

# The face validity of a final professional clinical examination

Michael Tweed<sup>1</sup> & John Cookson<sup>2</sup>

**Objective** To develop new methods of evaluating face validity in the context of a revised final professional examination for medical undergraduates, organized on three sites, over 2 days.

**Methods** The opinion of the students and examiners was surveyed by Likert-style questionnaires, with additional open comments. Expert opinion was gathered from external examiner reports and a recent Quality Assurance Agency (QAA) Subject Review Report.

**Results** The questionnaires had an overall response rate of 84%. Internal reliability, assessed by comparing responses to appropriate questions, was good with an equivalence of 45% (weighted kappa 0.54) for the students and 33% (weighted kappa 0.41) for the

assessors. There was little evidence of inconsistency between days or sites. The majority of the opinions from the students, examiners and external experts were positive. Negative comments related to time pressure and case mix.

**Conclusion** The measurement of face validity proved feasible and valuable and will assist in the further development of the course and the examination.

**Keywords** Education, medical, undergraduate, \*standards; \*educational measurement; faculty; knowledge, attitudes, practice; questionnaires; reproducibility of results.

*Medical Education* 2001;35:465–473

## Introduction

Evaluating the reliability, validity and practicality of any assessment tool, or examination, is vital. Components of validity are defined in different ways: internal and external; content, criterion (predictive and concurrent) and face validity.<sup>1–4</sup> Content validity is the degree to which the sample assessed represents the domain of the tasks to be measured. Criterion validity is a measure of how the current test performance predicts future performance (predictive) and how it compares with other tests (concurrent).

Face validity is an important component of validity.<sup>5</sup> It has been defined as a complex, multidimensional construct which is useful for evaluating how test items appear to respondents and others.<sup>6</sup> More simply it is the judgements that students, examiners and experts make<sup>4,7</sup> and it may be viewed as a measure of credibility.<sup>8</sup> Optimal evaluation of test validity requires the test to have been undertaken and the scores known.<sup>2</sup>

Even extensive evaluations of new assessment tools do not always report face validity,<sup>9</sup> although it may have been measured.<sup>10</sup> A combined search for 'face validity or credibility' with 'examinations or assessments' with 'clinical competence' using MEDLINE, from 1960 to May 1999, and Cinahl, from 1982 to May 1999 revealed only eight articles. None of these concerned the measurement of face validity by students, examiners and experts together, in assessments of clinical competence. Although this search strategy is not comprehensive, it is clear that the reporting of face validity is not widespread.

Face validity judgements are perceptions and do not have to be correct. Whatever the true validity of the assessment tool, if students, examiners or outside experts do not believe it to be good, then the tool and the results produced may not be taken seriously. For students, motivation and test performance will be diminished.<sup>5,11</sup> Teachers and examiners will not feel that the time and effort used, in test development and scoring, is justified. Face validity may be improved if students are told something of the results of any evaluation of the examination, such as examiner reliability. However, the assumption that students will understand the significance of any evaluation may not be justified.<sup>12</sup>

<sup>1</sup>Specialist Registrar, Department of Respiratory Medicine, Glenfield Hospital, Leicester, UK

<sup>2</sup>Professor of Clinical Education, Division of Medical Education, University of Leicester, Leicester, UK

Correspondence: M J Tweed, Division of Medical Education, Leicester Warwick Medical School, University of Warwick CV4 7AL, UK

### Key learning points

Measurement of face validity of examinations is important.

Measurement of face validity for students and examiners was feasible, by the use of questionnaires and by the use of external examiner and the Quality Assurance Agency (QAA) reports.

The majority of the results were positive for this examination and will prove useful in future planning of the course and examinations.

In a world of evidence-based medicine and evidence-based education,<sup>13,14</sup> evidence-based assessment tools will be necessary. The intention is to evaluate the reliability and validity of the assessment tools in use at Leicester.

The purpose of this study was therefore twofold: to gain experience in the development and use of tools for determining face validity, and by these means to evaluate a new method of assessment for final-year students.

