Adv in Health Sci Educ (2008) 13:593–605
DOI 10.1007/s10459-007-9066-2

ORIGINAL PAPER

Differences in knowledge development exposed by multi-curricular progress test data

Arno M. M. Muijtjens · Lambert W. T. Schuwirth · Janke Cohen-Schotanus · Cees P. M. van der Vleuten

Received: 1 November 2006 / Accepted: 13 April 2007 / Published online: 4 May 2007

Abstract Progress testing provides data on the growth of students' knowledge over the course of the curriculum obtained from the results of all students in the curriculum on periodical similar tests pitched at end-of-curriculum level. Since 2001, three medical schools have jointly constructed and administered four progress tests annually. All students in the 6-year undergraduate curricula of these schools take the same tests resulting in 24 distinct measurements per academic year (four tests for six student year groups), which may be used to compare performance between and within schools. Because single point measurements had proven unreliable, we devised a method to use cumulative information to compare schools' test performance. This cumulative deviation method involves calculation of the deviations of schools' scores from the cross-institutional average score for 24 measurement moments in 1 year. The current study shows that it appears to be feasible to use a combination of the cumulative deviation method and trend analysis for subdomains of medical knowledge to detect strengths and weaknesses in knowledge development in medical curricula. We illustrate the method by applying it to data from 16 consecutive progress tests administered to all students (4,300) of three medical schools in the academic years 2001/2002 through 2004/2005.

Keywords Assessment · Progress testing · Curriculum diagnosis · Between-school comparison

A. M. M. Muijtjens (✉) · L. W. T. Schuwirth · C. P. M. van der Vleuten
Department of Educational Development and Research, Faculty of Medicine, Maastricht University, P.O. Box 616, Maastricht 6200 MD, The Netherlands
e-mail: a.muijtjens@educ.unimaas.nl

J. Cohen-Schotanus
Institute for Medical Education, Center for Innovation and Research of Medical Education, University Medical Center Groningen, Groningen, The Netherlands

Introduction

Progress testing enables monitoring of the growth in students' knowledge over time by comparing the results of similar tests taken periodically by the total student population of a medical school. Students at different levels of education sit the same tests, pitched at the level of knowledge expected at the end of undergraduate medical education, simultaneously. Individual results on consecutive tests can be combined to reveal individual patterns of knowledge growth (Arnold and Willoughby 1990; Blake et al. 1996; Van der Vleuten et al. 1996; Föller et al. 2004; Mahadev et al. 2004).

The concept of progress testing was developed in the 1970s by the universities of Maastricht and Missouri (Arnold and Willoughby 1990; Van der Vleuten et al. 1996). At Maastricht University (then University of Limburg) progress testing was introduced to counteract test-driven strategic learning by students. The problem-based learning curriculum of Maastricht medical school has always focused strongly on self-directed learning and students developing their own learning goals. Progress testing fits with these concepts, because, unlike mastery-orientated end-of-module examinations, progress testing rewards different individual learning pathways. In addition, the method has proven to be an effective instrument for measuring individual students' knowledge development and—due to its longitudinal character—for early identification of students whose study progress is at risk (Van Til 1998; Verhoeven et al. 2002).

Two other Dutch medical schools recognised the strengths of progress testing and have joined Maastricht medical school in a collaborative effort on the production and organisation of progress tests. Inter-university collaboration on assessment is possible, because the national statutory requirements for undergraduate medical education in the Netherlands are based on a consensus blueprint of the end objectives of undergraduate medical education which has been endorsed by the eight Dutch medical schools (Metz et al. 2001). Collaboration on progress testing started in 1999 and since 2001 the three medical schools have jointly produced four progress tests every year and organised simultaneous identical tests for all the students of their schools (Van der Vleuten et al. 2004).

This collaboration introduces a new potential benefit of progress testing by creating an opportunity to benchmark the three curricula. The student population of the three schools in this study offers an excellent opportunity for benchmarking, because on entering the undergraduate medical curriculum student cohorts are highly homogeneous, all having passed the Dutch national secondary school examination at the same level and admittance to all the Dutch medical schools being regulated by a national lottery procedure (Cohen-Schotanus et al. 2006). Although there is uniformity of end objectives of medical education in the Netherlands, the curricula of the different medical schools are not identical. One might say that the destination of undergraduate medical education is the same for all medical schools, but their itineraries differ. However, quite soon after we started to compare the results of the inter-university tests, we discovered that comparisons at one single point in time were seriously unreliable. Depending on the moment of comparison and the year groups compared, conclusions concerning the relative superiority of the curricula varied (Muijtjens et al. 2007a).

