Are medical postgraduate certification processes valid? A systematic review of the published evidence

Linda Hutchinson,[1] *Peter Aitken*[1] *& Tom Hayes*[2]

Objective To collate the published works on validation of assessments used in postgraduate medical certification.

Design Systematic review of original papers on reliability and validity of assessments used in medical postgraduate certification.

Setting Medical and education research databases.

Results Fifty-five papers were identified from 1985 to 2000. A wide range of approaches to validation were employed. Inter-rater reliability and internal consistency were the most reported foci for validation. There were just two papers on consequential validity, and only a few on construct validity. These two forms of validity are considered central in recent general education writing. The majority of papers were from general and family practice. There was a noticeable lack of papers from the UK Royal Colleges (except the Royal College of General Practitioners), despite 5 years of the new unified grade and the renewed emphasis on the role of the Royal Colleges in setting assessment criteria.

Conclusions There is a relative scarcity of published papers on validation of assessment for postgraduate medical certification considering the influence these high stakes processes have on doctors career progression and employment opportunities. General and family practice institutions in a number of English speaking countries have set an example to others, by showing that rigour and transparency in assessment development and implementation can be reflected in publication.

Keywords Certification; education, medical graduate/*standards; educational measurement; literature review (PT); reproducibility of results; validation studies.

Medical Education 2002;36:73–91

Introduction

Responsibility for the assessment of doctors prior to certification or licensure for practice falls to many different medical colleges, boards and associations throughout the world. The assessments they set are high stakes hurdles with considerable implications for candidates' career progression, future employment and remuneration. To be fit for this purpose the process should be fair so that there is confidence that the results are comparable amongst institutions, amongst markers and over time.[1] The traditional checks on assessment are reliability and validity. There are many subcategories of both terms some of which are not easily explained and are not used in an equivalent manner universally. The most common terms are given in Table 1 and Table 2. Other terms used commonly in assessment are given in Table 3.

This study was designed to identify from the published literature validation studies on certification assessment processes in postgraduate medical education in order to collate the different methodologies employed and examples of good practice.

Methodology

The techniques of systematic review and meta-analysis were first described at a meeting of the American Educational Research Association. Glass detailed a method of removing biases and random errors through extensive and systematic searching of the literature in preference to the selection of papers to augment one line of argument.[15] A decade later a similar case was made for clinical research.[16,17]

[1]St George's Hospital Medical School, London, UK,
[2]University of Wales College of Medicine, Cardiff, UK

Correspondence: Linda Hutchinson, Director of Education and Workforce Development, University Hospital Lewisham, Lewisham High Street, London, SE13 6LH, UK.
E-mail: linda.hutchinson@uhl.nhs.uk

Key Learning Points

Medical postgraduate certification processes are high stakes assessments with considerable implications for career progression, future employment and remuneration opportunities.

Validation of assessments in general/family practice are more represented than other fields of medicine, but there is a relative lack of published work on certification assessment validation world wide.

Apart from general practice there is a complete lack of published original studies on the new UK unified grade (Calman) assessment processes.

Few studies looked at consequential and construct validity, the two concepts of central importance in recent general education literature.

In an era of increasing transparency it is expected that award giving bodies will need to demonstrate the rigour of their assessment processes.

The systematic search strategy for this paper was set up to identify any paper on the reliability or validity of a postgraduate certification process. The strategy was applied to Medline for 1985–2000. Modifications of the strategy as appropriate were then applied to the Institute for Scientific Information (ISI) Citation Index, Embase, Educational Resources Information Center (ERIC) and Topics in Medical Education (TIME) References of included papers were also checked. Details of these searches are given in Table 4. The examinations' officers of the UK Royal Colleges were contacted by letter to ask for details of published or unpublished reports on reliability or validity of their assessment strategies. Of 17 institutions contacted, 8 responded, of which 2 offered meetings. Only the Royal College of General Practitioners was able to offer more material. The remainder of the respondents were not aware of any studies or were undertaking internal work yet to be completed.

Titles and abstracts identified by each of the searches were read by the researcher. Papers that were easily identifiable as outside the scope and remit of this study

Table 1 Reliability terms

Term	Definition	Notes
Reliability coefficient	The relationship between true variability in score (due to variable attributes of the subjects) and true plus error variability	Lies between 0 and 1, with 1 = perfect reliability. Intra class (classical), Pearson (regression analysis), weighted or unweighted kappa (for dichotomous results), Spearman rank (for rank results) usually used.[2]
Standard error of measurement	Links standard deviation and reliability coefficient	Allows confidence intervals of true scores to be presented.
Internal consistency	Homogenicity of the test, assuming observations are fixed (all candidates take same items)or random but multiple (so variance due to observations is cancelled out)	Strongly recommended by some.[3] Several formulae – Cronbach or α coefficient, split half, Kuder-Richardson.[4,5] No consensus on level – minimum 0·80 or 0·85 usually suggested for high stakes testing.
Inter-rater reliability	Correlations between scores from different raters – observer variance.	Also referred to as inter observer or inter examiner reliability.
Intra-rater reliability	Correlations between scores from the same rater on different occasions.	
Test-retest reliability	The correlation between scores for a subject tested at different times.	Assumes the subject and construct are static between testing, which is not the aim in education, where continuous growth and development is an intrinsic aim.[6]
Alternate or parallel form	Comparison with second test with nonidentical items.[5]	
Generalizability theory	Measure of examination stability or equivalence, by allowing potential sources of variance, e.g. inter-rater, intra-rater, tiem of day, site, language, to be estimated within single study.[7,8]	Increasingly popular method for generating a generalizability coefficient (G). Also allows decision studies (D studies) to predict effect of increasing numbers of stations, raters, etc.

Table 2 Validity terms

Term	Definition	Notes
Face validity/fidelity	Does it appear to test what it is intended to test?[4] Fidelity is similar, the extent to which test conditions reflect real world circumstances.[9]	Superficial concept. Distinction needs to be made between what a test appears to measure and what it actually measures.
Content validity	The extent to which the tasks tested represent an adequate sample of the entirety of the domain to be tested. Can results be generalised?[6]	Most assessments use retrospective focus, by ensuring sampling against curriculum content. Licensure assessments require prospective focus – sampling against functions of the new role. Task analysis of the role important, especially in USA.[10]
Criterion referenced validity	Comparison with alternative measure of same construct - either contemporaneous, **concurrent validity**, or future, **predictive validity**.	Lack of suitable alternative can be problematic. Few gold standards available.
Construct validity	How accurately the test measures the unobservable qualities (the constructs) it is designed to elicit.	Requires a range of methods to 'build a case'.[5.11.12] Recent emphasis on proving that other factors are not confounders – e.g. preparation for an assessment rather than actual clinical performance leads to the best results.
Consequential validity	The effect the assessment has on learning, and the political use of test results.	A recent and increasingly important concept, bringing in ethical aspects of assessment.

Table 3 Other assessment terms

Term	Definition	Notes
Summative assessment	Assessment for a scoring, grading, pass/fail decision.	Usually at wider level than classroom, either school, college, local or national.
Formative assessment	Assessment for planning teaching and learning and for student development.	Usually at classroom level, to provide information to learner and teacher.
Competence assessment	Assessment of what the candidate can do or could do.	Poor performance may not always be due to poor competence.[13] Alternatively, competence may be assessed as satisfactory in 'ideal circumstances', but the results not transferable to real life situations.[14]
Performance assessment	Assessment of what the candidate actually does.	Some define performance assessment as only applicable to real life situations, similar to 'authentic assessment'[1]
Norm referenced	Relates an individual's performance to that of his or her peers.	Might be used if a fixed number of places are available for those who pass, for example for entry to a course, but it is not easy to defend in other circumstances. Candidates have no control over the performance of other candidates.[1]
Criterion referenced	Relates performance to predetermined criteria.	Some form of standard setting is required if the criterion is not a dichotomous present or absent observable entity.

were excluded (Fig. 1). Remaining papers were passed onto the next stage and read by two workers independently to determine suitability for inclusion.

