CHAPTER 5

Are specialist certification examinations a reliable measure of physician competence?[7]

Abstract

Introduction. High stakes postgraduate specialist certification examinations have considerable implications for the future careers of examinees. Medical colleges and professional boards have a social and professional responsibility to ensure their fitness for purpose. To date there is a paucity of published data about the reliability of specialist certification examinations and objective methods for improvement. Such data are needed to improve current assessment practices and sustain the international credibility of specialist certification processes.

Purpose. Determine component and composite reliability of the Fellowship examination of the College of Physicians of South Africa, and identify strategies for further improvement.

Methods. Generalizability and multivariate generalizability theory were used to estimate the reliability of examination subcomponents and the overall reliability of the composite examination. Decision studies were used to identify strategies for improving the examination composition.

Results. Reliability coefficients of the component subtests ranged from 0.58-0.64. The composite reliability of the examination was 0.72. This could be increased to 0.8 by weighting all test components equally or increasing the number of patient encounters in the clinical component of the examination. Correlations between examination components were high, suggesting that similar parameters of competence were being assessed.

Conclusion. This composite certification examination, if equally weighted, achieved an overall reliability sufficient for high stakes examination purposes. Increasing the weighting of the clinical component decreased the reliability. This could be rectified by increasing the number of patient encounters in the examination. Practical ways of achieving this are suggested.

[7] Burch VC, Norman GR; Schmidt HG, van der Vleuten CPM. Are specialist certification examinations a reliable measure of physician competence? Advances in Health Sciences Education. Submitted.

Introduction

Postgraduate specialist certification and licensure is the responsibility of a large number of medical colleges, professional boards and other associations throughout the world. These examinations are "high stakes hurdles with considerable implications for candidates' career progression, future employment and remuneration".[1] Given the high stakes nature of these examinations, certification bodies have a social and professional responsibility to ensure that they are robust, fair and defensible.[2-4] Key factors which determine fairness and psychometric defensibility of assessment processes include the reliability and validity of the assessment process. Other factors including acceptability of the assessment process to the various stakeholders, and the resources required to conduct the examination also need to be considered.[5]

Given the importance of this issue, it is thus surprising to find that a recent review of the literature identified only 55 publications, out of more than 7 000 titles and abstracts screened, describing the evaluation of postgraduate certification processes.[1] Most papers described the psychometric adequacy of specific assessment methods. Only three composite examination processes were comprehensively (both written and clinical components) evaluated: the Fellowship examination of the Royal Australian College of General Practitioners,[6] the Membership of the Royal New Zealand College of General Practitioners – Part I,[7,8] and the certification examination of the College of Family Physicians of Canada.[9] These findings are of concern for two reasons: (1) they suggest a reluctance on the part of institutions to subject themselves to external scrutiny,[1,2] and (2) evaluation of individual assessment instruments persists, despite evidence that a variety of test methods are required to adequately assess clinical competence.[4,5,10]

In this paper we evaluate the reliability of Part II of the Fellowship examination of the College of Physicians (FCP) of South Africa, a composite examination comprising written and clinical subtests. The reliability of the component subtests and the overall reliability of the composite examination are enumerated separately. We also suggest changes, on the basis of a number of decision studies (D-studies) that could further improve the quality of this high stakes certification examination.

Methods

Structure of the FCP examination

The FCP certification process comprises two parts. Part I consists of two written short-answer question tests (SAQT) assessing basic sciences knowledge relevant to clinical practice. Each 3-hour SAQT comprises four sections, and candidates are required to answer three out of

four questions in each section. Different examiners set and mark each section and assign a final score to each section expressed on a percentage scale. The overall final score achieved for Part I represents the average score achieved across all eight sections of the two papers (four sections per paper) expressed on a percentage scale. Candidates must achieve a final score of 50% or more to be awarded Part I of the FCP examination.