Since all Dutch universities offer 6-year undergraduate medical curricula, progress testing provides 24 measurements annually for each school, i.e. the results of four progress tests in each of six curriculum years. The results of the progress tests can be used for inter-university comparisons. Because we had found that single-point measurements varied and

were lacking in reliability, we looked for an acceptable method to detect more robust trends in comparisons between the curricula.

In this paper we present a statistical method, which we have named the cumulative deviation method and which is intended to elicit trends in longitudinal knowledge growth across the undergraduate curriculum and can be used for benchmarking. Before presenting the statistical method, we describe the progress testing procedure which is at the centre of this study in greater detail. Finally, we illustrate the cumulative deviation method by applying it to progress test data gathered during 4 years of inter-university collaboration on progress testing.

Methods

Progress testing

The Interuniversity Progress Test (IPT) analysed in this study contains 250 true-false questions (although currently multiple choice questions are used) representing a predefined blueprint based on disciplines and categories (Table 1). Every year four tests are produced which are administered in September, December, March and May to all the undergraduate medical students in the 6-year curricula of the three collaborating medical schools. The questions assess knowledge at graduate level. A don't know option is provided and formula scoring is used, i.e. the test score is defined as the percentage of correctly answered questions minus the percentage of incorrectly answered questions (%correct minus incorrect). Standards to distinguish between pass, fail and distinction are determined for each year group separately, because knowledge levels are expected to increase as students progress through the curriculum. Individual students' test results are fail, pass or distinction. At the end of the academic year, the combined marks on four progress tests, i.e. 1,000 items, determine whether or not a student can proceed to the next level.

Test production and collaboration

Each test is jointly produced by the three medical schools in accordance with a consensus blueprint. Each school has its own test review committee, comprised of a chair and five faculty members who represent basic science and clinical science disciplines. The committees review all the items produced by their own schools. The progress test is accorded comparable status in the examination regulations of the three schools.

After each test, the results are analysed and students can hand in comments on test items. The results of item analysis and students' comments can lead to removal of items and changes in answer keys. These decisions are always made jointly by the three review committees. Mean test scores are calculated for each test by year group both per school and across schools. These scores are the input for the cumulative deviation method.

Cumulative deviation method

The cumulative deviation method uses data from four tests in one academic year, representing 24 measurement moments spread over six curriculum years. Measurements 1–4 are

Table 1 Blueprint of the inter-university progress test

Disc	Clus	Categories																Total
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	
AN	BS	2	2	2		1	2	1			2	1	1			1			15
BC	BS	1	1	1			1	1	1		1	1	1	1					10
CH	CK	2	1	5			2	1	1		2	5	1				2		22
DE	CK								6										6
EP	MI														8			1	9
FA	BS	1	1	1		1	1	1			1	1		2					10
FY	BS	2	1	2		1	2	1			1	1	1						12
GY	CK					6		3						1		1	2		13
HG	CK	1	1	1	1	1	1	1	1	1	1	1	1			1	10	2	25
IN	CK	8	2	2			8	2	2		4	3	1	1		2	1	1	37
KG	CK	1	1			1	1	1			1	1	1			1	1	1	11
KN	CK	1											4						5
MC	BS		1			3		1						4					9
ME	MI				3					5					2			1	11
NE	CK			2									6				1		9
OH	CK												3			1	1		5
PA	BS	1	1	1		1	1		1		2	1	1	2					12
PS	MI				11					6			1			1			19
RE	CK	1		1			1												3
SG	MI	0								3						1		3	7
Total		21	12	18	15	15	20	13	12	15	15	15	22	11	10	9	18	9	250

The numbers in the cells are the numbers of test items for combinations of discipline and category. The numbers in the margins are the numbers of items per discipline (rows), and per category (columns). Column Clus indicates the three clusters of disciplines (Disc).

Disc, Discipline; AN, Anatomy; BC, Biochemistry; CH, Surgery; DE, Dermatology; EP, Epidemiology; FA, Pharmacology; FY, Physiology; GY, Obstetrics and gynaecology; HG, Family medicine; IN, Internal medicine; KG, Paediatrics; KN, Ear, nose, throat; MC, Clinical genetics; ME, Metamedical sciences; NE, Neurology; OH, Ophthalmology; PA, Pathology; PS, Psychology and psychiatry; RE, Rehabilitation medicine; SG, Surgery.