The criteria for inclusion were:

- the paper applies to postgraduate medical education
- the assessment(s) is(are) under investigation
- the assessment or assessment strategy is being tested either as a methodology already in use in certification or licensure in postgraduate medical education or is explicitly being developed for that purpose.

The last criteria required most deliberation. Attempts to limit the included papers to only those that pertained

Table 4 Database search details

Medline 1985–2000 was searched in early and late March 2000 using the following search strategy:

#1	reliab*.mp
#2	valid*.mp
#3	feasib*.mp
#4	1 or 2 or 3
#5	assess*.mp
#6	explode educational measurement/all subheadings
#7	explode evaluation studies/all subheadings
#8	education medical/standards
#9	5 or 6 or 7 or 8
#10	4 and 9
#11	explode education, medical, graduate/all subheadings
#12	explode ‘internship and residency’/all subheadings
#13	medical staff/education
#14	explode societies, medical/all subheadings
#15	family practice/education
#16	trainee*.mp
#17	registrar*.mp
#18	house officer*.mp
#19	intern*.mp
#20	interna*.mp
#21	19 not 20
#22	residen*.mp
#23	residential*.mp
#24	residenc*.mp
#25	23 or 24
#26	22 not 25
#27	11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 21 or 26
#28	explode clinical competence/all subheadings
#29	explode professional competence/all subheadings
#30	explode competency-based education/all subheadings
#31	28 or 29 or 30
#32	10 and 27
#33	10 and 31
#34	27 and 31
#35	32 or 33 or 34
#36	limit 35 to english language
#37	limit 36 to comment
#38	limit 36 to editorial
#39	limit 36 to letter
#40	limit 36 to review
#41	37 or 38 or 39 or 40
#42	36 not 41

A further supplementary Medline search was performed in early June 2000 using additional terms.

#1	exp Specialty Boards
#2	exp Certification
#3	exp licensure/or licensure, hospital/or licensure, medical/
#4	qualif*.mp
#5	1 or 2 or 3 or 4
#6	#5 not (#42 above)

EMBASE

A similar format to the Medline one was entered, with minor modifications for variations in MeSH and keywords.

Educational Resources Information Center (ERIC) database

The ERIC database was searched in March 2000 with combinations of medicine (keyword), measurement (keyword), higher education (descriptor), graduate medical education (descriptor), competence (keyword), measurement techniques (descriptor), professional education (descriptor).

Institute for Scientific Information Citation Indexes Databases

Search terms were graduate medical education and measur*, and graduate medical education and competenc*.

Topics in Medical Education (TIME)

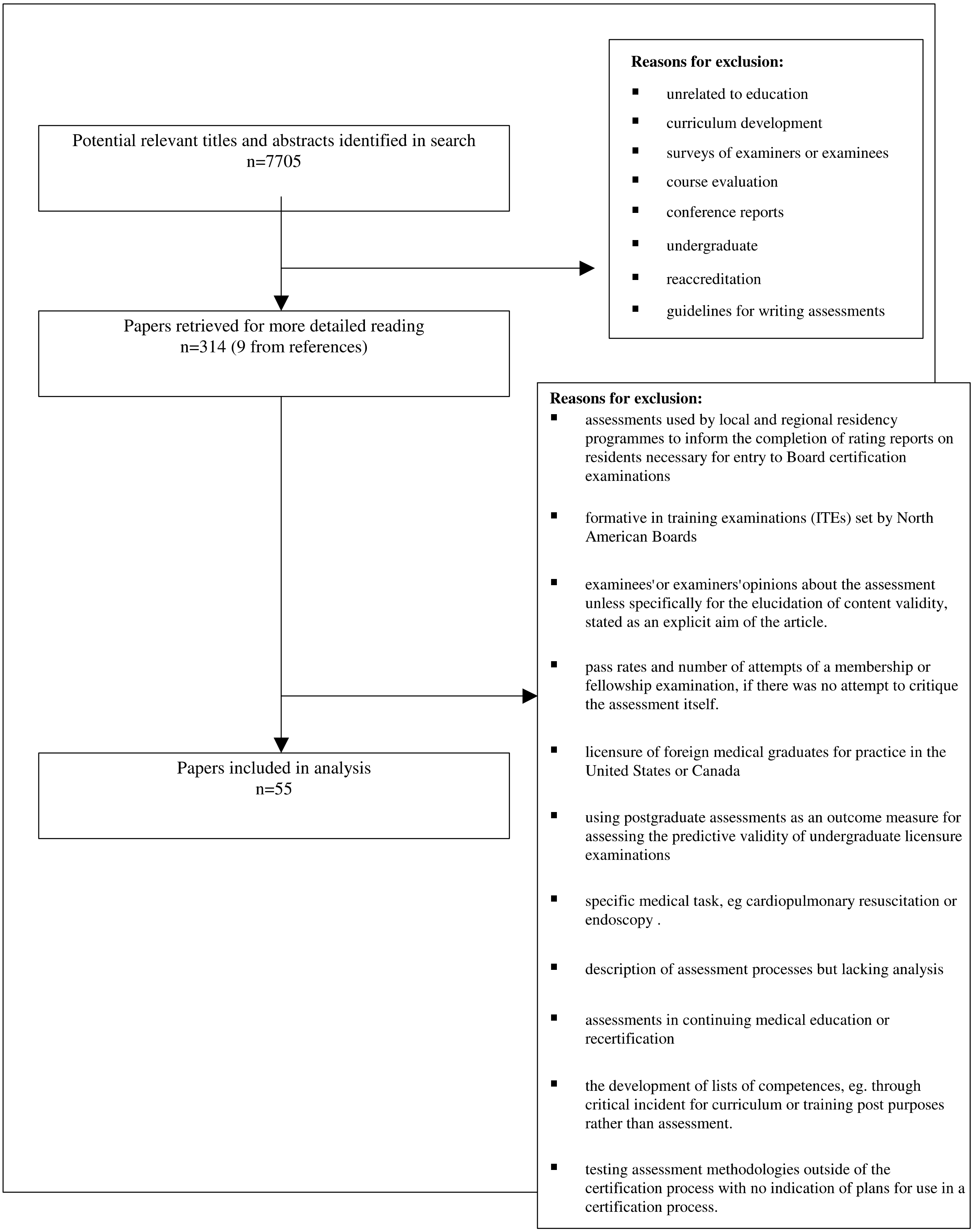
The TIME database from Dundee was accessed in late May with Keywords: competence assessment, postgraduate assessment, certification, accreditation, reliability and validity.

to certification that was essential for career progress were hampered by the quasi compulsory nature of many systems. It seemed relevant to include any assessment process that led to an award that advantaged the holder, whether for employment, finance, status, etc. Consequently, for example, the examination for Membership of the Royal College of General Practitioners (MRCGP) is included even though it is not ‘compulsory’ in the same way as summative assessment for Certificate of Completion of Vocational training (CCVT) for UK general practice (GP) trainees.

Results

Fifty-five papers analysing postgraduate medical certification processes met the criteria (Table 5). General details about country of origin, certification process involved and types of assessment methodology are given in Tables 6–10.

The papers were categorised into their main focus: candidate factors, examiner factors and examination factors, which will be discussed separately.



Potential relevant titles and abstracts identified in search
n=7705

Reasons for exclusion:

- unrelated to education
- curriculum development
- surveys of examiners or examinees
- course evaluation
- conference reports
- undergraduate
- reaccreditation
- guidelines for writing assessments

Papers retrieved for more detailed reading
n=314 (9 from references)

Reasons for exclusion:

- assessments used by local and regional residency programmes to inform the completion of rating reports on residents necessary for entry to Board certification examinations
- formative in training examinations (ITEs) set by North American Boards
- examinees'or examiners'opinions about the assessment unless specifically for the elucidation of content validity, stated as an explicit aim of the article.
- pass rates and number of attempts of a membership or fellowship examination, if there was no attempt to critique the assessment itself.
- licensure of foreign medical graduates for practice in the United States or Canada
- using postgraduate assessments as an outcome measure for assessing the predictive validity of undergraduate licensure examinations
- specific medical task, eg cardiopulmonary resuscitation or endoscopy .
- description of assessment processes but lacking analysis
- assessments in continuing medical education or recertification
- the development of lists of competences, eg. through critical incident for curriculum or training post purposes rather than assessment.
- testing assessment methodologies outside of the certification process with no indication of plans for use in a certification process.