## Methods

### The Assessment test

The course at Leicester Medical School changed in 1994. The new integrated course has a two-phase design.<sup>15</sup> Phase 1 is modular, covering aspects of human disease and health in an integrated fashion. Phase 2 comprises clinical blocks in various hospital placements and general practice.

The final examination consists of a written test concerning patient management problems and a clinical examination. Although evaluation of the written examination is underway, this study is concerned with the clinical component. The clinical examination is sequential<sup>16</sup> and the marking format and criteria are modified from the Leicester Assessment Package.<sup>7</sup> Each student sees four patients on two hospital sites over 2 days. Those students judged not to be entirely satisfactory see four more patients on a third day. Final judgements are based on eight cases. The examination tests consultative, clinical examination, problem solving, patient management and communication skills. The process is fully observed by pairs of examiners who score performance, according to clear descriptors. It takes place over three hospital sites.

The examiners are all in clinical practice as NHS consultants, honorary consultants or general practi-

tioners. Many examiners have undergone training in assessment methods and the results of this have been reported.<sup>17,18</sup> In the examination, new assessors are invited to observe experienced colleagues before they themselves score students' performance. Less experienced examiners are paired with more experienced colleagues. External examiners who are invited by the Faculty, are recognised as being experts in the field of assessment of medical undergraduates.

The patients in the study were real, not simulated, and came from a wide range of clinical practice. Clinical teachers in Leicester had been given guidelines for the identification of suitable patients. The latter should present common clinical problems, reflecting the main body systems and include patients with psychological problems. The problems selected were to be explicitly related to the course objectives, with history and examination findings that related well together. Patients with just 'a good sign' were not generally considered to be suitable.

### Data collection: participants and questionnaires

The sources of data were students, the examiners, external examiners and the Subject Review Report for the Quality Assurance Agency (QAA) for Higher Education.

Information was collected from the students and examiners (assessors) using a questionnaire. Because of a lack of literature on face validity measurement tools, the experience of others was sought using a medical education Listserver.<sup>19</sup> This produced a questionnaire and some results (personal communication, Professor M Albanese, University of Wisconsin). This questionnaire has been applied to a year-end skills examination (an objective structured clinical examination). Also, a questionnaire has been used to evaluate the Leicester Assessment Package (LAP).<sup>7</sup> Both use a 7-point Likert scale, the LAP questionnaire having more extensive score descriptors.

Our questions were loosely based on these previous questionnaires and were designed to study perceptions of the examination as a whole, as well as various components including its acceptability, its relationship to course work, and differences in student-examiner interaction across the hospital sites. We wanted to limit the questions to one side of A4 paper to avoid questionnaire fatigue and to include space for free-text comments if participants wished to make them. Two new questionnaires, for students and for examiners, were developed. These were piloted in an earlier clinical examination and proved to be feasible and reliable.<sup>20</sup> Minor alterations were made as a result (see Appendix).

Questionnaires were given to all students and examiners at the end of each of the first 2 days of the examination. Responses were anonymous and were collected immediately. At this time the students did not know whether they had passed this phase of the examination.

For ease of presentation and meaning, data from the Likert scale were analysed in two ways. Responses were categorized into 'agreement' (Likert 1–3), neither (4) or disagreement (Likert 5–7). The 95% confidence intervals for the proportion who were in the agreement category, were calculated. If the confidence interval did not include the midpoint then the 'agreement' or 'disagreement' was considered significant ( $P < 0.05$ ). Mean scores were also calculated, although the nature of the data does not allow further analysis.

Responses to similar questions were compared to assess internal reliability. The number of respondents in each category was compared to assess day-to-day consistency, as this should be independent of the other confounding factors.

Open comments on the questionnaire were categorized by one author (M.T.), by topic and into positive and negative. This process was repeated with the topics generated, to ensure consistency and subject of categorization. Questionnaire response validity was assessed by comparing the responses to two different types of data collection, the Likert results and the open comments.

Since most examiners stayed at one site and the students moved, it was possible to evaluate perceptions of student–examiner interactions by sites.

