Composite undergraduate clinical examinations: how should the components be combined to maximize reliability?

Val Wass,[1] *David McGibbon*[2] *& Cees Van der Vleuten*[3]

Background Clinical examinations increasingly consist of composite tests to assess all aspects of the curriculum recommended by the General Medical Council.

Setting A final undergraduate medical school examination for 214 students.

Aim To estimate the overall reliability of a composite examination, the correlations between the tests, and the effect of differences in test length, number of items and weighting of the results on the reliability.

Method The examination consisted of four written and two clinical tests: multiple-choice questions (MCQ) test, extended matching questions (EMQ), short-answer questions (SAQ), essays, an objective structured clinical examination (OSCE) and history-taking long cases. Multivariate generalizability theory was used to estimate the composite reliability of the examination and the effects of item weighting and test length.

Results The composite reliability of the examination was 0·77, if all tests contributed equally. Correlations between examination components varied, suggesting that different theoretically interpretable parameters of competence were being tested. Weighting tests according to items per test or total test time gave improved reliabilities of 0·93 and 0·81, respectively. Double weighting of the clinical component marginally affected the reliability (0·76).

Conclusion This composite final examination achieved an overall reliability sufficient for high-stakes decisions on student clinical competence. However, examination structure must be carefully planned and results combined with caution. Weighting according to number of items or test length significantly affected reliability. The components testing different aspects of knowledge and clinical skills must be carefully balanced to ensure both content validity and parity between items and test length.

Keywords Education, medical, methods; education, medical, undergraduate, *standards; educational measurement; reliability of results.

Medical Education 2001;35:326–330

Introduction

In response to recommendations from the General Medical Council,[1] most United Kingdom medical schools are broadening their educational objectives. More emphasis is being placed on skills training, communication and attitudinal development. This raises important issues in planning assessment procedures. The valid assessment of students' knowledge, skills and attitudes, the core elements of most curricula, requires different forms of test.[2] A multiple-choice question (MCQ) paper is a good test of student knowledge, whereas objective structured clinical examinations (OSCEs)[3] are increasingly used to examine practical skills. An expanding range of formats is now available to test applied knowledge and problem solving,[2] although the assessment of student attitude remains a challenge.

To reflect these curriculum changes, many medical schools are developing a battery of tests. It is essential for the assessment to be valid; the examination must truly test the learning it sets out to test. However reliability, i.e. the consistency of candidate performance on each test, is equally crucial. Other factors are also important. The feasibility of running and resourcing the examination cannot be ignored. Thus when setting these examinations, tensions exist between selection of

[1]Department of General Practice and Primary Care, Guy's, King's and St Thomas' School of Medicine, London, UK
[2]Guy's, King's and St Thomas' School of Medicine, London, UK
[3]Department of Educational Development and Research, University of Maastricht, Maastricht, Netherlands

Correspondence: V Wass, Department of General Practice and Primary Care, Guy's, King's and St Thomas' School of Medicine, Weston Education Centre, 10 Cutcombe Road. London SE5 9RJ, UK

 © Blackwell Science Ltd

Key learning points

Estimating the reliability of medical examinations is complicated as a battery of tests is often used to assess the requisite knowledge, skills and attitudes.

Using multivariate generalizability theory, variances in test length and composition can be accounted for and an index of overall reliability obtained.

To achieve acceptable overall reliability, careful structuring of papers to balance the length and format of individual tests is crucial.

the test format and the practicalities of delivering it, for example a 3-hour MCQ test requires considerably less resourcing than a 3-hour OSCE.

A key problem is achieving an acceptable balance between reliability and validity. If the examination is an important end-of-year or course assessment, i.e. a high-stakes one for the student, a reliability of greater than 0·8 is essential to ensure a fair pass/fail decision. Herein lies the problem. The reliability of different examination formats varies. A 3-hour MCQ test includes a large number of items and reliability should be high (above 0·8). For a 3-hour OSCE, this level of reliability is difficult to achieve[4] and essay papers, unless carefully scored, are unreliable.[5] Combining the results from the different tests, rather than assessing them individually, may perhaps achieve a better reliability but this can be difficult to do because the formats of the individual tests may be very different. There is little information on composite undergraduate examinations, and on how to construct them to minimize cost and maximize validity and reliability.[6]

When using these high-stakes composite tests, how should the overall composite reliability be estimated? Given the limited amount of overall time and the variety of available formats, how should the papers be constructed and combined to achieve maximum reliability? Questions arise relating to the optimal number of items to include in written papers or the appropriate length for clinical tests. Answers will depend on the contribution of these components to the overall reliability. When using a battery of tests, what weight should be given to the different components? For example, it may be felt that the clinical skills component should have more weight than a basic knowledge test. What effect does weighing components equally or differentially have on the composite reliability? The aim of this study is to address these questions.