Cat, Category; 01, Respiratory system; 02, Blood and lymph system; 03, Musculoskeletal system; 04, Mental health care; 05, Reproductive system; 06, Cardiovascular system; 07, Hormones and metabolism; 08, Dermis and connective tissue; 09, Personal and social aspects; 10, Digestive system; 11, Kidneys and urinary system; 12, Nervous system and senses; 13, Molecular and cellular aspects; 14, Epistemology and methodology; 15, Stages of life; 16, Knowledge of skills; 17, Preventative health care.

Clus	Cluster of sciences	Size
BS	Basic science	68
MI	Behavioural/miscellaneous	46
CK	Clinical knowledge	136

the test results of the first-year students on the four tests, measurements 5–8 the test results of the second-year students, and measurement moment 24 is test 4 in year 6, i.e. the final progress test in the curriculum. Overall mean scores per school and across schools are calculated for each of the 24 measurement moments (Fig. 1, upper panel). The next step of the cumulative deviation method entails calculating the deviation scores for each school by subtracting the overall mean score from each school's mean score (Fig. 1, middle panel). Following this step, cumulative deviation scores are calculated for each measurement moment for each school. This is done by averaging the deviation scores at the measurement moments leading up to the measurement moment in question. Thus the cumulative deviation of school A at measurement moment 12 is arrived at by averaging school A's deviation scores at measurement moments 1 through 12.

The deviation scores can be graphically displayed as curves across 24 measurement moments. The lower panel of Fig. 1 shows the cumulative deviation scores for the three schools, which henceforward will be referred to as schools A, B, and C. The sum of the

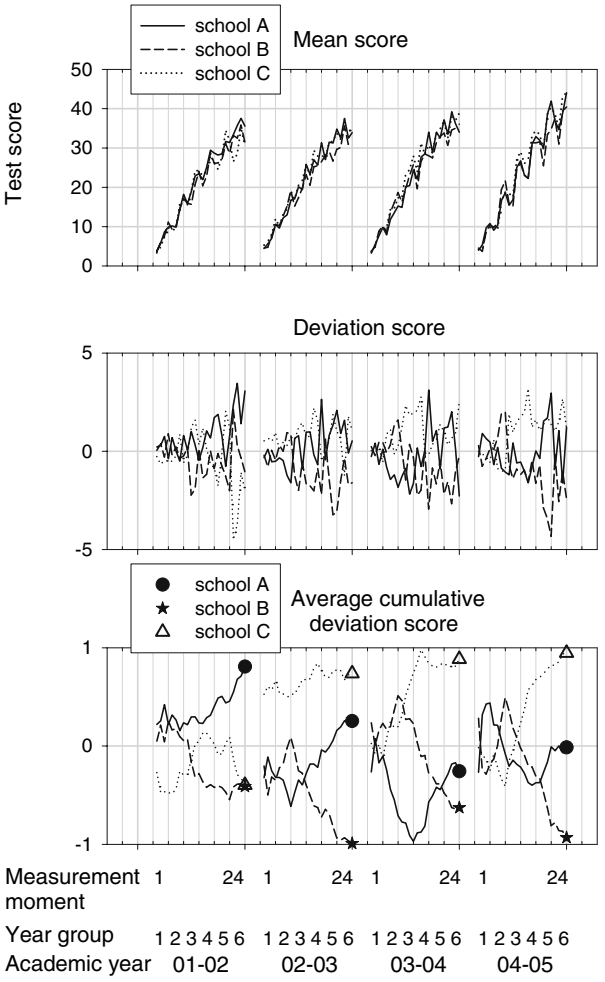

Fig. 1 Scores of the three medical schools on the inter-university progress test in the academic years 2001–2002 through 2004–2005. In each academic year test scores are obtained for 24 measurement moments (four tests in 6 year groups). Upper panel: mean test score (%correct minus incorrect) per school. Middle panel: the corresponding deviation scores (school mean minus across school mean). Lower panel: cumulative deviation scores; the symbols indicate a school's average deviation score across all 24 measurement moments in one academic year

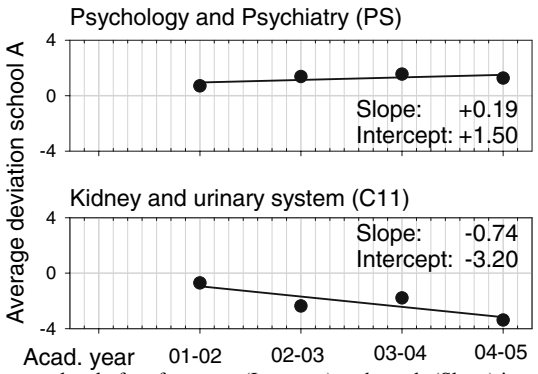

Fig. 2 Estimation of current level of performance (Intercept) and trends (Slope) in performance by fitting a straight line through the sequence of four points representing the cumulative deviations of test scores in 2001/2002–2004/2005. School A's results are shown for the subdomains of clinical psychology and psychiatry, and kidney and urinary system. The results indicate that performance on psychology and psychiatry is strong and on the rise, whereas performance for kidney and urinary system is weak with a downward tendency