Papers included in analysis
n=55

Figure 1 Search details.

Table 5 Included papers – summary

Ref no	Country	Specialty	Assessment/ certification	Assessment type	Subjects	Focus/foci
18	Australia	General practice	FRACGP	Whole exam	495 candidates	Construct validity
19	Australia (pilot)	General practice feasibility	FRACGP	Videotaped consultations	12 trainees/46 consultations	Inter-rater reliability
20	Australia	General practice	FRACGP	Whole exam – 8 subtests	286 candidates	Reliability and scoring through generalizability
21	Australia	General practice	FRACGP	Whole exam – 8 subtests	286 candidates	Reliability and scoring through generalizability
22	Australia	General practice	FRACGP	Whole exam – 8 subtests	79 examinees	Reliability and scoring through generalizability Internal consistency
23	Australia	Surgery	FRACS Part 1	MCQ	12 Part 1 exams (approx 220 candidates each)	Concurrent validity (within test)
24	Canada	Family physicians	CFPC	Whole exam	1324 candidates	Internal consistency Concurrent validity (between segments) Examination stability
25	Canada	Family physicians	CFPC	SAMPs	1324 candidates	Internal consistency Generalizability Examination stability
26	Canada	Family physicians	CFPC	SOO	1324 candidates	Internal consistency Generalizability Examination stability
27	Canada	Sport Medicine	Dipl Sport Medicine	OSCE	25 candidates	Internal consistency Inter-rater reliability
28	Canada	Internal medicine	RCPSC certification	Oral exam	287 candidates	Examination stability and reliability through generalizability Inter-rater reliability
29	Canada	Internal medicine	RCPSC certification	Oral exam	176 candidates	Inter-rater reliability Reliability and scoring through generalizability Examination stability
30	Canada – Quebec	Family physicians	Licensure for family practice	Whole exam – CFPC	539 over 3 exams	Examination stability Internal consistency Feasibility
31	Canada – Quebec	Family physicians	Licensure for family practice	OSCE	235 candidates	Examination stability Internal consistency Generalizability
32	Canada – Quebec	Family physicians	Licensure for family practice	OSCE	205 candidates	Construct validity
33	Canada – Quebec	Family physicians	Licensure for family practice	OSCE	13 physicians	Content validity
34	Canada – Quebec	Family physicians	Licensure for family practice	OSCE	172 candidates	Concurrent validity (between segments) Internal consistency
35	Canada – Quebec	Family physicians	Licensure for family practice	Whole exam – CFPC	614 physicians	Predictive validity
36	Israel	Family physicians	NFME Board	Oral exam	94 examiners	Examiner characteristics/inter rater reliability

Table 5 *Continued*

Ref no	Country	Specialty	Assessment/ certification	Assessment type	Subjects	Focus/foci
37	New Zealand	General practice	MRNZCGP Part 1	Whole exam – 6 parts	97 candidates	Inter-ater reliability Intra-rater reliability Concurrent validity (between segments) Examination stability
38	New Zealand	General practice	MRNZCGP Part 1	Majority of exam – 4 parts	93 candidates	Content validity Construct validity Inter-rater reliability Concurrent validity (between segments)
39	New Zealand	General practice	MRNZCGP Part 1	Simulated patients	63 candidates	Inter-rater reliability
40	New Zealand	General practice	MRNZCGP Part 1	Simulated patients	109 candidates	Construct validity (case specificity)
41	New Zealand	General practice	MRNZCGP Part 1	Simulated patients	109 candidates	Inter-rater reliability
42	New Zealand	General practice	MRNZCGP Part 1	Simulated patients	12 consumer examiners	Intra-rater reliability
43	UK	Palliative medicine	Dipl Palliative Medicine	Simulated patients	96 candidates	Inter-rater reliability
44	UK	General practice	MRCGP	SAQ paper	336 + 298 candidates	Consequential validity
45	UK (pilot)	General practice Generalizability	MRCGP	Simulated surgery	87 trainees	Reliability and scoring through
46	UK	General practice	MRCGP – Paper 1	SAQ (MEQ)	1391 candidates	Content validity
47	UK	General practice	Summative assessment (pilot/development)	Videotaped consultations	3 consultations	Inter-rater reliability Internal consistency
48	UK	General practice	Summative assessment (pilot/development)	Simulated surgery	73 course organisers	Iinstrument development (content validity)
49	UK	General practice	Summative assessment (pilot)	Simulated surgery	5 doctors, 6 assessors	Inter-rater reliability Intra-rater reliability Generalizability
50	UK	General practice	Summative assessment (pilot)	Videotaped consultations	10 trainee consultations/ 25 assessors	Concurrent validity (between segments) Feasibility Inter rater reliability
51	UK	General practice	Summative assessment (pilot)	Audit paper	104 trainees	Consequential validity Feasibility
52	UK	General practice	Summative assessment (development)	Audit paper	135 GPs and 18	Instrument development (content validity)
53	UK	General practice	Summative assessment (development)	Trainer's report	974 trainers	Instrument development (content validity)
54	UK	General practice	Summative assessment (development)	Trainers report	27 trainers	Instrument development (content validity)
55	UK	General practice	Summative assessment	Whole package	359 trainees	Concurrent validity (between segments)
56	UK	General practice	Summative assessment (development)	Written project	186 stakeholders	Instrument development (content validity)

Table 5 *Continued*

Ref no	Country	Specialty	Assessment/ certification	Assessment type	Subjects	Focus/foci
57	UK	General practice	Summative assessment (development)	Trainers report	159 trainees	Instrument development (content validity)
58	UK	General practice	Summative assessment (pilot)	Written project	20 projects	Sensitivity and specificity Feasibility
59	UK	General practice	Summative assessment (pilot)	Trainer's report	54 registrars	Inter-rater reliability Feasibility
60	UK	General practice	Summative assessment (pilot)	Audit paper	102 audit submissions	Sensitivity and specificity
61	UK	General practice	Summative assessment (pilot)	Simulated surgery	15 trainees	Test-retest reliability Construct validity
62	UK	General practice	Summative assessment	Trainers report	29 registrars	Concurrent validity (external to test)
63	USA	Anaesthetics	ABA certification (within test)	Written paper	2449 candidates	Concurrent validity
64	USA	Anaesthetics	ABA certification	Written paper	1310 residents	Concurrent validity (external to test)
65	USA	Family medicine	ABFM certification	MCQ	3 exams	Content validity
66	USA	Internal medicine	ABIM certification	Written paper	3 years of exam – 7–8000/year	Internal consistency Construct validity Concurrent validity (between segments and external to test)
67	USA	Internal medicine	ABIM certification	Written paper	185 + 74 physicians	Predictive validity
68	USA (pilot)	Internal medicine	ABIM certification	Written paper	2975 candidates	Concurrent validity (within and external to test)
69	USA	Nephrology	ABIM nephro board certification	Written paper	514 candidates	Concurrent validity (within test)
70	USA	Psychiatry	ABPN certification	oral exam/patients	1422 candidates	Inter-rater reliability
71	USA	Psychiatry	ABPN certification	oral exam/patients	1422 candidates	Examination stability Inter-rater reliability
72	USA	Psychiatry	ABPN certification (pilot)	Oral exam/live patient and AV examination	363 candidates	Internal consistency Inter-rater reliability Construct validity Instrument development

Table 6 Specialty and country

	Australia	Canada	Israel	New Zealand	UK	USA	Total
General/family practice	5	9	1	6	19	1	41
Surgery	1						1
Internal medicine		2				3	5
Sports medicine		1					1
Palliative medicine					1		1
Psychiatry						3	3
Anaesthetics						2	2
Nephrology	1	1					
Total	6	12	1	6	20	10	55