The final qualifying examination for medical students on the Guy's and St Thomas' campus of a London medical school, recently merged with King's College, is a composite one, aiming for validity with regard to as many facets of the undergraduate curriculum as possible. We have analysed the composite reliability of the June 1998 examination, using multivariate generalizability theory, and investigated the effect of different weightings of the results on the overall reliability of this high-stakes examination.

Methods

The study was carried out on the June 1998 Final MBBS examination for undergraduates completing clinical training. Since 1996, the Guy's and St Thomas' campus had taken over responsibility from the University of London for its own final examination. The examination was aimed at confirming students' clinical competence before they started pre-registration house officer (PRHO) appointments.

Examination structure

The school had designed a specific composite test format to assess knowledge, skills and attitudes as given in the core curriculum. This consisted of four different written test formats and two clinical tests.

Written tests

A multiple-choice paper (MCQ), lasting 180 minutes, consisted of the following.

1 *True/false questions*. A total of 90 question stems from a pre-tested university bank, each with five associated true/false items, were designed to test basic factual knowledge in medicine, surgery, general practice, psychiatry and public health. Each correct answer scored one. A mark was subtracted for an incorrect response. Candidates were allocated 160 minutes for these questions.

2 *Extended matching questions (EMQ)*. Six additional extended matching questions (25 single items) were used to assess problem-solving skills.[7] Candidates were allocated 20 min for the extended matching questions.

Students also took the following tests.

3 *A short-answer question paper (SAQ)* (3 hours) with 10 questions, designed to assess problem solving and data interpretation skills. Two questions used public health data and eight used clinical scenarios. Each question was first marked independently out of 20 by two examiners, who agreed a final score.

4 *An essay paper* (2·5 hours) of three questions, designed to assess both the ability to present written

debate and to communicate with professional colleagues. Candidates answered one compulsory question on writing a discharge letter and had two essay choices, from 10 broad philosophical topics and from 10 more knowledge-based titles. Each essay was marked independently by two examiners using a closed fixed percentage range (65/60/55/50/48/45/40/35), where 65% was excellent, 48% borderline and 45% or below was a fail.

Clinical tests

The two clinical tests were as follows.

1 *An OSCE* (2 hours and 20 minutes) of 20 stations of 7 minutes each. The examination was blueprinted from the clinical core curriculum with eight clinical examination, six communication, four practical skills and two radiology stations. Each station was marked against a checklist by one examiner.

2 *Two history-taking long cases* (21 minutes each). These assessed the candidate's interaction with real unstandardized patients. Candidates had 14 minutes, observed by the examiner(s), to interview the patient. Physical examination was not carried out. They then presented the case in 7 minutes to the same examiner(s). A checklist was used to measure the data-gathering process and global scores given for the presentation and candidate's attitude to the patient. Each candidate had two cases with different examiners.

Statistics

The reliability of the composite examination was estimated using multivariate generalizability theory.[8] This allows estimation of multiple true and error score variances, each true and error score being associated with each subtest. The approach pools variance components and covariance components across subtests to a single composite estimate. All scores on items within subtests were expressed on the same percentage scale. Variance components per subtest were then estimated using a one-facet generalizability design with items nested within persons (students). Covariance components were estimated for each subtest combination from the product of the respective variance components weighted by their intercorrelation.

Thus a matrix of person variance components and error variance components was obtained and used to estimate a composite reliability coefficient. The reliability coefficient can be interpreted as appropriate for absolute score interpretation. It is a more demanding interpretation of examination scores, yielding lower reliability estimates than a more common relative score interpretation (e.g. norm referencing). The approach allows optional subtest weighting and assessment of the contribution of each subtest to composite reliability. The latter was used to find directions for improving the overall reliability by changing the weights and number of items within each of the subtests. The approach also allows estimation of 'true' or disattenuated correlations between subtests. More detailed technical information can be found in Brennan[8] and Hays *et al.*[9]

Results

A total of 214 candidates took the examination. The total test time was 11 hours 32 minutes, comprising 8 hours written and 3 hours clinical. Table 1 gives details of the number of items, length, average percentage score, standard deviation (SD) and the lowest and highest scores obtained by candidates in each test.

