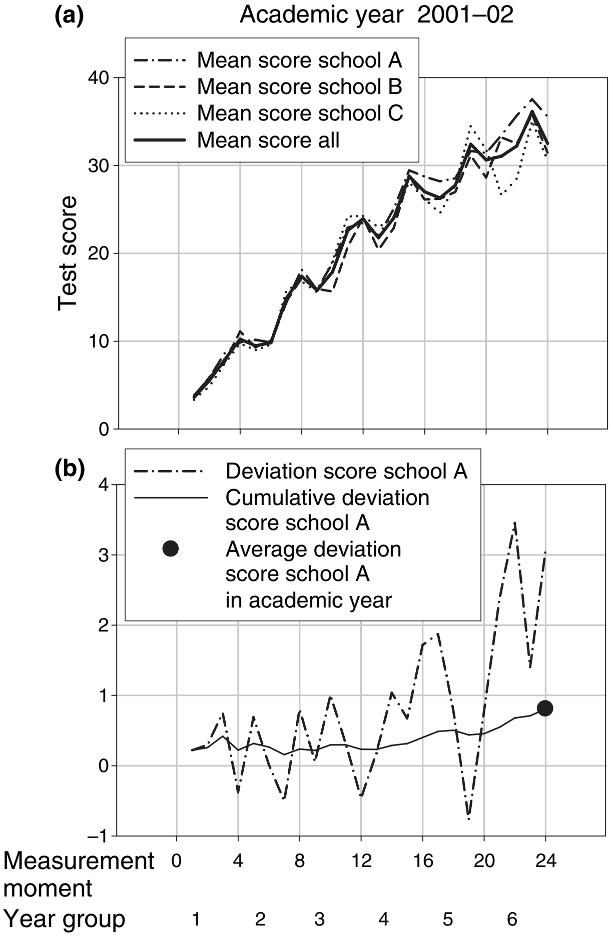

Figure 1 Scores for the 3 medical schools on the cross-institutional progress test in the academic years 2001–02 for 24 measurement moments (4 tests in Years 1–6). (a) Mean test score (percentage correct minus incorrect) per school and overall. (b) Deviation from the overall mean for school A, and corresponding cumulative deviation

scores of measurements 1–12 divided by 12. Cumulative deviation scores can be obtained for each measurement moment to indicate the average deviation up to that measurement moment. A positive cumulative deviation indicates better than average performance compared with peer schools up to that point and a negative value indicates the opposite. The cumulative deviation score for measurement 24 represents a school's average deviation across all year groups and tests taken in an academic year. The cumulative deviation score, analogous to a moving average, suppresses irregular fluctuations in single deviation scores, whilst maintaining systematic gradual variation. This is illustrated in Fig. 1(b). Whilst the deviation scores (dash-dot line) of school A follow an irregular noisy pattern, the cumulative deviation (solid line) score ignores the noisy fluctuations and shows the systematic gradual variation. The cumulative deviation score indicates a school's comparative performance in relation to the other schools in an academic year up to a certain measurement moment. The pattern of the cumulative deviation scores also supplies information about performance per student year group. A rising pattern indicates relatively good performance by a year group in that school during a certain period, whereas a falling pattern indicates the opposite.

RESULTS

Figure 2(a) shows the patterns of the mean percentage of correct minus incorrect scores for the 3