Table 7 Certification processes by country

Country	Specialty	Certification process	Number of papers
Australia	General practice (GP)	Fellowship of the Royal Australian College of General Practitioners	5 Same data set for 2
	Surgery	Fellowship of the Royal Australian College of Surgery Part 1	1
Canada	GP	Certification Examination of the College of Family Physicians of Canada	3
	GP	Quebec licensure for family practice	6 Same data set for 2
	Internal medicine	Royal College of Physicians and Surgeons of Canada certification	2
	Sports medicine	Diploma of Sports Medicine	1
Israel	GP	National Family Medicine Examination	1
New Zealand	GP	Membership of the Royal New Zealand College of General Practitioners Part 1	6
United Kingdom	GP	Certification of Completion of Vocational Training (summative assessment)	16
	GP	Membership of the Royal College of General Practitioners	3
	Palliative medicine	Diploma in Palliative Medicine	1
United States of America	Internal medicine	American Board of Internal Medicine (general internal medicine certification)	3
	Family medicine	American Board of Family Medicine certification	1
	Psychiatry	American Board of Psychiatry and Neurology certification	3
	Anaesthetics	American Board of Anesthetics certification	2
	Nephrology	American Board of Internal Medicine nephrology certification	1

Candidate factors

Comparison of candidate groups

Spike and Veitch subjected the scores of four years of the Fellowship Examination of the Royal Australian College of General Practitioners (FRACGP) to analysis against demographic and experience characteristics of the 495 candidates.[18] Five behavioural attributes (knowledge, interpretation, problem solving, affective behaviour and psychomotor skills) are assessed in the various components of the examination process. No variation in performance was found for gender. Younger age was associated with better knowledge scores. Australian graduates performed better than foreign graduates in interpretation, problem solving and affective behaviour. Family Medicine Programme status was associated with higher interpretation status but paradoxically, a decrease in interpretation scores with increasing number of years in general practice. The authors speculate whether younger candidates are more 'exam fresh' and more used to MCQ format than older candidates, and whether foreign medical graduates are less familiar with the assessment techniques used or may have language difficulties.

Thomson looked at vocational compared with non-vocationally trained trainees and their scores on the

Table 8 Certification process by specialty

Specialty	Country	Certification process	No. of papers
Anaesthetics	USA	ABA certification	2
General practice	Australia	FRACGP	5
	Canada	CFPC	3
		Quebec licensure	6
	Israel	NFME	1
	New Zealand	MRNZCGP Part 1	6
	UK	CCVT (summative assessment)	16
		MRCGP	3
	USA	ABFM	1
Internal medicine	Canada	RCPSC	2
	USA	ABIM general certification	3
		ABIM nephrology certification	1
Psychiatry and neurology	USA	ABPN certification	3
Palliative medicine	UK	Diploma in Palliative Medicine	1
Sport medicine	Canada	Diploma in Sport Medicine	1
Surgery	Australia	FRACS	1

Table 9 Type of assessment under investigation

Type of assessment	Certification process	Number of papers	Total
Whole or several parts of process	FRACGP	4	
	Certification Examination of the CFPC	1	
	Quebec licensure for family practice	2	
	MRNZCGP Part 1	2	
	UK GP CCVT (summative assessment)	1	10
Multiple choice question papers	FRACS Part 1	1	
	ABFM certification	1	2
Short answer questions	MRCGP (UK)	2	2
Short-answer management problems	Certification Examination of the CFPC	1	1
Written paper – various question types	ABA certification	2	
	ABIM certification	3	
	ABIM nephrology certification	1	6
Oral (using real patients)	RCPSC certification	2	
	NFME Board	1	
	ABPN certification	3	6
Observed structured clinical examination (OSCE)	Quebec licensure for family practice	4	
	Diploma of Sports Medicine (Canada)	1	5
Simulated surgery/patients	Certification Examination of the CFPC	1	
	MRNZCGP Part 1	4	
	UK GP CCVT	3	
	MRCGP (UK)	1	
	Diploma in Palliative Medicine (UK)	1	10
Videotaped consultations	FRACGP	1	
	UK GP CCVT	2	3
Audit paper/written project	UK GP CCVT	5	5
Trainer's report	UK GP CCVT	5	5
			55

Membership examination for the Royal New Zealand College of General Practitioners (MRNZCGP) Part 1.[38] The former had better mean scores overall (72·1% vs. 67·3% $t = 2{\cdot}17$, $P = 0{\cdot}03$) and in the management interview component. The author makes the point that this small difference may reflect practice with the format rather than improved competence.

Table 10 Validation foci

Focus	Examination process	Number of papers
Inter-rater reliability	FRACGP[19] Diploma of Sport Medicine (Canada)[27] Canada RCPSC[28,29] Israel NFME[36] MRNZCGP[37–39,41] Diploma in Palliative Medicine (UK)[43] UK GP CCVT[47,49,50,59] ABPN[70–72]	17
Internal consistency	FRACGP[22] CFCP[24–26] Diploma of Sport Medicine (Canada)[27] Quebec licensure for family practice[30,31,34] UK GP CCVT[47,55] ABIM[66] ABPN[72]	12
Examination stability	FRACGP[21,22] CFPC[24,26] RCPSC[28,29] Quebec Licensure for Family Practice[30,31] MRNZCGP[37] UK GP CCVT[61] ABPN[71]	11
Scoring systems	FRACGP[20–22] RCPSC[29] MRCGP[45] UK GP CCVT[48,49,58,60,61]	10
Instrument development	FRACGP[19] UK GP CCVT[47,48,52–54,56,57] ABPN[72]	9
Content validity	Quebec licensure for practice[33] MRNZCGP[37] MRCGP[46] UK GP CCVT[48,57] ABFM[65]	6
Candidate factors (construct validity)	FRACGP[18] Quebec licensure for family practice[32] MRNZCGP[38] UK GP CCVT[61] ABIM[66] ABPN[72]	6
Concurrent validity – between instruments	FRACGP[22] CFPC[24] Quebec licensure for family practice[34] MRNZCGP[38] UK GP CCVT[55] ABIM[66]	6
Concurrent validity – within instruments	FRACS[23] ABA[63] ABIM[68] ABIM (nephrology)[69]	4
Concurrent validity – external	UK GP CCVT[62] ABA[64] ABIM[66,68]	4

Table 10 *Continued*

Focus	Examination process	Number of papers
Intra-rater reliability	MRNZCGP[37,42] UK GP CCVT[49]	3
Case specificity	MRNZCGP[39,40]	2
Consequential validity	MRCGP[44] UK GP CCVT[51]	2
Predictive validity	Quebec licensure for family practice[35] ABIM[67]	2

Licensing examinations for family physicians in Quebec are run in English or French.[32] The authors investigated whether candidates' first language affected scores in the objective structured clinical examination (OSCE) sessions. Using generalizability theory they compared causes of variance and found that variance by site of examination without other factors (cases or persons) was almost nil, confirming that sites are equivalent whether run in English or in French.

In the American Board of Psychiatry and Neurology (ABPN) Part 2,[72] a comparison of passing and failing candidates through regression analysis of items on an 18 item scale during live patient examination revealed that treatment plan, informational cues, mental state examination and control of interview were most influential.

For UK general practice summative assessment, the Leicester assessment package developers claim construct validity as it was tested on first, second and third year registrars.[61] However the numbers are small and no analysis was performed on the descriptive results.

Norcini *et al.* used extreme groups methodology[2] in comparing two subsets of candidates for the 1980–82 American Board of Internal Medicine certifying examination in general internal medicine.[66] The 1448 candidates in the high criterion group were US graduates, had attended high quality residency programmes and received superior programme director ratings. The low criterion group (1803 candidates) consisted of foreign medical graduates who were not at selected residency programmes and received no more than satisfactory ratings. All components of the examination, the three different sorts of multiple choice questions and the patient management problems, could discriminate between the groups but composite scores, not surprisingly, were most discriminatory.