deviation scores of the three schools at any one moment is zero. The end point of the curve of one school represents the average relative performance across four progress tests by the six cohorts of that school in one academic year, thus summarising the relative performance of that school on the progress test during that year. Positive and negative end points indicate relatively superior and inferior performance, respectively. Logically, downward trends reflect deteriorating performance, stable negative lines indicate stable below average performance, upward trends reflect improving performance, stable positive lines reflect consistently better than average performance, and if the curve is horizontal and close to the horizontal axis it indicates an episode of approximately average performance.

When we applied the method to the overall scores, we observed a considerable reduction in the presence of noise in the deviations compared to our earlier single-point measurements (Muijtjens et al. 2007a). Apart from ascertaining differences between overall scores, however, the method can also be used to examine differences at the more detailed level of subdomains within tests. Furthermore, both the current state and the tendency of schools' relative performance can be estimated by fitting a straight line to the end points of the curves for the different schools in four consecutive academic years. The intercept, which is defined as the height of the line at the most recent end point, (Fig. 2) serves as an immediate estimate of current relative performance. The slope indicates the size and direction of the expected development of schools' relative performance in the following academic year(s).

Analysis

We applied this method to the progress test results of the students of three collaborating Dutch medical schools in four academic years (2001/2002–2004/2005), using the data of 16 consecutive progress tests, the first in September 2001 and the 16th in June 2005. The total numbers of students from 6 year groups attending the progress tests in the four academic years we investigated ranged from 1211–1530 (School A), 1186–1389 (School B), and 1318–1900 (School C). Analyses were made for the overall scores, scores on

subtests consisting of the three clusters distinguished in the test blueprint, i.e. clinical knowledge (CK), basic science knowledge (BS), and behavioural sciences/miscellaneous knowledge (MI), and scores on subtests relating to the 20 different disciplines and 17 categories of the blueprint (Table 1).

Statistical significance of the average cumulative deviation score

We determined the significance of the differences between the schools' cumulative deviation scores using a *Z*-test to compare the amplitude of the pattern with an estimate of the standard error (SE_d). Year group sizes per school were in the order of 200–300. The SDs of the overall test scores varies from 3 to 9% for measurement moments 1 through 24. Thus the corresponding standard error of the mean (SE_m) varies from 0.19 to 0.57. For a set of three schools the standard error of the deviation score is equal to $\sqrt{4/3} \cdot SE_m$, and division by the square root of the number of measurement moments results in SE_d varying between 0.22 and 0.13 for measurement moments 1 through 24. As a consequence, for measurement moments 1 and 24 average cumulative deviation values of >0.44 and >0.26, respectively are significant at a level of 5%. For the other measurement moments the boundaries for statistical significance take intermediate values.

Results

Figure 3 shows the patterns of the average cumulative deviation scores of progress test results for schools A, B, and C in the academic years studied. The results are shown per school for the full tests (panel 1), for the basic science, clinical science and behavioural science/miscellaneous clusters (panels 2–4), for a discipline-related subtest, i.e. items on psychology and psychiatry (panel 5), and for a category-related subtest, i.e. items on the kidney and urinary system (panel 6).

The upper panel of Fig. 3 shows the patterns for the overall test scores. Application of the significance boundaries (gradually decreasing from 0.4 to 0.24 between measurement moments 1 and 24) shows that large parts of the cumulative deviation pattern are statistically significant, with patterns varying within a range of −1% to +1%. For comparison, average progress test scores range from 3% at measurement moment 1 to 35% at measurement moment 24, with SDs gradually increasing from 3% at moment 1 to 9% at moment 24. So, in terms of Cohen's effect size the between school effect is of the order of ES = |mean difference|/SD = $2/9 \approx 0.2$, which according to Cohen's classification is to be considered a small effect (Hojat and Xu 2004).

The pattern of the end points for school A (dots in the upper panel) indicates a downward trend in performance over the four academic years, albeit that there is a slight upturn in the fourth year.