Reports from the seven external examiners were written independently from this study in a standardized format in response to questions. Comments pertinent to the clinical examination were noted. The comments were categorized by one author (M.T.), by topic and into positive and negative. This process was repeated to ensure consistency and subject of categorization.

Further external assessment was obtained from the recent Subject Review Report section on Teaching, Learning and Assessment.<sup>21</sup> The Quality Assurance Agency for Higher Education is an independent body which provides an integrated quality assurance service for higher education institutions throughout the UK.<sup>22</sup> Appointed assessors visit universities to study all aspects of courses, including: curriculum design, content and organization; teaching, learning and assessment; student progression and achievement; student support and guidance; learning resources, and quality management and enhancement.

As questionnaires were anonymous and this formed part of an ongoing evaluation of the course and

assessment, consent forms and ethical approval were not felt to be necessary.

### Statistical analysis

The 95% confidence intervals for the proportions in agreement were calculated from the standard errors of the proportions.

The agreement was measured in terms of equivalence of the seven Likert scores and weighted kappa. The calculation,<sup>23</sup> value<sup>24</sup> and application<sup>25</sup> of this technique have been described.

Internal consistency for the question responses between days is calculated by comparing the number of respondents who answered with any particular Likert score or category between the 2 days. This was accomplished by calculating the standard error for the proportion in any response group and from this the 95% confidence intervals of the difference between the 2 days. Also differences between all the responses for the group as a whole were evaluated using the Mann–Whitney test.<sup>23,24</sup>

The chi-squared test<sup>23,24</sup> was used for the comparisons between proportions of responders between the hospital sites.

## Results

### Response rates

The response rate of the students was 97% (338/347), that for the examiners was 58% (110/188). Details are shown in Table 1. The variations occurred because one student was ill on one day and some examiners only examined for half a day.

**Table 1** Response rates of the students and assessors by examination site and day

	Day 1	Day 2
<i>Assessors</i>		
Site 1	21/37	20/31
Site 2	21/30	19/30
Site 3	16/30	20/30
Total response		115/188 (58%)
<i>Students</i>		
Site 1	56/57	58/59
Site 2	57/57	54/58
Site 3	58/59	55/57
Total response		338/347 (96)%

### Reliability of results

Questions 6 and 10 of the student questionnaire, which dealt with this assessment of student competence as future doctors, were used to measure internal reliability. The Likert score equivalence was 45% (weighted kappa 0.54) and the categorical score equivalence was 72% (weighted kappa 0.53).

Similarly question 9 and questions 4–8, of the assessor questionnaire concerned assessment of core competence and its component categories. The mean of questions 4–8 (0.5 rounded up) was compared with question 9. The Likert score equivalence was 33% (weighted kappa 0.41) and the categorical score equivalence was 82% (weighted kappa 0.35).

### Consistency of the questionnaire over time

Six of the student questions (3, 4, 5, 6, 9 and 10) should be independent of patients, hospitals and assessors as they deal with the overall organization of the examination, and can therefore assess changes over

time; others are more dependent on individuals and local hospital organization. The former questions were analysed for differences between day 1 and day 2. There were seven possible Likert responses and three possible categories to six questions, giving 42 possible Likert responses and 18 possible categories. The number of respondents in each was compared between day 1 and day 2. Of the 42 response proportions, 40 were not different ( $P > 0.05$ ), and none of the 18 categories were different. There was no difference in response scores to any of these questions, between days, for the respondents as a whole ( $P > 0.05$ ).

Seven of the assessor questions (1, 4, 5, 6, 7, 8 and 9) should also be independent of patients, hospitals and students. They were analysed similarly. There was no significant difference between the proportions of respondents who answered with any particular Likert score or within each response category between the 2 days ( $P > 0.05$ ). There was no difference in response scores to any of these questions, between days, for the respondents as a whole ( $P > 0.05$ ).