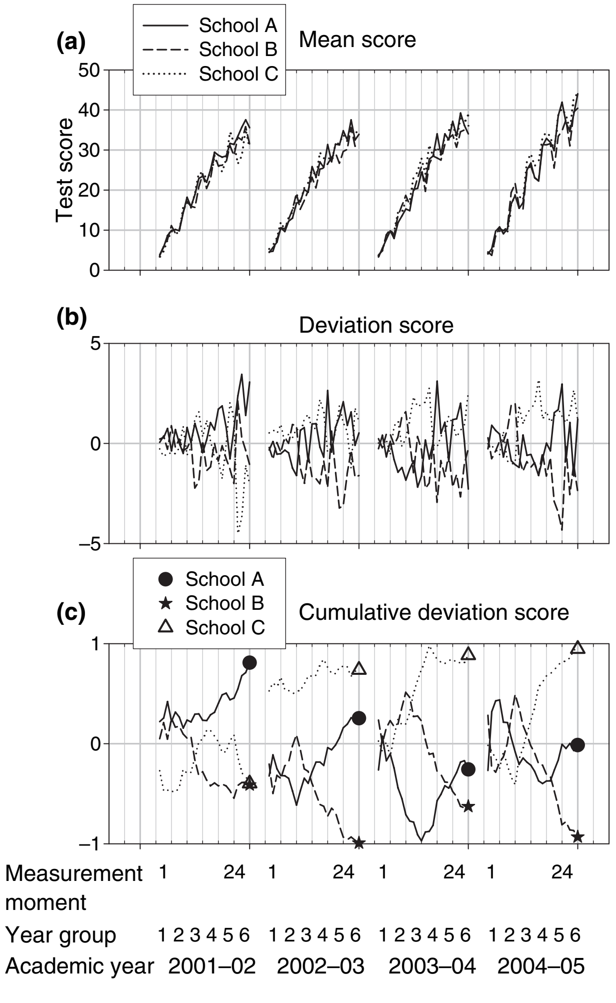

Figure 2 Scores for the 3 medical schools on the cross-institutional progress test in the academic years 2001–02 through 2004–05. Per academic year test scores are obtained for 24 measurement moments (4 tests in 6 year groups). (a) Mean test score (percentage correct minus incorrect) per school. (b) Corresponding deviation scores (school mean minus mean across schools). (c) Cumulative deviation scores: the symbols indicate a school's average deviation score across all 24 measurement moments for an academic year

medical schools for 24 measurement moments in each of 4 consecutive academic years. Generally, the mean score is about 3% at measurement moment 1, increasing to about 35% at measurement moment 24. This reflects the growth in medical knowledge during the curriculum (35% correct minus incorrect score is equivalent to 67·5% correct).

Single-point benchmarking

Figure 2(a, b) presents the results of cross-institutional comparisons with single-point benchmarking. The pattern of the overall average scores at 24 measurement moments in 1 year (Fig. 2a) reveals a combination of a trend and a (more or less) periodic pattern that recurs every 4 measurement moments. The trend represents the general increase in students' medical knowledge[14] over time. The periodic component is a combination of seasonal effects and variations in test difficulty. For instance, most of the curves dip at the first test (September), which can be explained as the seasonal effect of the summer holidays in July and August. The periodic pattern is not identical in every academic year because of random variation in test difficulty. The curves display deviations from a strictly periodic pattern, reflecting between-year group differences in test difficulty, as well as seasonal effects.

Figure 2(b) shows the deviation scores for the data for each school shown in Fig. 2(a). As the graph shows, cross-institutional comparison is hampered by pattern irregularity.

Longitudinal benchmarking

Figure 2(c) shows the results for the longitudinal benchmarking method. The cumulative deviation scores afford a clearer view of the systematic differences between the 3 schools. The symbols at the end-points of the curves (measurement moment 24) represent the cumulative deviations over 1 academic year per school. For instance, university A was top performer in 2001–02, showed a decline in 2002–04, and an upturn in performance in 2004–05.

Inspection of the curves reveals more detailed information about the relative performance of student year groups. For instance, school A's curves in Fig. 2(c) reveal low test performance by year groups 1 and 2 in 2002–03, by year groups 1–3 in 2003–04 and by year groups 2–4 in 2004–05. These results reflect consistently lower performance by the entering cohorts of 2001–04 compared with the preceding cohorts, as reflected by the increasing curves in year groups 3–6 in 2002–03, year groups 4–6 in 2003–04 and year groups 5 and 6 in 2004–05. These findings can be explained by a major curriculum reform at school A. It started in Year 1 in 2001 and includes 1 additional year every year until completion at the end of 2007.