Further inspection of the patterns across measurement moments reveals remarkable developments in school A's performance, with negative slopes of the patterns reflecting low test performance by year groups 1 and 2, 1–3, and 2–4 in the academic years 2002/2003, 2003/2004, and 2004/2005, respectively. These results reflect consistently lower performance by the entering cohorts in 2001–2004 compared to the performance of preceding cohorts as reflected by the increasing curves for year groups 3–6 in 2002/2003, year groups 4–6 in 2003/2004, and year groups 5 and 6 in 2004/2005. It is interesting to note

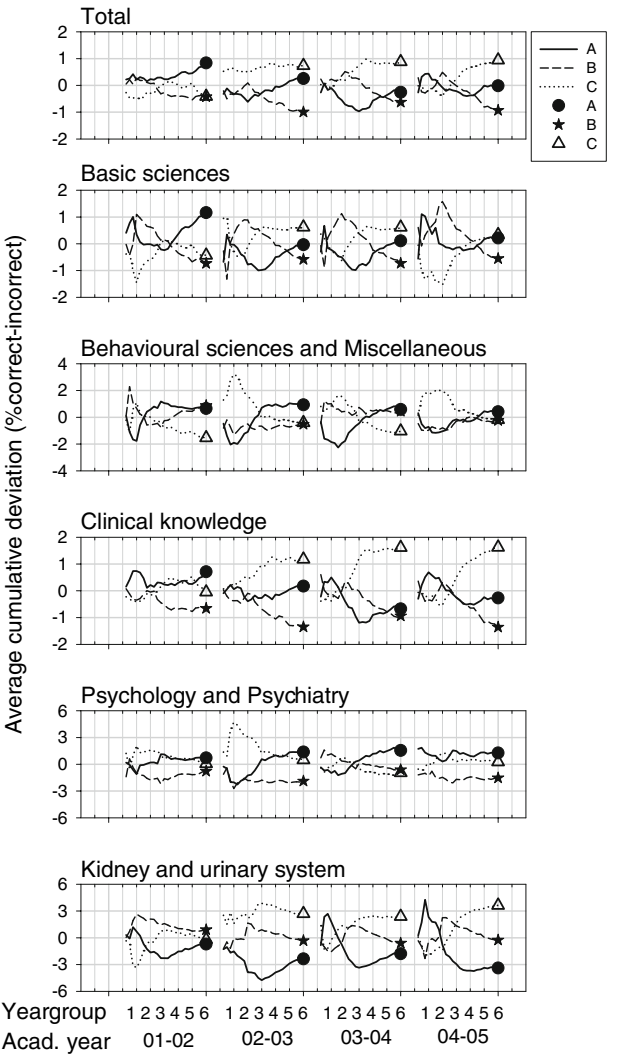

Fig. 3 Average cumulative deviation scores on the inter-university progress test for the participating medical schools A, B, and C in the academic years 2001/2002–2004/2005. Panel 1 shows the deviation scores for the overall test score, panels 2–4 show the scores for the subtests on the clusters of basic sciences, behavioural sciences/miscellaneous, and clinical knowledge. Panels 5 and 6 show the deviation scores for a discipline-based subtest on clinical psychology and psychiatry and a category-based subtest on kidney and urinary system. The three curves represent the relative performance of the three schools over 24 measurement moments (results of 6 year groups on four tests) per academic year. By definition the three curves add up to zero for each measurement moment. The end points represent the average relative performance of the schools per academic year

that the decrease in school A's performance on tests of medical knowledge started in 2001, for 2001 is the year the school launched its renewed curriculum. The curriculum was first implemented in year 1, followed by implementation 1 year at a time in the successive curriculum years until all years of the new curriculum will be in place in the academic year

2007/2008. The revision of the problem-based curriculum of School A in 2001 resulted in a more gradual vertical integration of the curriculum which until then had been divided into four mainly theory-oriented years followed by two clinical years. In the revised curriculum patient contacts were introduced earlier while attention for basic sciences continued during the clinical clerkships which started in year 4. The results of the current study suggest that these curricular changes were accompanied by a decreasing performance on knowledge development for school A compared to the other two schools. These indications caused school A to seek remedial measures based on more detailed information on those subdomains for which the school appeared to lag behind the other schools.

This information is provided in the other panels of Fig. 3. It enables us to examine whether school A was lagging behind in some subdomains but not in others. Panels 2–4 show the patterns for the BS, MI, and CK clusters. The end points of the curves for school A (dots) show a dramatic drop in performance on CK with negative deviation values for the last two academic years. BS shows a similar, albeit less dramatic, decline followed by signs of recovery in the last 2 years. MI appears to be unproblematic with hardly any variation across the academic years. Furthermore, the graphs reveal that for MI and BS the differences in performance between the three schools decrease across consecutive academic years, whereas for CK they increase.