**Table 2** Student and assessor percentage responses and mean score\*

	Agree	95%CI	Neither	Disagree	Mean
<i>Student questions</i>					
1 The examiners were helpful and non-intimidating	82	78–87	10	8	2.23
2 The instructions from the examiners were clear	86	82–90	6	8	2.33
3 The history skills assessed corresponds with what I was taught	79	75–83	12	9	2.62
4 The clinical examination skills assessed corresponds with what I was taught.	77	73–82	13	10	2.63
5 The problem list/management assessed corresponds with what I was taught.	77	73–82	13	10	2.75
6 The exam fairly and accurately assessed my ability	67	62–72	14	19	3.11
7 There was enough time to complete what I wanted to do/demonstrate.	39	34–45	9	51	4.22
8 The clinical scenarios were realistic.	77	73–82	11	12	2.67
9 The core competencies objective list was helpful in preparing for the exam.	59	54–64	22	19	3.30
10 The exam was a good assessment of my competence as a future doctor.	61	56–66	19	20	3.26
11 The exam is what I expected.	68	63–73	18	14	3.03
12 I learnt from the exam experience.	77	72–81	16	7	2.54
<i>Assessors' questions</i>					
1 The mark sheet and criteria list were user-friendly	68	59–77	17	15	2.89
2 The case scenarios were realistic	87	81–93	6	6	2.40
3 I was prepared for being an assessor in this exam format.	90	84–96	4	6	2.34
4 The exam was a good assessment of history skills.	83	76–90	11	6	2.45
5 The exam was a good assessment of clinical examination skills.	71	62–79	14	15	2.96
6 The exam was a good assessment of communication skills.	87	81–94	5	7	2.50
7 The exam was a good assessment of cognitive skills.	86	79–92	7	7	2.68
8 The exam was a good assessment of clinical management.	86	79–92	8	5	2.69
9 The exam was a good assessment of core competence.	82	75–89	13	5	2.74
10 There was enough time to assess the students	53	43–62	16	31	3.64
11 There was enough clinical variety to assess the students	59	50–68	19	22	3.29
12 The students' abilities were above my expectations	33	23–42	28	39	4.14

\*Percentage of respondents who were in agreement (Likert 1–3), neither agreement nor disagreement (Likert 4) or disagreement (Likert 5–7). The 95% confidence intervals for the percentage agreement and mean Likert score are also given

### Evaluation of the examination process

Table 2 shows the student and assessor views on the examination process. For students the main concern was time, and for assessors it was time and clinical variety.

To address the possible differences in examiner bias between hospitals, responses from student questions 1 and 2 and assessor question 12 were compared between the various sites. None of the differences were statistically significant ( $P > 0.05$ ). Assessors were equally helpful, non-intimidating and clear in their instruction in each of the hospitals. The assessors in each hospital had equal expectations of the students prior to the examination.

### Open comments

One or more comments were made by 41 of the assessors (35.6% of the responders) and 86 (24.8%) of the students. The topics covered are shown in Table 3.

Of the assessors' comments, 44% (15% of the total) expressed concerns about the adequacy of time and 19% about patient selection. The main student anxiety (53% of responses) concerned time.

**Table 3** Questionnaire open comments by topic\*

	Positive	Negative
<i>Assessor topics</i>		
Assessment sheets	1	10
Format of the examination	5	1
Grade system for pass/fail	2	
Grade system for pass/higher		3
Patient selection		13
Hospital environment		1
Scope of testing		2
Time for examiners		2
Time for students		28
<i>Student topics</i>		
Examiners	7	4
Faculty information		4
Format of examination		3
Learning experience of examination	2	
Organization of examination		2
Patient selection		8
Previous practice		6
Hospital environment		3
Time for students		44

\*Topics covered by the open comments, categorized into positive and negative regarding the examination process, based on 338 student and 110 examiner responses

### The external examiner reports

Comments from the seven external examiner reports were categorized by topic and are shown in Table 4.

The majority of the positive comments related to the assessment of student ability and the organization of the examination. The majority of the negative comments concerned the marking schedule and grading system, especially the initial pass/fail criterion which was felt to be unduly harsh.

### Subject review report

The 6000-word QAA Report considered many other matters besides this examination. However the final professional examination was regarded as 'an innovative and effective assessment of the course objectives'.