DISCUSSION

The results of the cross-institutional comparisons with the 2 benchmarking methods highlight the potential error associated with single-point benchmarking and show that benchmarking based on cumulative deviation scores allows more detailed, accurate and informative comparisons.

Apparently, conclusions based on single-point benchmarking can vary substantially in both direction and size depending on the moment of the analysis. The fact that we observed these inconsistencies in an analysis of data from schools with strong similarities in setting, population and curriculum substantiates the doubts raised by our results with regard to the reliability of single-point benchmarking. To illustrate how precarious single-point benchmarking can be, we will reconsider the results for school A (Fig. 1b) in 2001–02. School A's relative performance varies across measurement moments (i.e. depending on test and year group). At measurement moments 4, 7, 12 and 19, it was lower than that of the other schools. However, the opposite is observed for the other measurement moments. On average, as expressed by the positive cumulative deviation (solid line) score for that year, school A outperformed the other schools. Single-point benchmarking would give rise to inaccurate conclusions when based on measurement moments 4, 7, 12 or 19. Inspection of the scores for other schools and academic years shows similar results.

The integrative use of multiple information sources thus appears to lead to a more stable, reliable and convincing conclusion. The cumulative deviation score reveals smooth, gradual variations, whilst 'suppressing' irregular, noisy variations. As students' levels of medical knowledge are unlikely to change abruptly, the gradual smooth variations resulting from the cumulative procedure can be assumed to be a better representation of differences between schools in students' knowledge levels than the unstable, abruptly varying results obtained with single-point benchmarking.

The cumulative deviation, indicating a particular medical school's relative average performance compared with that of the other schools up to a certain measurement moment appears to be readily interpretable. The integrative use of information from several measurement moments proved to supply clear and convincing evidence for effects of educational interventions (curriculum change), allowing detailed analysis, diagnosis and intervention. This is illustrated by school A, where a new curriculum was introduced in 2001. School A showed a decline in performance of the year groups in the new curriculum relative to the peer schools' performances. This prompted school A to set up a committee to review the new curriculum and identify gaps in medical content. It is unlikely that this quality control cycle would have been initiated based on data for a single measurement moment or if cross-institutional comparison had not been possible at all. The committee was set up early in the academic year 2004–05 and remedial actions did not start until the end of the academic year, so these actions did not affect the test results used in the current study.

The cumulative deviations of schools B and C showed patterns that were consistent with the curriculum changes at these schools. The curriculum change that was introduced in school C in 2003 was accompanied by a decline in performance similar to that observed in school A in 2001 (Fig. 2c). The curriculum change in both schools involved a change in direction towards a competency-oriented curriculum. The decline in progress test performance at these schools suggests that more attention for competencies may have resulted in less attention for theoretical knowledge development as assessed by the progress test. School B's curriculum was stable from 1995–96 until 2005–06. Starting in 2003–04, school B showed improved performance in Years 1 and 2, which may be indicative of a relative benefit regarding progress test scores that reflects *not* changing to a competency-oriented curriculum.

The benchmarking procedure with cumulative deviation scores could be extended to examine every subscore of the progress test (e.g. disciplines) to yield more detailed information on school performance.

Our results suggest that longitudinal benchmarking provides a useful outcome-based criterion for evaluating school or curricular performance.

A drawback of this study is that benchmarking is norm-referenced rather than criterion-referenced, because the relative performance of the cooperating schools is considered. However, assuming that on average the participating schools deliver relevant and adequate educational programmes and write relevant and adequate test items, we are inclined to accept cross-institutional average test performance as a credible and defensible benchmark.

Another limitation of the benchmarking procedure is that it is based on progress test data. Progress tests are only concerned with *knowledge* and do not comprise any of the other aspects buttressing the competence of a medical doctor.