Table 2 gives the disattenuated correlations, i.e. the true correlations after factors contributing to the variance between the tests have been corrected for, between the individual examination components. The

Table 1 Descriptions of the individual examination components

Examination component	No. of items	Testing time, minutes	Average score, %	SD	Lowest score obtained, %	Highest score obtained, %
True/false items	450	160	66·7	7·9	44·4	85·3
EMQ	25	20	68·7	9·7	36·0	96·0
SAQ	10	180	64·9	5·4	53·5	79·4
OSCE	20	140	69·9	5·0	55·6	82·2
Long cases	2	42	67·6	10·2	39·0	92·9
Essay	3	150	58·7	8·3	28·6	85·7
Total	510	692	66·1	7·8	42·9	86·9

The examination was undertaken by 214 candidates.
SD, standard deviation; EMQ, extended matching questions; SAQ, short-answer questions; OSCE, objective structured clinical examination.

Table 2 Disattenuated correlations between the individual examination components

	EMQ	SAQ	Essay	OSCE	Long case
True/false items	0·43	0·46	0·33	0·21	0·01
EMQ		0·60	−0·08	0·83	0·48
SAQ			0·49	0·76	0·54
Essay				0·30	0·51
OSCE					0·59

Table 3 Reliability scores for different weightings of examination components

Applied score weighting	Generalizability coefficient for combined components
Components weighted equally	0·77
Double weighting for OSCE and long cases	0·76
Weighted according to number of items	0·93
Weighted according to testing time	0·83

correlation coefficients for the MCQ factual knowledge test with the clinical components were 0·28 for the OSCE and 0·04 for the long cases. Correlation coefficients for the extended matching questions with the short-answer questions and OSCE were much higher; at 0·72 and 0·77, respectively. Correlations for the short-answer written paper showed the most consistent relationship with the other components: OSCE 0·78, long cases 0·54, MCQ 0·56 and essay 0·54.

The composite reliability scores estimated for the different weightings of the examination components are given in Table 3. If each examination format has an equal contribution to the reliability, regardless of the number of items or test time length, the overall reliability is 0·77. If reliability is estimated by taking into account the number of items in each test, the reliability increases to 0·93. However this capitalizes on the very large number of items in the MCQ. If test length is taken into account instead, i.e. the long cases and extended matching questions contribute less, a reliability of 0·83 is achieved. In this examination, the contribution of the clinical test was doubled. Weighting the composite test in this way reduced the reliability, slightly, to 0·76.

Discussion

This final examination aimed for high content validity in assessing the skills required of a final-year medical student about to qualify as a pre-registration house officer. As a result, estimation of its overall reliability was difficult because of the composite nature of the tests used. The problems relate to the large choice of essay questions, the random allocation of the long cases, the very large number of items in the MCQ compared with the other papers and the differences in test length. Application of multivariate generalizability theory enabled us to take these variances into account, estimate the overall reliability and achieve a more meaningful analysis of the impact of each test on the examination overall. We have shown that the composite reliability of the examination was 0·77 when each component was given equal weight and differences in test structure accounted for. For a high-stakes examination this should be taken as the minimum acceptable level. A higher value would be desirable.

We have also shown the impact of the different test structures on the composite reliability. When candidate scores were weighted according to the number of test items, the contribution of the MCQ component dominated and the reliability increased to 0·93. This reliability is more acceptable for a high-stakes test but the examination also has a important accountability function, i.e. it was designed to ensure students were clinically competent to 'pass out' of medical school. It could be argued that the content validity of the overall test should be adjusted so that, despite the smaller number of items, clinical tests were equally important. This is further supported by the analysis of this examination as the MCQ factual knowledge test correlated so poorly with the other components. In the final examination the adjustment was made by doubling the weight of the clinical test scores. We have shown that this resulted in a slight fall in the overall reliability, which means that care must be taken to balance the content of these examinations. The number of MCQ items used could have been reduced by half without significantly affecting the composite reliability of the examination.