### Discussion

The overall response of 84% was excellent, probably because the questionnaire was short and could be completed quickly. The high response rate echoes the findings of the pilot study.<sup>20</sup> This high response from the students probably relates to the feedback culture which the Faculty has developed. The students give feedback to the Faculty, at several points during the course: they reply to questionnaires, attend staff-student forums and/or send delegates to Faculty committees. Currently there is no such culture amongst examiners. Responses to open questions were few and mostly negative. It is likely that respondents were more likely to comment if they had something negative to say. A more structured set of comments was obtained from the external examiner reports, and 73% of their comments were positive compared with just 12% of the assessors' open comments.

**Table 4** Comments from external examiners' reports\*

Topic	Positive	Negative
Assessment of student ability	11	1
Coherence with course objectives	6	
Feedback/learning	2	1
Marking schedule/grading	8	8
Organization of examination	12	
Patient selection	2	1
Reflection of clinical practice	1	
Sufficient time for examiners	1	5

\*Topics covered by the seven external examiners, taken from their structured reports and categorized into positive and negative regarding the examination process

Attempts to evaluate the true reliability, validity and practicality of this examination are underway, and it will be interesting to compare the results with the perceptions gathered from this study. As there is no gold standard, true validity is a difficult concept to evaluate. Although some questions related to content, the study remains one of face, not content, validity as perceptions and not true content are being assessed. If the content is truly defective, then the assessment will not be rescued by a high level of face validity. The reverse is not of course true. Again, as this study was measuring perceptions, the word 'good' was used in the questionnaire without definition, as face validity is a matter of judgement rather than measurement of some exact criterion.

There are some limitations associated with using Likert scales to collect these data, as they are prone to bias and other analytical difficulties. Response bias occurs when participants, rather than answer honestly, respond in a manner that they believe is acceptable or expected. Making questionnaires anonymous reduces this bias.<sup>26</sup> Placing the positive response on the left of a Likert scale will lead to more positive results with less variance.<sup>29a</sup> This bias could have been avoided by using four randomly allocated questionnaires with the positive result as the right or left, high or low score but this would have reduced the practicality of the questionnaire. Interpretation of Likert scale data is not uncomplicated.<sup>28,29</sup> Even though, as in this study, appropriate statistical analysis is used, a result may be of statistical significance but of no useful importance. We do not know what degree of change in the Likert scale is significant and requires action.

Data collected may be improved by using other scoring systems. A visual analogue scale (VAS) may be better than a Likert scale for these purposes. Compared with a five-point and possibly a seven-point Likert scale, a VAS has greater variance and theoretically greater responsiveness.<sup>27</sup> However, it is less practical, requiring either a computerized optical reader or more time for analysis.

The internal reliability of the questionnaire was good, with agreement between similar questions, and little variation in the proportion of respondents answering each option over the 2 days. Nevertheless it is possible, that these groups conceal individual differences. The internal reliability was an improvement on that of the pilot study,<sup>20</sup> probably due to changes in the wording of the questions and improvement in completion rates by the examiners. Repeating the questionnaire later might serve as a further test of reliability. However, the students' knowledge of the results could affect their opinion.<sup>11</sup> The open comment response was poor, but provides some confirmation of the questionnaire responses.

Some responses were difficult to analyse. Responses to assessor question 12 were troublesome. If an assessor disagreed with this statement, it is not clear if they believed the students were as they expected or worse. Modifying the questionnaire to overcome this by changing the scale descriptors would complicate and also lengthen it.

The use of external examiners and the QAA Report as expert opinions on validity has benefits and limitations. External examiners have been shown to maintain reliability and validity.<sup>4</sup> Although impartial, they are chosen by the institution. The institution does not choose QAA assessors. They are impartial and objective, but their reports consider much else apart from assessment.

Our study indicates that it is feasible to collect data on the face validity of an assessment tool. Further, the methods used have produced useful data, which could be used to improve the course and assessment in the future, whilst reassuring students and staff of the positive outcomes. As the face validity of an assessment tool is under the control of those who set the test,<sup>5</sup> when the results of the other evaluations of reliability and validity are known, the Faculty will be able to make changes and further reassure candidates and examiners, where appropriate.