Inspection of school A's curves for the three clusters (Fig. 3, solid line in panels 2–4) reveals that the effect of the new curriculum that was apparent from the overall test scores is not reproduced in each domain. The effect is most pronounced for CK, suggesting that the most powerful impact of curriculum change has been on clinical knowledge. For BS the pattern is different, showing low performance in year groups 2 and 3, and high performance in year groups 4–6 in each academic year. The curves for MI show low performance in year 1 and in the first semester of year 2, followed by an upswing in the second semester of year 2, which persists through year 3 and is followed by steady good performance in years 4–6. So, for MI, school A did well throughout the years studied.

To illustrate the interpretation of an average cumulative deviation curve and the kind of information that can be extracted from it, we will focus on school C's curve for MI (Behavioural Sciences and Miscellaneous) in the academic year 2004–2005 (Fig. 3, third panel). The increase in year 1 and the persisting positive level in year 2 indicate that school C students (compared to their peers in schools A and B) gain more MI knowledge in these years. However, year 3 students of school C appear to do relatively poorly on MI as is indicated by the steep decrease of the curve for the students in year 3 (deviations are: 2.3, −1.8, −5.8, −4.0). However, it is entirely possible that this poor performance reflects that the two other schools are catching up on school C by giving more attention to MI subjects in year 3. School C shows a consolidation of the scores in year 4 and a slight decrease in years 5 and 6, ending in an overall slightly negative result across the six cohorts in the academic year 2004–2005.

We also plotted and analysed the deviation patterns for the 17 categories and the 20 disciplines included in the test blueprint. Two examples of the resulting graphs are shown in panels 5 and 6 of Fig. 3. In order to capture the essence of the findings we have fitted a straight line through the four consecutive end points of one school and presented the intercept and the slope of the line in a bivariate plot. Figure 2 shows the results of this procedure for school A's results on the subtests for clinical psychology and psychiatry, and kidney and urinary system. For clinical psychology and psychiatry both intercept and slope were positive, indicating current good performance which is expected to persist in the next

academic year, whereas for kidney and urinary system both parameters are negative indicating low performance now and a further decreasing tendency in the near future.

Figure 4 visualises the lines fitted through the four end points of school A's deviation scores for the overall test (Tot) and for the BS, MI, and CK clusters in years 2001/2002–2004/2005. The horizontal axis represents the intercept, indicating the school's average deviation (i.e. relative performance) at the end of 2004/2005. The vertical axis indicates the slope, reflecting the trend or change in relative performance per year. In this way it becomes easy to compare the results of various analyses. The better the straight line fits to the four points, the more reliable the indications of the performance and its future development are.

Since a negative intercept value indicates that the current deviation is negative and a negative slope indicates an expected decrease, all points below the horizontal axis and to the left of the diagonal dashed line ('the critical region') correspond to 'problematic' subdomains. This means that for these subdomains the current deviation estimate is either