Part II of the certification process (Table 1) may only be attempted after obtaining Part I and completing at least 36 months of training in a recognized training institution. Component subtests of the examination focus on the integrated assessment of clinical competence, broadly defined as "the degree to which an individual can use the knowledge, skills and judgement associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice".[11] For this reason, testing methods that approximate clinical practice are used as far as possible. The SAQTs determine the ability of candidates to formulate rational, cost-effective management (investigation and treatment) plans for problems commonly encountered in specialist practice. Questions are usually phrased as clinical problems or patient scenarios. Each 3-hour SAQT comprises four sections. Two sections usually each describe a clinical problem or case scenario requiring a comprehensive, multi-component answer. The remaining two sections contain four questions each. Candidates are required to answer three questions in each section. Different examiners set and mark each section of each paper and submit a final score for each section expressed on a percentage scale. The final score for sections comprising three separate questions represents the average score achieved across the three questions expressed on a percentage scale. The data interpretation test (DIT) comprises 20 clinical scenarios accompanied by data (e.g. images of clinical signs, electrocardiograms, radiographic studies, etc.). Each test item requires interpretation of the given information and short written responses (a word, phrase, or list) to accompanying questions. This test is set and marked by a pool of examiners using standardized marking criteria. Each test item is awarded a numeric score and the final score represents the average score achieved across all test items expressed on a percentage scale.

Candidates obtaining an average score of 50% or more, for the three written subtests, are invited to the clinical component of Part II of the examination.[12] This comprises three unobserved real patient encounters (PE): one 60-minute PE followed by a 30-minute oral examination conducted by two examiners; two 30-minute PEs followed by a 15-minute oral examination of each case conducted by a different pair of examiners. Each PE is awarded a consensus score, using a criterion-referenced rating scale, expressed on a percentage scale. Successful candidates must achieve a weighted (Table 1) final score of 50% or more for Part II,[12] including a score of 50% or more for at least two PEs. Candidates are awarded an FCP after successfully completing Part I and II of the certification process.

Data analysis

The results of all FCP Part II candidates examined in September 2004, May 2005 and September 2005 were entered onto an Excel (Microsoft) spreadsheet. Of 79 examinees, 69 candidates completed all the component subtests of the examination. The item scores within each subtest of these 69 examinees were used.[6] Test item scores were entered as a mark out of 100. Subtest scores were calculated by averaging test item scores, defined as the scores assigned to the smallest independent test units within a component subtest.[6] For the SAQT, the final score represented the average score achieved across eight test sections, four per test, expressed on a percentage scale. For the DIT, the final score represented the average score achieved across 20 test items expressed on a percentage scale. For the PEs, the final score represented the average score achieved across three PEs expressed on a percentage scale. The final overall score was calculated using the weighting strategy outlined in Table 1. Descriptive statistics were carried out using Statistica Version 7 (Statsoft Inc., Tulsa, USA) software, including disattenuated correlations between subtests, i.e. correlations between component subtests after factors contributing to the variance between subtests have been removed.[13]

Reliability coefficients for the component subtests of Part II were determined using generalizability theory (GENOVA version 2.2).[14,15] The dataset was subjected to one-facet generalizability studies using an items-nested-within-persons design (i:p design) in order to estimate the reliability coefficients of component subtests. In generalizability theory terms, these represent dependability coefficients and indicate a reliability estimate from a domain-or criterion-referenced score interpretation. The standard error of measurement (SEM) of subtests was determined using the square root of the error score variance. The SEM reflects the magnitude of measurement error on the original score scale, and the 95% CI for individual test scores was estimated by multiplying the SEM by 1.96 and adding /subtracting the value to /from individual examinee scores. In the case of borderline students the SEM provides additional useful information regarding test scores since it indicates the amount of error that needs to be considered when evaluating the reliability of an individual test score.[6,10,16,17] To facilitate meaningful comparison across subtests, reliability coefficients were standardized for one hour of testing time.