The opinions of the students, assessors and experts about the examination process were mostly positive. Issues to be addressed are the perceived shortage of time, for both students and examiners, and the need for a better spread of cases which was identified by the examiners. Time constraints are inevitable, given the need to maintain practicality. Further assessment of these issues will include asking current house officers who took the examination how the time pressures and case mix of the examination compare with those of real consultations. Case mix could be improved by more rigorous application of the guidelines, reaffirming these in the documentation sent to teachers in the search for cases, leading to better 'blueprinting' of cases to the course objectives.

We continue to develop this method for assessing face validity in the course of evaluating our examinations.

## Contributors

Dr Tweed's contribution to this paper was the conception of the idea, development of the questionnaires, analysis of the results and initial writing of the paper. Professor Cookson's contribution was supervision and advice during the research and the subsequent writing of the paper.

## Funding

There was no external funding for this study.

## References

- 1 Wainer H, Braun HI, eds. *Test Validity*. New Jersey: Lawrence Erlbaum Associates; 1988.
- 2 Linn RL, ed. *Educational Measurement*. Phoenix: Oryx Press; 1993.
- 3 Gronlund N. *Assessment of Student Achievement*. 6th edn. Boston: Allyn and Bacon; 1998.
- 4 Brown G, Bull J, Pendlebury M. *Assessing Student Learning in Higher Education*. London: Routledge; 1997.
- 5 Chan D, Schmitt N, DeSchon RP, Clause CS, Delbridge K. Reactions to cognitive ability tests: the relationships between race, test performance, face validity perceptions, and test-taking motivation. *J Appl Psychol* 1997;**8**:300–10.
- 6 Thomas SD, Hathaway DK, Arheart KL. Face validity. *Western J Nursing Res* 1992;**14**:109–12.
- 7 Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritized criteria in the Leicester assessment package. *Br J Gen Pract* 1994;**44**:109–13.
- 8 Matsell DG, Wolfish NM, Hsu E. Reliability and validity of the objective structured clinical examination in paediatrics. *Med Educ* 1991;**25**:293–9.
- 9 Newble DI, Hoare J, Elmslie RG. The validity and reliability of a new examination of the clinical competence of medical students. *Med Educ* 1981;**15**:46–52.
- 10 Newble DI. Eight years experience with a structured clinical examination. *Med Educ* 1988;**22**:200–4.
- 11 Chan D, Schmitt N. Video-based versus paper-and-pencil method of assessment judgement tests: subgroup differences in test performance and face validity perceptions. *J Appl Psychol* 1997;**82**:143–59.
- 12 Allison R, Katona C. Audit of oral examinations in psychiatry. *Med Teacher* 1992;**14**:383–9.
- 13 Petersen S. Time for evidence based medical education. *BMJ* 1999;**318**:1223.
- 14 Hart I. Best evidence medical education. *Med Teacher*. 1999;**21**:453–4.
- 15 Anon. *Course Structure of Medical Degree at Leicester*. Available from URL. <http://www.le.ac.uk/medicine/FMBS/medicine/costruct/costruct.htm>.
- 16 Cookson J, Petersen S. Sequential testing in clinical examinations. *Med Teacher* 1999;**21**:222.
- 17 Preston-Whyte ME, Fraser RC, McKinley RK. Teaching and assessment in the consultation: a hospital clinicians' preparatory workshop for integrated teaching of clinical method to undergraduate medical students. *Med Teacher* 1998;**20**:266–7.
- 18 Preston-Whyte ME, Fraser RC, McKinley RK. Teaching senior hospital clinicians to systematically assess and enhance undergraduate consultation skills. *9th International Ottawa Conference on Medical Education*. 2000; p. 205.
- 19 Access to the Med Ed listserver. Available from URL. <http://www.mailbase.ac.uk/lists/medical-education>.
- 20 Tweed M, Cookson J. Feasibility of testing for face validity of an assessment of clinical competence. *ASME Scientific Conference Abstracts*. 1999; p. 29–30.
- 21 Anon. *Subject Review Report for Medicine at Leicester University*. Available from URL. <http://www.niss.ac.uk/education/hefce/qar/q188-99.pdf>.
- 22 Anon. *Description of the QAA*. Available from URL. <http://www.qaa.ac.uk/aboutqaa.htm>.
- 23 Armitage P, Berry G. *Statistical Methods in Medical Research*. 3rd edn. Oxford: Blackwell Scientific Publications; 1994.
- 24 Anthony D. *Understanding Advanced Statistics*. London: Churchill Livingstone; 1999.
- 25 MacClure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987;**126**:161–8.
- 26 Bucher L. Evaluating the affective domain. Consider the Likert scale. *J Nursing Staff Develop* 1991; **7**(5):234–8.
- 27 Pfennings L, Cohen L, van der Ploeg H. Preconditions for sensitivity in measuring change: visual analogue scales compared to rating scales in a Likert format. *Psychol Rep* 1995;**77**:475–80.
- 28 van Alphen A, Halfens R, Hasman A, Imbos T. Likert or Rasch? Nothing is more applicable than good theory. *J Adv Nursing* 1994;**20**:196–201.
- 29a Albanese M, Prucha C, Barnet JH, Gjerde CL. The effect of right or left placement of the positive response on Likert-type scale used by medical students for rating instruction. *Acad Med* 1997;**72**:627–30.
- 29 Roberson PK, Shema SJ, Mundfrom DJ, Holmes TM. Analysis of paired Likert data: how to evaluate change and preference questions. *Fam Med* 1995;**27**:671–5.