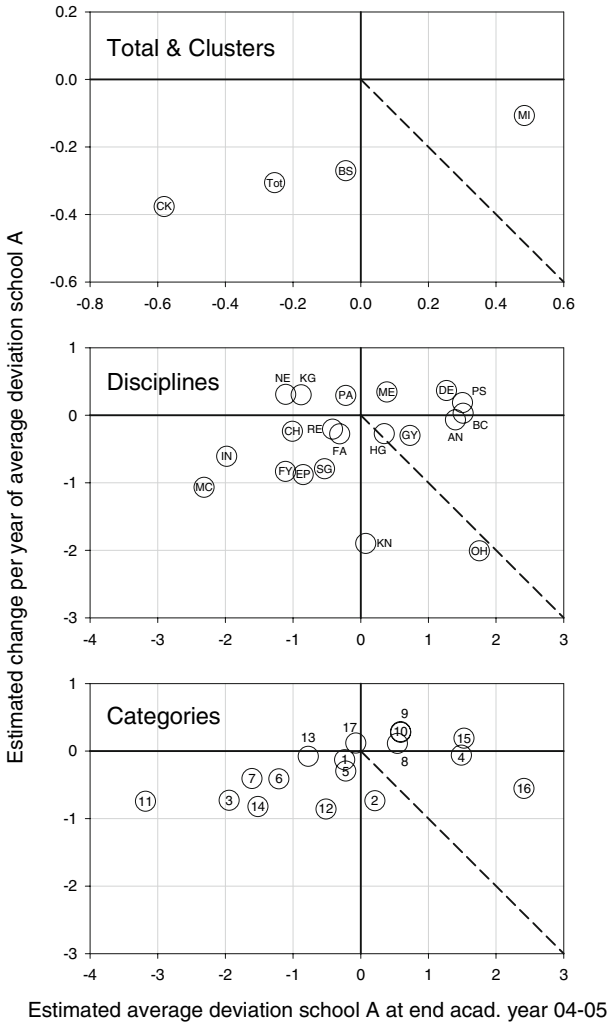

Fig. 4 Bivariate plots of intercept and slope estimated for the sequence of four points representing school A's average cumulative deviations for 2001/2002–2004/2005. The intercept (horizontal axis) indicates the school's average deviation (i.e. relative performance) at the end of 2004/2005, and the slope (vertical axis) indicates the trend (i.e. the change in relative performance per year). The upper panel shows the results for the total test (Tot), basic sciences (BS), behavioural sciences and miscellaneous (MI), and clinical knowledge (CK). The middle panel shows the results by discipline and in the lower panel the results are presented by category

already negative or expected to become so in the next year. As for size and severity of the problems, the lower and the more to the left in the lower left-hand quadrant points are located, the larger and more persistent problems are likely to be. Located in the upper left-hand quadrant, we find subdomains with low relative performance but positive slopes. In the upper right-hand quadrant subdomains have high relative performance with rising trends. Finally, the portion of the lower right-hand quadrant above the dashed line contains subdomains with currently relatively high performance but negative trends. So, although these subdomains' performance is at risk, low performance is not expected to occur in the immediate future.

The upper panel reveals that school A's overall performance is generally problematic (Tot), but the major problem is with CK, with a hint of future problems regarding BS. For MI, indications of current or expected problems are absent. The middle panel shows a subset of disciplines that are outside the 'critical region' (DE, PS, BC, AN, ME), a subset at risk of becoming problematic (GY, HG, OH), a subset showing problems as well as improvement (PA, KG, NE), and three subsets with increasing levels of problems, i.e. FA, RE, CH, SG, EP, FY, followed by KN, followed by IN and MC. In the lower panel we present the outcomes for category-based subsets, which reveal a problem free subset (15, 4, 9, 10, 8, 17, 16) and again three subsets with problems of increasing order of severity, i.e. categories 1, 5, 2, 13, 12, categories 6, 7, 14, 3, and category 11.

Discussion

Progress test results are known to be a potentially rich source of information for students, teachers, and administrators (Arnold and Willoughby 1990; Blake et al. 1996; Van der Vleuten et al. 1996; Föller et al. 2004; Mahadev et al. 2004). The focus of the current study is on the use of this information to investigate the effects of different curricula on the acquisition of medical knowledge. Exposing the strengths and weaknesses of a curriculum can make an important contribution to the quality cycle of curricular evaluation and improvement.

In this paper we introduced and illustrated a method for using cumulative information from progress tests to identify differences in performance between and within undergraduate medical curricula. Some clear problem areas were detected which may be attributable to curricular changes or flaws.

Obviously, test results can also be affected by factors other than curriculum characteristics. The most obvious one would be variation between student cohorts. However, in this study this effect is expected to be small, because the population of medical schools in the Netherlands is quite homogeneous due to admission by national lottery procedure among students who have successfully passed their final examinations at the required level of secondary education, which is highly comparable across the board (Cohen-Schotanus et al. 2006). It is important to note that for the average cumulative deviation to be a valid indicator of the relative performance of a school, it is essential that there should be no or only a negligible systematic difference in performance between the cohorts entering the schools that are compared. This can be illustrated as follows. Suppose cohorts entering one of the schools to be compared would systematically be of below average quality, this would result in consistently negative deviations at early measurement moments. In that case, even if the school would do a good job and its sixth year students would perform as well as those of the other schools, the negative deviations in the early years would add up

to a negative average cumulated deviation to summarise the school's performance in one academic year. Fortunately, the curves in the middle panel of Fig. 1 show that for progress test data the situation is quite the opposite with small deviations at the start and larger deviations at the end of the curves. For these data the cumulative deviation appears to be a valid indicator of a school's relative performance across the six cohorts in one academic year.

Another potential source of variation in schools' performance is test composition. This too is unlikely to have had an impact in this study, because the test blueprint remained unchanged throughout the study period.

So, for the years we studied, the most likely major systematic difference between the schools is the curriculum.