Effect on candidate behaviour

Two papers from UK general practice raise questions about consequential validity. Lough *et al.* asked trainees in the West of Scotland about the impact of 2 years of compulsory audit project submission.[51] The majority of the trainees reported that the assessed audit project was their first experience of audit. Undertaking the task had increased their confidence.

The second paper looked at the impact of the Critical Reading Question paper (CRQ) in the Membership examination of the Royal College of General Practitioners (MRCGP) on the study habits of trainees.[44] The CRQ paper has now been incorporated into the structured short answer paper, MRCGP Part 1. In consecutive years before and after the introduction of the CRQ, a questionnaire on learner behaviour associated with the MRCGP as a whole were sent to those who had registered for the examination. In both years routine practice work was more important to candidates than conferences, lectures, textbooks or videos. In the second year there was a statistically significant reduction in use of undergraduate texts and tabloid medical press, and an increase in reading the BMJ and British Journal of General Practice.

Examiner factors

Inter-rater reliability

The ABPN uses real patients in the oral examination for Part 2 of their specialty board certification. The candidates interview a patient and are then questioned by one pair of examiners. After viewing a videotape of another consultation they are questioned by a different pair of examiners. Analysis of the results of 1422 candidates (2844 individual examinations) revealed only fair to good association between pairs of examiners measured by weighted kappa (0·54–0·56).[70] More explicit grading criteria and extensive rater training are suggested remedies. In a later paper about a trial of different grading systems, inter rater intraclass correlation coefficients ranged from 0·37 to 0·77.[72]

Weingarten *et al.* took a different approach for the National Family Medicine Examination (NFME) oral examination in Israel, relating examiner characteristics to their marking tendencies.[36] From a large sample of 94 examiners over 5328 examination sessions correlation between 'toughness' and academic rank (Fisher exact 1-tailed test, $P = 0{\cdot}048$) was found. Specifically they found associations between failure rates in the clinical exam and length of experience as an examiner ($X^2 = 4{\cdot}68$, $P = 0{\cdot}03$), between failure rate overall and academic rank above lecturer ($X^2 = 8{\cdot}86$, $P = 0{\cdot}003$) and qualification in an English speaking country ($X^2 = 5{\cdot}08$, $P = 0{\cdot}024$).

After introduction of consumer observers to the MRNZCGP examination, one consumer and one medical examiner independently scored two 15 minute role play consultations on the same proforma.[41] Comparison of scores revealed no differences between the means of overall scores by the two groups but correlations (Pearson product-moment) were between 0·4 and 0·52. Agreement on pass/fail status was present in 92·2% of interviews (total 218 interviews) but the kappa coefficient was only 0·15. Where disagreements occurred, there were significant differences between scoring for several of the communication attributes.

Finlay and colleagues' study on actor patient and examiner ratings in a simulated consultation skills examination for the Diploma in Palliative Medicine (University of Wales) used yet another approach, Bland and Altman's 'limits to agreement'.[43,73] Large differences between scores given by actor patients and those given by examiners were demonstrated. The examination uses only one simulated consultation and the authors accept the lack of generalizability.

Absence of training and low numbers of raters was considered influential in the low reliability coefficients in an Australian FRACGP study.[19] Four examiners independently marked three real consultations videotaped by 13 trainees. Inter rater reliability was low to moderate, with correlation coefficients of scores 0·32–0·65 (Pearson), and correlation of rank order 0·42–0·61 (Spearman).

Inter-rater reliability is mentioned in the Canadian Diploma of Sport Medicine paper,[27] the Canadian oral certification examination for Internal Medicine of the Royal College of Physicians and Surgeons of Canada[29] and the pilot tests of two schedules for consultation scoring for UK GP summative assessment[47,49] and for the trainer's report.[59]

Intra-rater reliability

Intra-rater reliability (the correlation of results from same markers at different times) was investigated in two papers from New Zealand general practice.[37,42]

Examination factors

Instrument development

Several papers from UK general practice described in detail the processes of instrument development. Not all of the instruments are in current use in summative assessment although some have been accepted for inclusion in 2000–2001 (personal communication, Roger Neighbour). The papers described development of marking schedules for audit projects,[52] written projects,[56] videotaped or live (real or simulated) consultations[47,48,50] and trainer's report.[53,54,57]

The authors of the papers in 1991 and 1993 on the American Board of Psychiatry and Neurology certification examinations recommended the use of more explicit grading criteria.[70,71] These were subsequently developed and piloted.[72] The pilot focuses on the specific competencies that differentiated passing from failing candidates in order to validate the instrument.

Two instruments to rate videotaped consultations were compared in an Australian study.[19] Both had similar reliability coefficients and Pearson correlation coefficients between them were 0·88–0·93, with the higher levels for overall scores as compared to individual components of each rating form.

Content validity of instruments

The consensus development techniques in the above papers were used to justify both face and content validity. Other papers also used a range of techniques to justify content validity. Development process using experts and consensus[24,27,30,68] and matrices or blueprints[22,27,68] were among those described. Those papers that were specifically designed to test content validity used expert panels,[38,48] examiner and candidate surveys,[38] and factor analysis of scores.[46]

The Quebec OSCE group used recently certified practising family physicians[33] and the UK Royal College of General Practitioners used recent vocational trainees to canvass opinion about the content of their respective assessment methodologies.[57]

Re-categorisation of all questions against a new classification system was undertaken by the American Board of Family Practice.[65] The paper describes the process but gives no outcome data and minimal recommendations.

A detailed factor analysis was recently performed on the MRCGP Part 1 examinees' response.[46] Content variation between two sittings of the examination was uncovered despite the use of a blueprint for setting questions.

Scoring system and process

Two papers piloted proposed assessment instruments in order to develop referral processes. Sensitivity and specificity were compared to decide the optimum number of markers and requirements for referral.[58,60] Several papers used generalizability theory to analyse their assessment methodologies and calculate optimum numbers of assessors and cases or scoring systems.[45,49,61]

Three papers describe the same data set on three years of subtest results of the FRACGP examination.[20–22] The reproducibility coefficients were low to moderate for test scores, moderate to high for pass/fail decisions. This is in contrast to Turnbull *et al.* who postulated and proved lower reliabilities when moving from continuous to dichotomous calculations.[29]

Internal consistency

Many of the papers cite coefficient α or standard error of measurement (SEM) for component parts of their assessment processes. These are summarised in Fig. 2. Papers citing high levels of internal consistency for components or whole examinations include oral examinations with patients and videotaped simulated and real consultations. This is contrary to the general impression that good reliability is only achieved with multiple choice questions or OSCEs.

Examination stability

One paper from UK general practice specifically looked at test-retest reliability by comparing candidate performance and pass/fail decisions on the same format examination held 4 weeks apart.[61]

A number of papers used analysis of variance and generalizability theory to investigated variations in scores across site, case difficulty, day of examination and time of day.[21,22,28,29,37,71]

The Royal College of Physicians and Surgeons of Canada (RCPSC) internal medicine certification examination long case component showed a minimal variance for case difficulty as rated by examiners.[28] Short case, by contrast, showed no variance for cases, but more marked variance for day of the examination. The first day had the lowest reliability coefficients of 0·73 (other days 0·84–0·87). In another study, unacceptably low correlations between sessions for decision (0·57–0·69 for sites, 0·30–0·47 for morning vs. afternoon) was detected.[29] The authors suggest that examiner training was not deficient as there were higher inter rater reliabilities; limited sampling was more likely.

Analysis of the certification examination of the College of Family Physicians of Canada (CFPC) revealed a difference between three of the 9 sites in the 1993 examinations.[24,25] Brown *et al.* found no difference across sites for the simulated office oral component.[26]

In Quebec successful performance in an OSCE component, additional to the other CFPC examinations, is required prior to licensure for family practice.