Received 19 January 2000; editorial comments to authors 27 March 2000; accepted for publication 16 May 2000

## Appendix

### Assessor questionnaire

Site: LRI GH LGH (please circle)  
Day: Wed Thur

Thank you for taking a few minutes to answer this questionnaire. Please circle the number that you feel most appropriately reflects how you feel about the finals exam (clinical).

	1	2	3	4	5	6	7	
	Strongly agree	agree	tend to agree	Neither	tend to disagree	disagree	strongly disagree	
<b>1. The mark sheet and criteria list were user-friendly.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>2. The case scenarios were realistic.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>3. I was prepared for being an assessor in this exam format.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>4. The exam was a good assessment of history skills.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>5. The exam was a good assessment of clinical examination skills.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>6. The exam was a good assessment of communication skills.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>7. The exam was a good assessment of cognitive skills.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>8. The exam was a good assessment of clinical management.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>9. The exam was a good assessment of core competence.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>10. There was enough time to assess the students.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>11. There was enough clinical variety to assess the students.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>12. The students' abilities were above my expectations</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>PTO: Please write any comments or suggest ways to improve the examination.</b>								



**Student questionnaire**

Site: LRI GH LGH (please circle)  
 Day: Wed Thur

Thank you for taking a few minutes to answer this questionnaire. Please circle the number that you feel most appropriately reflects how you feel about the finals exam (clinical).

	1	2	3	4	5	6	7	
	Strongly agree	agree	tend to agree	Neither	tend to disagree	disagree	strongly disagree	
<b>1. The examiners were helpful and non-intimidating.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>2. The instructions from the examiners were clear.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>3. The history skills assessed correspond with what I was taught.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>4. The clinical examination skills assessed correspond with what I was taught.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>5. The problem list/management assessed correspond with what I was taught.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>6. The exam fairly and accurately assessed my ability.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>7. There was enough time to complete what I wanted to do/demonstrate.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>8. The clinical scenarios were realistic.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>9. The core competencies objectives list was helpful in preparing for the exam.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>10. This exam was a good assessment of my competence as a future doctor.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>11. This examination is what I expected.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>12. I learnt from the exam experience.</b>								
Strongly agree	1	2	3	4	5	6	7	Strongly disagree
<b>PTO: Please write any comments or suggest ways to improve the examination.</b>								