The composite reliability of the FCP Part II was estimated using multivariate generalizability theory (mGENOVA).[15,18,19] A one-facet generalizability design with items-nested-within-persons-within-subtests, in which subtests were considered a fixed facet ([i:p] x s, s fixed), was used. A matrix of person variance components, including covariance components (S_p), and error score variance ($S_{i:p}$) was constructed and used to determine the composite universe score variance and error variance by summation of the respective matrix elements after weighting the entries by their appropriate number of items and subtest weights.[6] These variance

components were used to calculate a composite (domain-referenced) reliability coefficient. Decision studies (D studies) were conducted to determine the impact of different weighting strategies on overall examination reliability, and the number of PE subtest items required to achieve a composite examination reliability of 0.8 or more.

Results

Examination success rate

During the period evaluated, 79 candidates attempted Part II of the FCP examination. Overall, 69 candidates (87%) passed the written component and 54 of the 69 candidates invited to the clinical component of the examination successfully completed all the component subtests of Part II of the FCP examination (78.3% success rate). Table 1 provides the mean (SD) percentage score for each of the component subtests completed by the 69 candidates invited to the clinical component of the examination.

Table 1. *Component subtests of the FCP examination Part II.*

Component subtests	Weighting of subtests	Number of items	Testing time, minutes	Average score, %	Standard deviation
Short-answer question tests	0.2	8	360	58.7	5.4
Data interpretation test	0.2	20	180	56.8	8.8
Patient encounters	0.6	3	180	58.9	10.3
Total	1.0	31	720	58.2	8.1

Reliability of the component subtests

The individual subtest variance components for true and error variance, shown in Table 2, were used to calculate the reliability coefficient and SEM for each of the subtests reported in Table 3.

Table 2. *Estimated variance and covariance components of p (S_p) and i:p ($S_{i:p}$)*

Subtest		SAQT	PE	DIT
S_p				
	SAQT	17.49		
		± 5.05		
	PE	26.63	61.29	
		± 5.04	± 18.53	
	DIT	31.68	48.99	50.11
		± 5.18	± 8.84	± 13.24
$S_{i:p}$				
	SAQT	96.02		
		± 6.21		
	PE		131.39	
			± 15.82	
	DIT			559.23
				± 21.99

Table 3. *Reliability coefficients and standard error of measurements for component subtests.*

	Reliability coefficient		Standard error of measurement	
Component subtest	**Actual testing time**	**1-hour testing time**	**Actual testing time**	**1-hour testing time**
DIT	0.64	0.37	5.29	7.02
SAQT	0.59	0.20	3.46	4.45
PE	0.58	0.32	6.62	8.45

The reliability coefficients of individual subtests (Table 3) varied somewhat. Using a standardised testing time of one hour, it can be seen that the DIT and PEs performed better than the SAQT. The margin of error of the SAQT was, however, the least of all component subtests.

Table 4. *Disattenuated correlations between the examination subtest component.*

	Covariance	Correlation	Disattenuated correlation
DIT - SAQT	31.68	0.66	1.00
DIT - PE	48.99	0.54	0.88
SAQT - PE	26.63	0.48	0.81

Correlation of examination component subtests

The disattenuated correlations between the individual subtest components of the examination are shown in Table 4. All the correlations between the subtest components were greater than 0.8, indicating that the various subtests were highly intercorrelated.

Reliability of the composite examination

The estimated true and error variance components were also used to determine the composite reliability of Part II of the FCP examination for the cohort studied. The first D-study listed in Table 5 reflects the actual test situation which achieved a reliability coefficient of 0.72. The remaining D-studies in Table 5 reflect the reliability coefficient changes achieved by manipulating the weighting of the subtest components. An equal subtest weighting strategy achieved a composite examination reliability of 0.8.