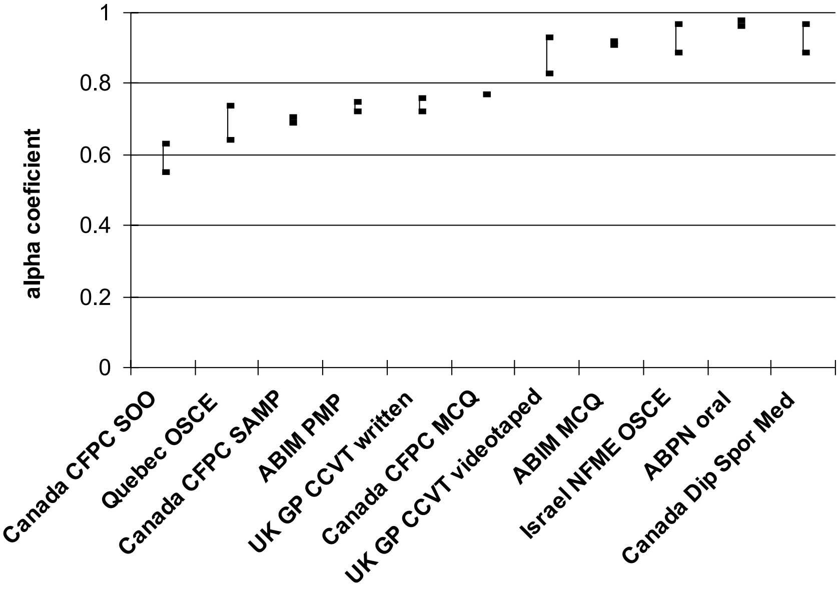


Figure 2 Alpha coefficient summary.

A variety of statistical techniques tested examination stability across sites, homogenicity of cases and the equivalence of parallel tracks.[30,31] The authors were satisfied that the logistical difficulties of running a large scale OSCE were surmountable and the test results reliable.

Case specificity

Thomson compared two different scoring mechanisms to reduce the case specificity of the clinical components of MRNZCGP Part 1 examination but found no difference.[39] He concluded that differences are 'real' differences in performance that cannot be manipulated by scoring method. He hypothesised various strategies to keep the number of sampling cases low. One of these is to remove the variance due to context by giving relevant knowledge to the candidate in advance. However, when tested, no correlation was found between performances on the cases suggesting that even generic communication skills remain context specific.[40] This adds to his premise that more than two simulated consultations are required for a high stakes assessment.

Concurrent validity

Within instrument

Two of the included papers, both from the USA Specialty Boards, one nephrology,[69] the other anaesthetics,[63] investigated the impact of separately scoring subsets of written questions. The authors identified a subset of questions with answers that could, if selected incorrectly, be potentially dangerous to patients. They then compared scores on the subset to composite scores, to find out whether any candidates who passed had 'failed' the subset. In the American Board of Internal Medicine's Subspecialty Board of Nephrology, the 6·6% of certified candidates who 'failed' the subset had no demographic differences to those who passed.[69] Further analysis of subset to composite score corrected for low reliabilities revealed that the subset did not measure unique properties. The American Board of Anaesthesiology (ABA) identified 29 multiple choice questions with potentially dangerous answers out of 175 in the 1983 certifying examination.[63] The 1036 candidates who passed selected a mean of 1·6 (SD = 0·3, range 0–7, no median given) dangerous answers, while the 1413 who failed selected a mean of 3·4 (SD = 0·4 range 0–10, no median given). Further analysis of those from both groups who selected four or more dangerous answers indicates that passing candidates had a higher ratio of dangerous answers to other incorrect answers, suggesting that guessing was a factor. It was concluded that separate scoring of a subset of dangerous answers is unnecessary.

The ABIM general internal medicine paper underwent a similar process but by adding a core subset.[68] The certifying examination used norm referencing and organisers had removed poorly discriminating questions (those all candidates should answer correctly) in order to maximise the spread of scores. Concerns about the resultant skewed content of the examination and the possibility that candidates could pass without mastery of important core components of internal medicine led to the development of a supplementary core examination. During a pilot application of the test to 2975 candidates, correlation with the main test corrected for unreliability was high at 0·91. 1% failed the core but passed the certifying examination.

Buzzard and Bandaranayake were thorough in investigating the relative difficulty of visual trigger material to verbal questions in the Fellowship examination of the Royal Australasian College of Surgeons.[23] The study demonstrated that visual questions were neither more difficult nor more discriminating than verbal ones testing equivalent content but could be justified on face validity.

Between parts of assessment package

The Quebec OSCE was compared with the parallel CFPC (multiple choice questions, short answer management problems and simulated office orals) examination.[34] Despite all parts being designed to test similar constructs, low correlations were found. Scores in components of the CFPC when compared to each other also did not show high correlations.[24] The authors take the view that this indicates the components measure complementary elements of competence.

A similar analysis of the six component parts of the MRNZCGP Part 1 examination revealed poor correlations even between two parts of the same component.[38] Thomson is more critical than the Canadian authors above and relates the low correlations to lack of adequate sampling in most of the component parts.

Norcini *et al.* were concerned that MCQs and PMPs had previously shown low correlations suggesting that each measured a different construct of clinical competence in the ABIM general internal medicine certifying examination.[66] However their extensive study demonstrated high correlations between both over three years of the ABIM certifying examination in the early 1980s with over 14 000 candidates. Regression analysis demonstrated that MCQs contributed a small unique variance component and PMPs the smallest unique contribution.

They extrapolate that MCQs are more valid, reliable and efficient than PMPs.

In the more detailed of the papers on FRACGP,[22] analysis of variance was used to identify the subsets that contributed most to the variance and from that, suggestions for improvements to reliability by concentrating on those subsets. The management interview (two role played consultations) contributed most to the variance.

The West of Scotland pilot for summative assessment of general practice trainees in the UK ran from 1993 to 1995.[55] The written papers showed low α coefficients relative to the other components. The different components identified different problems.

External to assessment

Norcini *et al.* used programme director's ratings as a surrogate for clinical competence, and analysed two extreme subsets of candidates for the ABIM general internal medicine certifying examination (see above).[66] Langdon *et al.* also used programme director ratings.[68] However, candidates with low ratings, 3 or under are not eligible to sit the examination, and several papers from residency programmes throughout the USA have found programme directors' ratings to be the least reliable of a range of measures used. It is questionable therefore whether they are a suitable criterion against which to compare other methodologies.

Even less impressive was the use of programme directors to determine which of three procedures they would allow their trainees to perform on themselves.[64] The ethical issues of this design (participants were unaware of the study intent) and of nonaction when programme directors identified nearly 100 doctors who they were sufficiently concerned about that they would not allow them to anaesthetise themselves for any of the operations, are worrying. The claim that it is evidence for the sensitivity of the ABA certifying examination in identifying competence in anaesthetic practice not just knowledge base is no more than speculative.

Kelly, Campbell and Murray used a better methodology for investigating concurrent validity of the trainer's report for summative assessment of UK general practice registrars by testing 29 registrars on a short practical skills workshop.[62] They found several registrars failed items even if these had been stated as satisfactory in their trainer's report.

Predictive validity

Two papers described attempts to elicit evidence of predictive validity. A Quebec study used the opportunity of a single universal health insurance plan and sophisticated encryption system to allow anonymity and inclusion of all physicians in the province.[35] A variety of data about practices, activities, quality of care (e.g. inappropriate prescribing and mammography screening rate) were compared against licensure results. The results of analysis of the 614 physicians showed correlation between diagnostic scores in the licensure examination and higher referral for consultation rates, higher disease specific (as opposed to symptom relief only) in elderly care, and less inappropriate prescribing. Unfortunately the time of follow up was only 18 months after entering practice and clearly, those who failed the licensure examination would not be part of the study.

