cohorts of the curriculum years between 1997 and 2001 (n=1947) was analysed. Each student took five progress tests over the last three years of the MB ChB programme. Each test consisted of 250 True/False questions representative of the 4 core modules within the problem-based curriculum to achieve face validity. For each student cohort, mean and standard deviations were calculated and scores compared (students t-test). The reliability of each test (Cronbach) was calculated and estimates of the average 'test-by-test' change were derived (STATA version 6.0).

Results: Each student cohort improved on their previous average (p=0.004). Moreover, during each round of testing, senior students scored higher than less senior students ({year 4 > year 3, p = 0.045}, {year 5 > year 3, p =0.002}, {year 5 > year 4, p=0.06}). Average reliability was 0.84 (range 0.79-0.89). Average test-by-test changes for each cohort demonstrated significant linear trends (P<0.001 in all cases)

Conclusion: Despite modification from the original Maastricht model, the progress test in Manchester is a valid, reliable and reproducible method for examining the acquisition of knowledge across the clinical undergraduate curriculum.

Cohort	Year 3		Year 4		Year 5
	Jan	Jun	Jan	Jun	Jun
97-99(n=288)	34.1	44.3	46.2	42.6	60.6
98-00(n=313)	31.6	32.5	50.2	58.5	51.9
99-01(n=314)	30.7	45.5	36.2	45.3	55.7
00-02(n=272)	21.2	33.3	42.9	52.7	57.7

The Reliability of the Operative Competence Assessment for Surgical Trainees Using Video and Direct Observation

Keywords: *In-service training, clinical competence, educational measurement*

Authors: *Burt CG, Ricketts C, Grant JR, Wilkins DC.*

Institution: *Peninsula Medical School*

Summary: Objectives: Case variation[1] and subjective judgments affect reliability when assessing surgical trainees. The aim was to investigate the reliability of the Operative Competence Assessment using video and direct observation.

Methods: A validated global rating scale called the Operative Competence Assessment was used to score surgeons performing inguinal hernia repairs. Surgical assessors blinded to the operator's seniority scored the edited operative videos only if they considered that they had obtained sufficient information. Other surgical trainees were scored by direct observation in theatre, where external surgeons participated to reduce rater bias. The results were analysed using Generalisability Theory.

Results: Seventy assessors (75%) felt able to judge the videos and scored all 6 operations. The assessors' contribution to variance was 0.6%, indicating a negligible hawk or dove effect. 78.7% of variance was due to the assessor-surgeon interaction, exceeding the 20.7% variance between the operating surgeons. The assessors' comments suggest that their judgments were based on a comparison with their own technique. Thirteen assessors scored 5 surgeons by direct observation in theatre. The inter-surgeon variation was 91.0% and the case variation was 9.0%. The Generalisability coefficients for the theatre assessment sessions ranged from 0.90 to 0.99.

Conclusions: Assessment of surgical skill by video alone is insufficient. Direct observation provides additional information necessary for accurate assessment. The Operative Competence Assessment was a highly reliable model when used in the operating theatre.

1. Norman GR, Tugwell P, Feightner JW et al. Knowledge and clinical problem-solving. Med Educ 1985; 19: 344-56.

Detecting cheating in medical examinations

Keywords: *Cheating; examinations; postgraduate*

Authors: *McManus, Chris*

Institution: *University College London*

Summary: In any high-stakes examination, some candidates will be tempted to pass by cheating, which can take many forms (Cizek 1999). Many undergraduate and postgraduate medical examinations are multiple-choice, and are taken by candidates sitting at desks in large examination halls. Computer-marked answer sheets make it surprisingly easy for candidates to see the answers of candidates seated alongside or in front, so there is a risk that a candidate's answers are not entirely their own. Any form of cheating threatens an examination's validity, since the mark scored does not relate to the candidate's true knowledge. It is also a threat to health care, since candidates who cheat are certified as competent, despite having insufficient knowledge. Statistical methods for detection of cheating look at the statistical pattern of results, assessing whether candidates are unduly similar to one another, particularly if they are adjacent in their seating. Although this can be done using known theoretical distributions of statistical distributions (e.g. Frary, Tideman, & Watts 1997), methods using empirical distributions make fewer assumptions. I will describe the application of a program called Acinonyx to a medical examination. The program implements and extends the methods of Angoff (1974), which looks at the answers of all possible pairs of candidates. That number can be large, as if 1000 candidates take an exam there are 499,500 pairs, and hence the computational load is large, and the high risk of type I errors needs to be taken into account.

Changes in standard of candidates taking the

Keywords: *Postgraduate examinations; standards; marker questions; MRCP(UK)*

Authors: *Chris McManus, Jennifer Mollon, Oliver Duke, Allister Vale*

Institution: *MRCP(UK) Central Office, St Andrews Place, London NW1 4LE, UK*

Summary: Maintenance of standards is a problem for postgraduate medical examinations, particularly if norm-referencing is the only method of standard setting. The MRCP(UK) Part 1 Examination includes marker questions in each diet, which are unchanged from questions used in a previous diet. Here we describe two complementary studies of marker questions in diets of the Examination over the years 1985 to 2002, to assess whether standards have changed. Study 1analysed routinely collected information on the performance of 4405 marker items, using a statistical model to assess changes in performance across diets. Study 2 compared performance of individual candidates on 28 individual marker items which were shared by the 1996/2 and 2001/3 diets. Study 1 found evidence that candidate performance on the MRCP(UK) Part 1 Examination improved gradually between 1985 and 1997, and then declined sharply until 2001. The 'dog-leg' at 1997/3 did not result from changes in Examination Regulations or candidate mix. Study 2 confirmed that 2001/3 performance was significantly worse than 1996/3 performance in graduates of UK medical schools, and that passing candidates in 2001/3 performed less well than passing candidates in 1996/2. Setting the pass mark by norm-referencing allowed candidates to pass the Examination who had performed less well than previous cohorts. As a result, the current MRCP(UK) Part 1 and Part 2 Examinations use criterion-referencing. The reasons for the declining performance of UK medical school graduates are not clear, but have wider implications for medical education. Further studies are needed of other postgraduate and undergraduate examinations.