In conclusion, the current study shows that benchmarking based on single measurements is precarious and prone to error as a result of the low reliability of single-point comparisons. Our results suggest that findings based on single-point comparisons should be interpreted with caution as they have been shown to be susceptible to incidental effects that are not representative of a school's overall performance. With longitudinal benchmarking, however, the use of cumulative deviations allows for identification of possible causes of aberrations, which may identify problem areas in schools without leading to unjustified general conclusions.

The longitudinal nature of progress testing combined with multi-institutional cooperation, as described in this paper, offers excellent opportunities for credible and defensible benchmarking. Further studies might investigate whether the results can be generalised to other situations where longitudinal data are available and cumulative deviations can be used.

In conclusion, this study unequivocally shows that single-point benchmarking is dubious at best and its results should be interpreted with caution. Benchmarking based on longitudinal data and cumulative deviations appears to be the method of preference wherever it is feasible, as it provides information that is superior in both richness and accuracy.

Contributors: AMMM conceived the study, developed the outline and method, and analysed and interpreted the data. All authors participated in data acquisition, and in drafting or revising the paper. All authors read and approved the final manuscript.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 Prideaux D, Gordon J. Can global co-operation enhance quality in medical education? Some lessons from an international assessment consortium. *Med Educ* 2002;**36** (5):404–5.

2 Byrne GJ, Owen A, Newble D, Barton R, Garden A, Roberts T, O'Neill PA. Sharing resources for UK undergraduate written assessments – 1 year of UMAP. In: Maldonado MB, ed. *11th International Ottawa Conference on Medical Education*. Barcelona: Spain. *Educacion Medica* 2004;**7**:192.

3 van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM, Thoben A, Cohen-Schotanus J, van Boven CPA. Cross-institutional collaboration in assessment: a case on progress testing. *Med Teach* 2004;**26**:719–25.

4 Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med* 2002;**77** (2):156–61.

5 Albano MG, Cavallo F, Hoogenboom R, Magni F, Majoor S, Manenti F, Schuwirth L, Stiegler I, van der Vleuten C. An international comparison of knowledge levels of medical students: the Maastricht progress test. *Med Educ* 1996;**30**:239–45.

6 Ripkey DR, Swanson DB, Case SM. School-to-school differences in step 1 performance as a function of curriculum type and use of step 1 in promotion/graduation requirements. *Acad Med* 1998;**73** (Suppl 10):16–8.

7 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Holdrinet RSG, Oeseburg B, Bulte JA, van der Vleuten CPM. An analysis of progress test results of PBL and non-PBL students. *Med Teach* 1998;**20** (4):310–6.

8 Verhoeven BH, Snellen-Balendong HAM, Hay IT *et al.* The versatility of progress testing assessed in an international context: a start for benchmarking global standardisation? *Med Teach* 2005;**27** (6):514–20.

9 Schmidt HG. Innovative and conventional curricula compared: what can be said about their effects? In: Nooman ZM, Schmidt HG, Ezzat ES, eds. *Innovation in Medical Education: an Evaluation of its Present Status*. New York: Springer Publishing Company 1990;1–7.

10 Cohen-Schotanus J, Muijtjens AMM, Reinders JJ, Agsteribbe J, van Rossum HJM, van der Vleuten CPM. The predictive validity of GPA scores in a partial lottery medical school admission system. *Med Educ* 2006;**40**:1012–19.

11 Metz JCM, Verbeek-Weel AMM, Huisjes HJ. *Blueprint 2001: Adjusted Objectives of Undergraduate Medical Education in the Netherlands*. Nijmegen, The Netherlands: Driemediagroep 2001:1–62.

12 Diamond J, Evans W. The correction for guessing. *Rev Educ Res* 1973;**43**:181–91.

13 Lord FM. Formula scoring and number-right scoring. *J Educ Meas* 1975;**12** (1):7–11.

14 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, van der Vleuten CPM. Growth of medical knowledge. *Med Educ* 2002;**36** (8):711–7.

Received 3 July 2006; editorial comments to authors 2 January 2007, 16 May 2007; accepted for publication 19 June 2007