The other study of predictive validity did have the benefit of the certified and the noncertified.[67] In this instance it was internists in 6 states in the USA who responded to an invitation to undergo a Medical Knowledge Self-Assessment Program. Of the 1476 physicians eligible, 392 agreed to take part and 259 were eventually included. Analysis of volunteers against nonvolunteers showed little difference in ABIM scores but no other characteristics were given. The written examination was a multiple choice question paper, making the outcome measure identical to the examination under study. However, other outcome measures, a patient questionnaire, ratings by professional associates and review of a sample of medical records and practice characteristics were collected by a research assistant.

A wide range of associations and non associations are given but, in brief, ABIM certification was found to be predictive of performance in the written examination and professional rating scores but not parameters of patient care. The ratings from other professionals may have been biased by other factors, but the authors were careful to examine these possibilities and exclude them after regression analyses. Interestingly, no difference between patient satisfaction scores were found between the two groups.

Other issues

Feasibility[50,51,58,59] and cost[19,30] were mentioned in several papers. Efficiency of MCQs and PMPs were analysed in detail in a paper on the ABIM certifying examinations.[66]

Discussion

Despite the systematic approach to identification of papers, omissions are likely. Non-English papers were excluded by design; documents internal to institutions and expert contact were not sought. Papers not coded by the terms used in the search strategy would have been inadvertently missed and the many variations of assessment terminology increase this possibility. Educational research papers are underrepresented in medical journals generally,[74] and negative results may be excluded from publication either through publication bias or through sensitivity on the part of the certifying body.

No attempt was made to exclude papers of poor quality. Meta-analysis techniques were not performed as it would be misleading to draw definitive conclusions from the diversity of settings and heterogenicity of assessment methodologies.

Inter-rater reliability and internal consistency are the commonest foci for validation. There is some more recent use of generalizability theory to test examination stability. It is notable that the emphasis has been on these quantifiable aspects of reliability, ones that originated in the psychometrics movement earlier last century. Their suitability when applied to assessments of complex performance and competence is starting to be questioned in the education literature.[1,11,12] It is also unlikely that inter laboratory or operator reliability of a clinical diagnostic test would be studied before the initial work on determining whether the test is the right test in the first place. Construct validity is central to test validation,[1,11,75] but there is little evidence of extensive work into this aspect.

Some papers claimed construct validity by demonstrating the expected increase in scores with increasing experience despite the flaws in this method.[2] Only the Australian study of candidate characteristics[18] and the Canadian study of the impact of language[32] question whether other factors could be confounders to performance in assessment. A more complete understanding and search for proof and disproof of underlying constructs in medical competence and performance assessment seems imperative.

Although a number of different specialties are represented, general or family practice predominates (Tables 6 and 8). The UK RCGP has demonstrated that it is prepared to be thorough in assessment instrument development and pilot testing, providing other institutions with an example to follow.

The under-representation of hospital specialties is striking. Apart from the MRCGP, no papers were found for United Kingdom membership or fellowship examinations for the Royal Colleges. Summative assessment procedures for award of Certificate of Completion of Specialist Training have been in operation since 1996, but no papers studying process or outcome were identified. Concern about legal challenges and traditional closed societal values may have restricted the amount of material in the general public domain. It will be interesting to repeat the search strategy in several years time to determine whether the new culture of transparency creates an impetus for more published work on assessment validation.

The papers in this study demonstrate that good practice in test development and implementation is present in medicine but there is insufficient evidence to support the validity and reliability of any single assessment process. It is encouraging that some institutions are demonstrating that they are actively looking at their assessments. Others may have undertaken investigations and consequently amended their assessments but without seeking publication. However the total absence of any sign of willingness for external scrutiny from many of the institutions that have a powerful and unopposed role in the career paths of doctors in training is a major concern.

Acknowledgements

The authors would like to thank Elizabeth Limb, Statistician in Public Health Sciences, St George's Hospital Medical School, for statistical advice.

Contributors

LH had the original idea, performed the search, analysed the data, wrote the paper and is guarantor. PA assisted in reading the papers to validate inclusion decisions. TH advised on the project throughout.

Funding

There was no external funding for this study.

References

1 Gipps CV. *Beyond Testing: Towards a Theory of Educational Assessment.* London: Falmer Press; 1994.

2 Streiner DL, Norman GR. *Health Measurement Scales. A Practical Guide to Their Development and Use*, 2nd edn. Oxford: Oxford University; Press 1995.

3 Norcini J. What should we do about unreliable scores? *Med Educ* 2000;**34**:501–2.

4 Tuckman BW. *Conducting Educational Research*, 3rd edn. Orlando, Florida: Harcourt Brace Jovanovich Inc; 1988.

5 Aiken LR. *Psychological Testing and Assessments*, 7th edn. Needham Heights, MA: Allyn and Brown; 1991.

6 Benett Y. The validity and reliability of assessments and self-assessments of work-based learning. In. Murphy P, eds. *Learners, Learning and Assessment.* London: OU & Paul Chapman; 1999.

7 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioural Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley; 1972.

8 Brennan RL. Generalizability theory. *Educational Measurement: Issues Prac* 1992;**11**:27–34.

9 Mulholland H. 'De facto' assessment: issues of validity and reliability. *Current Obstetrics Gynaecol* 1997;**7**:145–8.

10 American Educational Research Association American Psychological Association. *National Council on Measurement in Education, Standards for educational and psychological testing.* Washington DC: American Psychological Association; 1985.

11 Messick S. The psychology of educational measurement. *J Educational Measurement* 1984;**21**:215–37.

12 Kane MT. Validating interpretive arguments for licensure and certification examinations. *Evaluation Health Professional* 1994;**17**:133–59.

13 Rethens J-J, Sturmans F, Drop R, van der Vleuten C, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ* 1991;**303**:1377–80.

14 Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. *Educational Reser* 1995;**24**:35.

15 Glass GV. Primary, secondary, and meta-analysis of research. *Educational Reser* 1976;**5**:3–8.

16 Mulrow CD. The medical review article: state of the science. *Ann Intern Med* 1987;**106**:485–8.

17 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *Can Med Assoc J* 1988;**138**:697–703.

18 Spike NA, Veitch PC. Analysis of the RACGP Fellowship examination. *Aust Family Physician* 1990;**19**:767–75.

19 Hays RB, Jones BF, Adkins PB, McKain PJ. Analysis of videotaped consultations to certify competence. *Med J Australia* 1990;**152**:609–11.

20 Hays R, Fabb WE, van der Vleuten C. An analysis of the FRACGP exam. *Aust Family Physician* 1994;**23**:2147–9.

21 Hays RB, van der Vleuten C, Fabb WE, Spike NA. Longitudinal reliability of the Royal Australian College of General Practitioners certification examination. *Med Education* 1995;**29**:317–21.

22 Hays R, Fabb WE, van der Vleuten CP. Reliability of the Fellowship examination of the Royal Australian College of General Practitioners. *Teaching Learning Med* 1995;**7**:43–50.

23 Buzzard AJ, Bandaranayake RC. Comparison of the performance of visual and verbal multiple-choice questions. *Australia NZ J Surgery* 1991;**61**:614–8.

24 Handfield-Jones R, Brown JB, Rainsberry P, Brailovsky CA. Certification examination of the College of Family Physicians of Canada. Part 2: Conduct and general performance. *Can Family Physician* 1996;**42**:1188–95.

25 Handfield-Jones R, Brown JB, Biehn J, Rainsberry P, Brailovsky CA. Certification examination of the College of Family Physicians of Canada. Part 3: Short-answer management problems. *Can Family Physician* 1996;**42**:1353–61.

26 Brown JB, Handfield-Jones R, Rainsberry P, Brailovsky CA. Certification examination of the College of Family Physicians of Canada. Part 4: Simulated office orals. *Can Family Physician* 1996;**42**:1539–48.

27 Mohtadi NGH, Harasym PH, Pipe AL, Strother RT, Mah AF. Using an objective structured clinical examination to evaluate competency in sport medicine. *Clin J Sport Med* 1995;**5**:82–5.

28 McLean LD, Dauphinee WD, Rotman A. The oral examination in internal medicine of the Royal College of Physicians and Surgeons of Canada: a reliability analysis. *Ann Royal College Physicians Surgeons Can* 1988;**21**:510–4.