An alternative explanation of differences might be changes in test production. Before the start of the inter-university collaboration in 1999, the full test was constructed by Maastricht faculty. Since September 1999 faculty members of the Nijmegen and Groningen schools have gradually increased their contributions and an earlier study indicated that students scored better on items written by faculty of their own schools, despite the fact that all test items were tailored to shared end objectives of medical education (Muijtjens et al. 2007b). However, during the period of the study the percentages (and standard deviations over the years) of items produced by the Maastricht, Nijmegen and Groningen schools remained unchanged at 62% (7), 25% (3), and 13% (4), respectively. Therefore, this source too is not expected to have contributed substantially to the variation between the deviations. Moreover, different year groups in school A sat the same tests and only those in the new curriculum showed low performance. The pattern of school A's curve for the overall score appears to point clearly to a curriculum effect. The patterns of the cluster-related subdomains offer more specific information, showing that the decrease in knowledge growth is most pronounced for the clinical sciences. When we look at the still more detailed level of disciplines and categories, the patterns of the curves start to show associations with where and when subjects are scheduled in the curriculum. Furthermore, as the number of items in subtests diminishes, the resulting deviation patterns become more sensitive to item and item-school interaction effects.

Since the comparison is relative by necessity, it is quite conceivable that increasingly poor performance of one school reflects improved performance of other schools. This cannot be ruled out in this study, although it seems quite unlikely. Participation of more medical schools in the Inter-university Progress Test will help to improve the credibility and validity of the across schools average as the reference (benchmark) for comparisons between curricula. Fortunately, more medical schools already have joined the collaboration (University of Leiden, September 2006) or will do so in the near future (VU Amsterdam, September 2007).

The cumulative deviation method, which we propose in this paper, has so far been applied to knowledge tests only, because that is what progress testing measures. But we believe that the method is versatile enough to be applied to any longitudinal numerical measure of medical competence.

In summary, the results of our study appear to support the feasibility of using the method of average cumulative deviation to compare schools' performance on (subdomains of) medical knowledge, reveal the impact of curricular changes on knowledge acquisition, and diagnose strengths and weaknesses of current or developing curricula as regards the growth of medical knowledge.

Acknowledgements We thank Ron Hoogenboom and Guus Smeets for their contributions to the collection and analysis of data and Mereke Gorsira for editing the final version of the paper.

References

Arnold, L., & Willoughby, T. L. (1990). The quarterly profile examination. *Academic Medicine, 65*(8), 515–516.

Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., Cunnington, J., & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine, 71*(9), 1002–1007.

Cohen-Schotanus, J., Muijtjens, A. M. M., Reinders, J. J., Agsteribbe, J., Van Rossum, H. J. M., & Van der Vleuten, C. P. M. (2006). The predictive validity of GPA scores in a partial lottery medical school admission system. *Medical Education, 40*, 1012–1019.

Föller, T., Brauns, K., Fuhrmann, S., Hanfler, S., Hoffmann, J., Kölbel, S., et al. (2004). Five years of progress testing at Charité Universitätsmedizin Berlin, Germany. In M. B. Maldonado (Ed.), *11-th international Ottawa conference on medical education* (p. 192). Barcelona.

Hojat, M., & Xu, G. (2004). A visitor's guide to effect sizes. *Advances in Health Scienes Education, 9*, 241–249.

Mahadev, G. K., O'Neill, P. A., Owen, A. C., McCardle, P., Benbow, E., & Byrne, G. J. (2004). Seven years experience of progress testing in Manchester UK. In M. B. Maldonado (Ed.), *11-th international Ottawa conference on medical education* (pp. 192–193). Barcelona.

Metz, J. C. M., Verbeek-Weel, A. M. M., & Huisjes, H. J. (2001). *Blueprint 2001; adjusted objectives of undergraduate medical education in The Netherlands.* Nijmegen (The Netherlands), Driemediagroep.

Muijtjens, A. M. M., Schuwirth, L. W. T., Cohen-Schotanus, J., Thoben, A. J. N. M., & Van der Vleuten, C. P. M. (2007a). Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical Education* (under review).

Muijtjens, A. M. M., Schuwirth, L. W. T., Cohen-Schotanus, J., & Van der Vleuten, C. P. M. (2007b). Origin bias of test items compromises the validity and fairness of curriculum comparisons. *Medical Education* (under review).

Van der Vleuten, C. P. M., Schuwirth, L. W. T., Muijtjens, A. M. M., Thoben, A., Cohen-Schotanus, J., & Van Boven, C. P. A. (2004). Cross institutional collaboration in assessment: A case on progress testing. *Medical Teacher, 26*, 719–725.

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*(2), 103–109.

Van Til, C. T. (1998). *Voortgang in voortgangstoetsing (progress in progress testing).* Ph.D thesis, University of Maastricht, Maastricht.

Verhoeven, B. H., Verwijnen, G. M., Scherpbier, A. J. J. A., & Van der Vleuten, C. P. M. (2002). Growth of medical knowledge. *Medical Education, 36*(8), 711–717.