Table 5. *Composite examination reliability coefficients using different weighting strategies.*

		DI	SAQ	PE
D-Study 1: Subtests weighted in favour of PE				
Number of items		20.00	8.00	3.00
Test duration in hours		3.00	6.00	3.00
Weighting		0.20	0.20	0.60
Reliability coefficient	0.72			
Standard error of measurement	4.17			
D-Study 2: Subtests weighted equally				
Number of items		20.00	8.00	3.00
Test duration in hours		3.00	6.00	3.00
Weighting		0.33	0.33	0.33
Reliability coefficient	0.80			
Standard error of measurement	3.02			
D-Study 3: Subtests weighted according to number of items				
Number of items		20.00	8.00	3.00
Test duration in hours		3.00	6.00	3.00
Weighting		0.65	0.26	0.09
Reliability coefficient	0.76			
Standard error of measurement	3.59			

Finally, a D-study was done to determine the number of patient encounters required to achieve a composite examination reliability of 0.8 or more (Table 6). The weighting and item

number of the written subtests were kept constant i.e. unchanged from the first D-study listed in Table 5. As can be seen from the data, a minimum of five PE cases achieved a composite examination reliability of 0.8 and reduced the 95% confidence interval (CI) of the SEM from 8.2% to 6.5%.

Table 6. *Composite examination reliability and standard error of measurement using different numbers of items for the PE subtest.*

	Composite FCP examination Part II		
Number of PE items	**Reliability coefficient**	**Standard error of measurement**	**95% confidence interval**
3	0.72	4.17	± 8.17
4	0.77	3.66	± 7.17
5	0.80	3.33	± 6.53
6	0.83	3.08	± 6.04
8	0.86	2.74	± 5.37
10	0.88	2.52	± 4.94
12	0.89	2.35	± 4.61

Discussion

Achieving major changes in assessment practices is a difficult task because (1) medical educators seem reluctant to critically review the utility of assessment practices, (2) educational practice decisions are rarely based on research outcomes, and (3) assessment practices are often dictated by the opinions, sentiments and traditions of teachers, students and institutions.[1,2,5,20] To date, fewer than 20 postgraduate specialist certification examination processes have been psychometrically evaluated and subjected to external scrutiny (published).[1] This finding supports the previous statements made. While it could be argued that all certification examinations are likely to demonstrate similar reliability parameters and, hence, the need to publish the results of such examination evaluation procedures is redundant, this opinion may not be true for at least two reasons. Firstly, the composition of these assessment processes is highly variable, as demonstrated by the few existing publications, and assumptions regarding the uniform reliability of such variably composed multi-part examinations are likely to be speculative rather than correct. Secondly, certification and licensing bodies have a social and professional responsibility to publicly demonstrate the credibility of their examination processes.[3] Currently there is very little evidence supporting specialist certification practices

worldwide; this situation should be remedied. Once sufficient published data exist, it may be possible to determine basic principles regarding the optimal composition and testing time of high stakes certification examinations. This may serve to standardize the process worldwide. Currently there are insufficient published data to facilitate this advance in certification examination practices. As such, this paper contributes to the small body of evidence regarding high stakes postgraduate assessment practices. A number of findings merit brief discussion.

Of the three previously mentioned composite specialist certification examinations subjected to psychometric evaluation, only the paper by Hays and colleagues,[6] used multivariate generalizability theory to evaluate the composite reliability of the Fellowship examination of the Royal Australian College of General Practitioners (FRACGP). The general practitioner specialist certification examinations conducted in both New Zealand[7] and Canada[9] were only evaluated for internal consistency using classical test theory procedures. Thus, our findings can only be compared to the FRACGP examination data. The FRACGP examination comprises seven component subtests: a 200-item multiple choice question paper, an 80-item data interpretation test, two computerised diagnostic problems, two written case commentaries (1 500-2 000 word reports), five role-play interviews, five patient encounters and a 30-minute structured oral examination based on a logbook of 100 recorded cases. The reported total testing time was 15 hours per candidate and the composite domain-referenced reliability of the examination was 0.79. As can be seen from this brief description, the two examinations differ considerably in terms of both composition and duration. Based on such limited data it is not possible to make any international recommendations regarding the optimal composition and total testing time of high stakes specialist certification examinations. As previously suggested, more published data are required.