29 Turnbull J, Danoff D, Norman G. Content specificity and oral certification examinations. *Med Educ* 1996;**30**:56–9.

30 Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective, structured clinical examination for licensing family physicians. *Can Med Assoc J* 1992;**146**:1735–40.

31 Brailovsky CA, Grand'Maison P, Lescop J. A large-scale multicenter objective structured clinical examination for licensure. *Academic Med* 1992;**67**:S37–S39.

32 Marshall KG, Brailovsky CA, Grand'Maison P. French-English, English-French translation process of an objective structured clinical examination (OSCE) used for licensing family physicians in Quebec. *Teaching Learning Med* 1995;**7**:115–20.

33 Grand'Maison, p. Brailovsky CA, Lescop J. Content validity of the Quebec licensing examination (OSCE). *Can Family Physician* 1996;**42**:254–9.

34 Grand'Maison P, Brailovsky CA, Lescop J, Rainsberry P. Using standardized patients in licensing/certification examinations: comparison of two tests in Canada. *Family Med* 1997;**29**:27–32.

35 Tamblyn R, Abrahamowicz M, Brailovsky C, Grand'Maison P, Lescop J, Norcini J *et al.* Association between licensing examination scores and resource use and quality of care in primary care practice. *JAMA* 1998;**280**:989–96.

36 Weingarten MA, Polliack MR, Tabenkin H, Kahan E. Variations among examiners in family medicine residency board oral examinations. *Med Educ* 2000;**34**:13–7.

37 Thomson AN. An assessment of a postgraduate examination of competence in general practice: part 1 – reliability. *NZ Med J* 1990;**103**:182–4.

38 Thomson AN. An assessment of a postgraduate examination of competence in general practice: part 2 – validity. *NZ Med J* 1990;**103**:217–9.

39 Thomson AN. Case specificity of performance with simulated patients. *NZ Med J* 1990;**103**:372–4.

40 Thomson AN. Can communication skills be assessed independently of their context? *Med Educ* 1992;**26**:364–7.

41 Thomson AN. Consumer assessment of interview skills in a family practice certification examination. *Family Med* 1993;**25**:41–4.

42 Thomson AN. Reliability of consumer assessment of communication skills in a postgraduate family practice examination. *Med Educ* 1994;**28**:146–50.

43 Finlay IG, Stott NCH, Kinnersley P. The assessment of communication skills in palliative medicine: a comparison of the scores of examiners and simulated patients. *Med Educ* 1995;**29**:424–9.

44 Wakeford R, Southgate L. Postgraduate medical education: modifying trainees' study approaches by changing the examination. *Teaching Learning Med* 1992;**4**:210–3.

45 Bingham L, Burrows P, Caird R, Holsgrove G, Jackson N. Simulated surgery – using standardized patients to assess the clinical competence of GP registrars – a potential clinical component for the MRCGP examination. *Education for General Prac* 1996;7:102–11.

46 Munro N, Rughani A, Foulkes J, Wilson A, Neighbour R. Assessing validity in written tests of general practice – exploration by factor analysis of candidate response patterns to Paper 1 of the MRCGP examination. *Med Educ* 2000;**34**: 35–41.

47 Cox J, Mulholland H. An instrument for assessment of videotapes of general practitioners' performance. *BMJ* 1993;**306**:1043–6.

48 Fraser R, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritized criteria in the Leicester assessment package. *Br J General Prac* 1994;**44**:109–13.

49 Fraser R, McKinley RK, Mulholland H. Consultation competence in general practice: testing the reliability of the Leicester assessment package. *Br J General Prac* 1994;**44**: 293–6.

50 Campbell LM, Howie JGR, Murray TS. Use of videotaped consultations in summative assessment of trainees in general practice. *Br J General Prac* 1995;**45**:137–41.

51 Lough JRM, McKay J, Murray TS. Audit and summative assessment: two years' pilot experience. *Med Education* 1995;**29**:101–3.

52 Lough JRM, McKay J, Murray TS. Audit and summative assessment: a criterion-referenced marking schedule. *Br J General Prac* 1995;**45**:607–9.

53 Johnson N, Hasler J, Toby J, Grant J. Content of a trainer's report for summative assessment in general practice: views of trainers. *Br J General Prac* 1996;**46**:135–9.

54 Johnson N, Hasler J, Toby J, Grant J. Consensus minimum standards for use in a trainer's report for summative assessment in general practice. *Br J General Prac* 1996;**46**:140–4.

55 Campbell LM, Murray TS. Summative assessment of vocational trainees: results of a 3-year study. *Br J General Prac* 1996;**46**:411–4.

56 Evans A, Singleton C, Nolan P, Hall W. Summative assessment of general practice registrars' projects: deciding on criteria and developing a marking schedule. *Education for General Prac* 1996;7:229–36.

57 Johnson N, Hasler J. Content validity of a trainer's report: summative assessment in general practice. *Med Education* 1997;**31**:287–92.

58 Evans A, Nolan P, Bogle S, Hall W, Bahrami J. Summative assessment of general practice registrars' projects: reliability of the Yorkshire schedule. *Education for General Prac* 1997;**8**:40–7.

59 Johnson N, Hasler J, Toby J, Grant J. Pilot testing of a structured trainer's report for summative assessment in general practice. *Education for General Prac* 1997;**8**:308–15.

60 Lough JRM, McKay J, Murray TS. Audit and summative assessment. system development and testing. *Med Education* 1997;**31**:219–24.

61 Allen J, Evans A, Foulkes J, French A. Simulated surgery in the summative assessment of general practice training. results of a trial in the Trent and Yorkshire regions. *Br J General Prac* 1998;**48**:1219–23.

62 Kelly MH, Campbell LM, Murray TS. Clinical skills assessment. *Br J General Prac* 1999;**49**:447–50.

63 Slogoff S, Hughes FP. Validity of scoring 'dangerous answers' on a written certification examination. *J Med Education* 1987;**62**:625–31.

64 Slogoff S, Hughes FP, Hug CC, Longnecker DE, Saidman LJ. A demonstration of validity for certification by the American Board of Anesthesiology. *Academic Med* 1994;**69**:740–6.

65 Pisacano NJ, Veloski JJ, Brucker PC, Gonnella JS. Defining the content of Board certification examinations. *Proc Annu Conf Res Med Education* 1986;**25**:205–10.

66 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Med Education* 1985;**19**:238–47.

67 Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med* 1989;**110**:719–26.

68 Langdon LO, Grosso LJ, Day SC, Norcini JJ, Kimball HR, Fletcher SW. A core component of the certification examination in internal medicine. *J General Intern Med* 1993;**8**:497–501.

69 Webster GD, Goldfarb S, Norcini JJ, Shea JA, Murray LN. Performance of a dangerous answer subset within a subspecialty certifying examination. *Med Education* 1987;**21**:426–31.

70 McDermott JF, Tanguay PE, Scheiber SC, Juul D, Shore JH, Tucker GJ *et al.* Reliability of the Part II board certification examination in psychiatry: interexaminer consistency. *Am J Psychiatry* 1991;**148**:1672–4.

71 McDermott JF, Tanguay PE, Scheiber SC, Juul D, Shore JH, Tucker GJ *et al.* Reliability of the Part II board certification examination in psychiatry: examination stability. *Am J Psychiatry* 1993;**150**:1077–80.

72 McDermott JF, Streltzer J, Lum KY, Nordquist CR, Danko G. Pilot study of explicit grading criteria in the American Board of Psychiatry and Neurology Part II examination. *Am J Psychiatry* 1996;**153**:1097–9.

73 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.

74 Buckley G. Partial truths: research papers in medical education. *Med Education* 1998;**32**:1–2.

75 Moss PA. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Rev Educational Res* 1992;**62**:229–58.

Received 4 *May* 2001; *editorial comments to authors* 9 *July* 2001; *accepted for publication* 9 *July* 2001