Alternatively, by adjusting the calculations so that each test had the same length, a modest increase in reliability to 0·83 was obtained. Thus we have demonstrated that by carefully constructing composite examinations, adjusting test length, avoiding imbalance of test items and giving a large choice of questions, the reliability could be improved at the same time maintaining content validity. Hays *et al.* reported a similar experience when analysis of the Royal Australian College of General Practitioners' Certification Examination was carried out.[10]

Was the choice of a variety of tests justified? By adjusting for the variances in the tests, the disattenu-

ated correlations between the components give some idea of whether the components are testing similar or different skills. The short-answer questions were aimed at testing the candidate's problem-solving skills when faced with common clinical management problems, and the correlations of this test (around 0·5–0·7) with both the knowledge-based and clinical tests, suggest that the skill being tested was different but interrelated. The correlation seen between the basic knowledge test (the MCQ) and the clinical tests is surprisingly low. The MCQ used questions developed over years, testing straight factual, textbook knowledge about diseases. An explanation could be that this knowledge has little relation to more clinically based knowledge required by candidates for performance in the OSCE and long cases, and that the SAQ was a more appropriate test of the application of knowledge. The MCQ content may need review, with the inclusion of more extended matching questions as these showed a stronger correlation with the SAQ and OSCE tests. These may be more appropriate in these final stages of the curriculum.

Some members of the examination board, who felt that written debate was not a skill essential to this test of clinical competence, had questioned the inclusion of an essay paper. If the composite reliability of the examination is calculated excluding the essay paper and giving equal weight to the others, the overall reliability does not improve but falls to 0·75. Some clinicians were insistent that good writing skills are essential to future professional practice. We found no evidence to suggest that the inclusion of the essay paper detracted from the overall quality of the examination.

Thus using a variety of tests to improve the content validity of this final examination and test the range of skills required of a pre-registration house officer resulted in an examination of reasonable reliability. This could however, be improved by more careful balancing of the number of test items and length of each test. Given the complexity of the skills being tested, and as the emphasis on the development of professional attitudes, communication skills and a patient-centred approach to medicine gains momentum in the UK medical curriculum,[1] further research into the best format for testing clinical competence at the end of the undergraduate curriculum is urgently needed.

Acknowledgements

We thank the chairman of the examination board, Dr Charles Twort, Professor Gwyn Williams and Professor Roger Jones for their support with this study. Dr Mark Kinirions and Professor Amanda Ramirez organized the short-answer and essay papers. Mr Ron Hoogenboom carried out the statistical analysis.

Contributors

Dr Val Wass and Dr David McGibbon were involved in the running of the examinations and the collating of the results. Professor van der Vleuten undertook the statistical analysis and supported Dr Wass in the research project.

Funding

None.

References

1 The General Medical Council Education Committee. *Tomorrow's Doctors. Recommendations on Undergraduate Medical Education*. London: General Medical Council; 1993.
2 Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;**1**:41–67.
3 Harden RM, Gleeson FA. ASME medical educational booklet no. 8. Assessment of medical competence using an objective structured clinical examination (OSCE). *J Med Educ* 1979;**13**:41–54.
4 Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1996;**22**:325–34.
5 Frijns PHAM, Van der Vleuten CPM, Verwijnen GM, Van Leeuwen YD. The effect of structure in scoring methods on the reproducibility of tests using open ended questions. In: W. Bender, RJ Hiemstra, AJJA Scherbier, RP Zwierstra, eds. *Teaching and Assessing Clinical Competence.* Boekwerk; 1990; pp. 466–71.
6 Van der Vleuten CPM. The validity of final examinations. *BMJ* 2000;**32**:1217–9.
7 Case SM, Swanson DB. Extended matching items: a practical alternative to free response questions. *Teaching Learning Med* 1993;**5**:107–15.
8 Brennan RL. *Elements of Generalisability Theory*. Iowa: American College Testing Program; 1983; pp. 133–5.
9 Hays RB, Fabb WE, van der Vleuten CPM. Reliability of the Fellowship Examination of the Royal Australian College of General Practitioners. *Teaching Learning Med* 1995;**7**:43–50.
10 Hays RB, van der Vleuten CPM, Fabb WE, Spike NA. Longitudinal reliability of the Royal Australian College of General Practitioners' Certification Examination. *Med Educ* 1995;**29**:317–21.

Received 28 *July* 2000; *editorial comments to authors* 1 *August* 2000; *accepted for publication* 13 *November* 2000