Conventional wisdom around assessment of clinical competence suggests that defensible assessment requires the use of multiple test methods and an adequate number of test items.[4,5,10,21] The evidence from this analysis adds insight to both propositions. Although the test formats used in the examination are very different, in fact, the disattenuated correlations suggest that they really assess common elements. Hence, the different formats may be more essential from a perspective of content and face validity than psychometric necessity. Conversely, the value of multiple observations is not controversial; however, strategies to yield quantitative estimates of test length are rarely used. To date only two medical education publications demonstrate the value of using multivariate generalizability theory (G-theory) to determine the optimal composition of these multi-component assessment packages.[6,13] As demonstrated in our paper, and the two published papers, both the optimal weighting of component subtests and the optimal number of test items can be objectively determined using data derived from existing test results. It is, therefore, surprising that this evidence-based approach to the design and evaluation

of composite assessment strategies, described 30 years ago, is not more widely used.[22,23] Possible reasons for the lack of published data may include fear of litigation or a reluctance to publish negative findings.[1,23] and limited experience with the use of the statistical method.[1,22-24] However, medical education practices will only improve if they are subjected to regular review and improved on the basis of the best available evidence.[4,22,23]

While the composite reliability coefficient of the FCP examination is reasonable, as judged by currently suggested international practice norms,[25] the D-studies we conducted identified two potential mechanisms for further improving the overall reliability of this composite examination: (1) weight all component subtests equally or (2) increase the number of PEs. The superiority of equal weighting may appear counter intuitive at first inspection, however, it is consistent with a large body of evidence in psychometrics.[26] Still, although equal weighting of all component subtests would improve the overall exam reliability, examiners are unlikely to reduce the weighting of the clinical subtest given the primary function of the Fellowship examination, i.e. to certify professional competence. Similar views have been expressed with regard to undergraduate medical programme qualifying examinations.[13] Therefore, based on the available data the same outcome could be achieved by increasing the number of patient encounters. Published data suggest that a minimum of 10 PEs, of approximately 20-30 minutes duration each, is required to produce a reliability coefficient of 0.8 or more.[16,17,27,28] These estimates, however, were all derived using univariate analysis. Our multivariate analysis suggests that a minimum of five PEs would achieve a composite examination reliability coefficient of 0.8. By using single examiners rather than pairs, since this contributes little to improving overall reliability,[29-32] this improvement could be effected without recruiting additional examiners. This strategy would, however, lengthen the clinical testing time from 180 minutes to 225 minutes, a 25% increase. Furthermore, additional patients would need to be recruited to participate in the examination. Since each examination cycle currently requires 15 or more carefully selected cases, increasing this number to 20 or more is unlikely to be easily achieved given the multiple resource constraints faced in the public health care sector, the setting of the clinical part of the FCP examination. Thus, alternative options need to be explored.

The use of multiple workplace-based clinical assessments is an attractive alternative option.[4] This technique requires direct observation of unstandardised real PEs in the workplace.[16,17,33-35] A reliability coefficient of 0.8 or more can be achieved using 10 or more encounters of approximately 30 minutes each.[16,17] By obtaining 10 in-course assessment events per trainee, approximately three per year, borderline candidates could be identified prior to the final certification examination. Such candidates, for whom the standard error of measurement (SEM) of the assessment process is more important than its overall reliability, could selectively be

subjected to more PEs than candidates achieving an in-course rating well clear of the cutpoint.[6,10,16,17] These in-course assessment events could be conducted as part of the routine clinical duties of senior clinician-educators working at each of the training centres. Although internal examiner bias is always of concern, examiners may be less inclined to inflate the in-course scores of weaker candidates if the potentially negative outcome is clearly articulated, i.e. fewer cases in the final certification examination resulting in a greater margin of error (95% CI of SEM) and, therefore, an increased the risk of failure for borderline candidates. Not only would the accumulated in-course score contribute to the final composite examination score, thereby improving the composite reliability of the examination, but the educational impact may be considerable. Regular, directly observed in-course assessment would provide candidates with objective feedback regarding their actual observed performance and emphasize the critical importance of clinical competence.

In addition to implementing an in-course assessment strategy, the PEs in the final certifying examination should be directly observed, since this improves test validity and reliability.[28,32] Hamdy and colleagues recently reported that four directly observed 30-minute PEs followed by an oral examination of 15 minutes achieved a reliability coefficient of 0.82 using single examiners.[31] It may, thus, be possible to implement this strategy without lengthening the clinical testing time, except for a small minority of borderline candidates. Some may argue that the validity of the FCP examination may be compromised by the loss of the 60-minute PE. Evaluation of authentic workplace-based PEs, in various settings (Emergency Unit, general medical ward, outpatient clinic) using both first-time visits and repeat visits, indicates that most authentic PEs are approximately 30 minutes in duration.[16] Based on this observation it could be argued that the 30-minute PE more accurately represents current clinical practice. These two evidence-based strategies, both feasible within our context, could further enhance the reliability of the current FCP examination and deserve serious consideration.

An important consideration in this paper is the manner in which the pass /fail mark or cut score of this composite examination was determined. The cut scores for both the written component of the FCP Part II examination, as well as the composite score for all component subtests, were determined using a variation of the Hofstee method and the Contrasting Groups method.[36,37] This is described in detail elsewhere.[12] Review of the few published papers describing the evaluation of composite high stakes certification examinations, at both undergraduate and postgraduate level, did not provide helpful information regarding standard setting procedures used by other centres. The cut score of these multi-component examinations was often not mentioned, or where mentioned, no description of the method used to derive it was given.[6,13] Several years ago, Norcini and Shea emphasised that the credibility of standard setting for certification purposes is more dependent upon the use of absolute standards, expert

judgement, due diligence and supportive evidence, rather than the exact method chosen.[38] Encouraging certification bodies to publish information detailing the format, procedure and composite reliability of their examinations would facilitate the growth of a body of literature that could inform and support the way such decisions are made in the future.[38]

A limitation of this study is that the composite reliability estimate obtained did not completely mirror the actual procedure used in practice. In a composite reliability estimate the underlying assumption is made that scores on the component subtests are combined in a compensatory fashion. In our actual practice, admission to the clinical component was dependent on passing the written subtests. This does mean that some caution is needed when interpreting the data, particularly the disattenuated correlation coefficients reported. Owing to the sequential nature of the testing, candidates sitting the clinical component represent a somewhat homogeneous subset of the initial candidate pool. Consequently, the disattenuated correlations, although already very high, are, if anything, an underestimate of what would be observed if the patient component were administered to all candidates. While high correlation coefficients do not necessarily imply that two tests are measuring the same thing, it does mean that knowledge of one permits accurate prediction of the other.[39] Thus, from a statistical perspective it could be agued that little additional information is gained from the clinical subtest for candidates who have passed the written subtests.[39] Nevertheless, such a perspective is too narrow; the examination as a whole should adequately represent all domains of competence, regardless of psychometric considerations. Moreover, examiners conducting high-stakes certification examinations are unlikely to agree with this statement for good reasons, as previously articulated.

In closing then, we have added a contribution to the small body of literature describing the use of multivariate generalizability theory to evaluate the reliability of high stakes composite postgraduate certification examinations. We endorse the call for more published research in order to help certification bodies achieve a more uniform, evidence-based approach to the development of credible, defensible high stakes examinations.[38] In particular, future studies should focus on providing data regarding strategies used to determine the cut score of these examinations, an issue not described to any great extent in the existing literature.

References

1. Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Medical Education 2002; 36: 73-91.
2. Friedman M, Mennin SP. Rethinking critical issues in performance assessment. Academic Medicine 1991; 66: 390-395.
3. Lew SR, Page CG, Schuwirth LW, Baron-Maldonado M, Lescop JM, Paget NS, et al. Procedures for establishing defensible programmes for assessing practice performance. Medical Education 2002; 36: 936-941
4. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Medical Education 2005; 39: 309-317.
5. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education 1996; 1: 41-67.
6. Hays RB, Fabb WE, Van der Vleuten CPM. Reliability of the Fellowship examination of the Royal Australian College of General Practitioners. Learning and Teaching in Medicine 1995; 7: 43-50.
7. Thompson AN. An assessment of a postgraduate examination of competence in general practice: part I – reliability. New Zealand Medical Journal 1990; 103: 182-184.
8. Thompson AN. An assessment of a postgraduate examination of competence in general practice: part II – validity. New Zealand Medical Journal 1990; 103: 217-219.
9. Handfield-Jones R, Brown JB, Rainsberry P, Brailovsky CA. Certification examination of the College of Family Physicians of Canada. Part II: Conduct and general performance. Canadian Family Physician 1996; 42: 1188-1195.
10. Swanson DB, Norman GR, Linn RL. Performance based assessment: lessons from the health professions. Educational Researcher 1995; 24: 5-11.
11. Kane MT. The assessment of professional competence. Evaluation & the Health Professions 1992; 15: 163-182.
12. Burch VC, Norman GR. Turning words into numbers: establishing an empirical cut score for a letter graded examination. Submitted.
13. Wass V, McGibbon D, van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? Medical Education 2001; 35: 326-330.
14. Cronbach IJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioural measurements: generalizability for scores and profiles. New York: Wiley; 1972.
15. Brennan RL. Generalizability theory. New York: Springer-Verlag; 2001.

16. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Annals of Internal Medicine 1995; 123: 795-799.
17. Norcini JJ, Blank LL, Duffy D, Fortna GS. The mini-CEX: a method for assessing clinical skills. Academic Medicine 2003; 138: 476-481.
18. Jarjoura D, Brennan RL. A variance components model for measurement procedures associated with a table of specifications. Applied Psychological Measurement 1982; 6: 161-171.
19. Brennan RL. Manual for mGENOVA. Iowa testing programmes occasional paper number 47; 2001.
20. Nelson MS, Clayton BL, Moreno R. How medical school faculty regard educational research and make pedagogical decisions. Academic Medicine 1990; 65: 122-126.
21. Van der Vleuten CPM, Norman GR, de Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Medical Education 1991; 25: 110-118.
22. Crossley J, Davies H, Humphris G, Jolly B. Generalizability: a key to unlock professional assessment. Medical Education 2002; 36: 972-978.
23. Clauser BE, Harik P, Margolis MJ. A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. Journal of Educational Measurement 2006; 43: 173-191.
24. Tweed M, Moila J. Legal vulnerability of assessment tools. Medical Teacher 2001; 23: 312-314.
25. Royal College of Physicians and Surgeons of Canada (RCPSC). Handbook for Chairs and Members of Examination Boards. Ottawa: RCPSC; 2000.
26. Wainer H. Estimating coefficients in linear models: it don't make no nevermind. Psychological Bulletin 1976; 83: 213-217.
27. Daelmans HEM, Scherpbier AJJA, van der Vleuten CPM, Donker ABJM. Reliability of oral examinations re-examined. Medical Teacher 2001; 23: 422-424.
28. Wass V, Jolly B. Does observation add to the validity of the long case? Medical Education 2001; 35: 729-734.
29. Norcini JJ. The death of the long case? British Medical Journal 2002; 324: 408-409.
30. Wass V, Jones R, van der Vleuten C. Standardised or real patients to test clinical competence? The long case revisited. Medical Education 2001; 35: 321-325.
31. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination. Medical Education 2003; 37: 205-212.
32. Wass V, van der Vleuten CPM. The long case. Medical Education 2004; 38: 1176-1180.
33. Brennan BG, Norman GR. Use of encounter cards for evaluation of residents in obstetrics. Academic Medicine 1997;72(Suppl.):S43-S44

34. Hatala R, Norman GR. In-training evaluation during an internal medicine clerkship. Academic Medicine 1999; 74 (Suppl.): S118-S120.
35. Turnbull J, MacFayden J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. Journal of General Internal Medicine 2000; 15: 556-561.
36. Norcini JJ. Setting standards on educational tests. Medical Education 2003; 37: 464-469.
37. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teaching and Learning in Medicine 2006; 18: 50-57.
38. Norcini JJ, Shea JA. The credibility and comparability of standards. Applied Measurement in Education 1997; 10: 39-59.
39. Norman GR, van der Vleuten CPM, de Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Medical Education 1991; 25: 119-126.