BMC Medical Education

RESEARCH

Open Access

Medical specialty certification exams studied according to the Ottawa Quality Criteria: a systematic review

Daniel Staudenmann[1*], Noemi Waldner[1], Andrea Lörwald[1†] and Sören Huwendiek[1†]

Abstract

Background Medical specialty certification exams are high-stakes summative assessments used to determine which doctors have the necessary skills, knowledge, and attitudes to treat patients independently. Such exams are crucial for patient safety, candidates' career progression and accountability to the public, yet vary significantly among medical specialties and countries. It is therefore of paramount importance that the quality of specialty certification exams is studied in the scientific literature.

Methods In this systematic literature review we used the PICOS framework and searched for papers concerning medical specialty certification exams published in English between 2000 and 2020 in seven databases using a diverse set of search term variations. Papers were screened by two researchers independently and scored regarding their methodological quality and relevance to this review. Finally, they were categorized by country, medical specialty and the following seven Ottawa Criteria of good assessment: validity, reliability, equivalence, feasibility, acceptability, catalytic and educational effect.

Results After removal of duplicates, 2852 papers were screened for inclusion, of which 66 met all relevant criteria. Over 43 different exams and more than 28 different specialties from 18 jurisdictions were studied. Around 77% of all eligible papers were based in English-speaking countries, with 55% of publications centered on just the UK and USA. General Practice was the most frequently studied specialty among certification exams with the UK General Practice exam having been particularly broadly analyzed. Papers received an average of 4.2/6 points on the quality score. Eligible studies analyzed 2.1/7 Ottawa Criteria on average, with the most frequently studied criteria being reliability, validity, and acceptability.

Conclusions The present systematic review shows a growing number of studies analyzing medical specialty certification exams over time, encompassing a wider range of medical specialties, countries, and Ottawa Criteria. Due to their reliance on multiple assessment methods and data-points, aspects of programmatic assessment suggest a promising way forward in the development of medical specialty certification exams which fulfill all seven Ottawa Criteria. Further research is needed to confirm these results, particularly analyses of examinations held outside the Anglosphere as well as studies analyzing entire certification exams or comparing multiple examination methods.

†Andrea Lörwald and Sören Huwendiek shared last authorship.

*Correspondence:
Daniel Staudenmann
daniel.staudenmann@gmail.com
Full list of author information is available at the end of the article

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Keywords Medical education, Specialty certification examination, Validity, Reliability

Background

Patients rely on doctors for safe and effective medical care, yet preventable adverse events remain prevalent [1]. How can such events be avoided? One answer lies in professional assessments. Before being allowed to practice medicine independently, e.g., in their own private practice, doctors must pass a postgraduate exam which aims to test the skills, knowledge and attitudes relevant to their chosen medical specialty. Such medical specialty certification exams have long been used in many countries but differ greatly in their implementation. Historically, a simple oral examination by a senior colleague has often sufficed [2], but recent evidence supports the effectiveness of "triangulation", a more multifaceted approach including assessment methods such as multiple choice questions (MCQs) and objective structured clinical examinations (OSCEs) [3].

The terminology of specialty exams differs substantially by country, even in the peer-reviewed literature published for a world-wide audience [4]. "Specialist medical assessment", "board exam", "postgraduate certification process", "specialty certificate examination" and resulting acronyms are all commonly used. The institutions responsible for organizing the exams vary depending on country and medical specialty, as do the skills, knowledge and time spent training required as well as the privileges granted to a successful candidate [2].

In this article, we define a medical "specialty certification exam" as a high-stakes summative assessment of a candidate which takes place after completing postgraduate training such as residency, which is essential for career progression and – upon successful completion – typically allows the candidate to treat patients as an independent medical specialist. In the USA, 87% of physicians choose to get certified despite certification being voluntary [5]. Examples of specialty certification exams include the British "Royal College Membership exams", the American "Board Certification" and the Swiss "Facharztprüfungen".

Specialty certification exams are crucial to patient safety. Successful completion should guarantee the minimum level of competencies needed to diagnose and treat patients without a senior colleague readily available and formally responsible for ensuring the quality of the junior doctor's treatment. Previous research shows that certified doctors generally provide better medical care than non-certified ones [6–9]. A systematic review by Lipner et al. shows that certification status is correlated with various clinical measures such as defibrillator complication rates or acute myocardial infarction mortality. In the majority of 29 studies, certified physicians provided better patient care [10]. To provide just one example, a study by Reid et al. shows certified physicians performing 3.3 percentage points higher on a quality performance composite than non-certified physicians across 23 specialties [9]. However, medical errors remain common overall [11] and cases of professional misconduct are regularly discussed in the media [12]. In one retrospective study, patients received only 54.9% of recommended basic care [13]. In American hospitals alone, medical errors are estimated to cause over 400′000 premature deaths per year [14], making it the third highest cause of death [15]. As the final examination of legally required formal education in many countries, specialty certification exams provide the last opportunity to identify physicians who do not (yet) qualify for unsupervised practice. They therefore play a crucial role in publicly guaranteeing practicing physicians' competence.

In this study, we use the "Criteria for Good Assessment: Consensus Statement and Recommendations" from the Ottawa 2010 Conference ("Ottawa Quality Criteria") to evaluate different medical specialty exams. This consensus statement was developed by a working group of medical assessment experts from various countries including Norcini et al. [16] and revised in 2018 [17]. They recommend the following seven criteria (Table 1):

Previous research

Given how important specialty certification exams are, there is a surprising lack of evidence pertaining to their efficacy [18]. Current literature often focuses on subspecialty specific exams in individual countries. To the best of our knowledge, the last systematic review was published in 2002 by Hutchinson et al. The authors searched different databases for studies published between 1985 and 2000, initially found 7705 and excluded all but 55 from their analysis. Hutchinson et al. then summarized each paper regarding any form of validity and reliability analyzed within. They remark on the paucity of published data, finding the under-representation of hospital specialties in particular "striking". They call for a repeated analysis in the future and for increased openness "from many of the institutions that have a powerful and unopposed role in the career paths of doctors in training" [19].

Interest in the topic of effective medical education has increased sharply since then, yet there remains a gap in the literature concerning many specialties and countries. Hospital specialties are under-represented, while general

Table 1 Framework for good assessment according to the consensus statement and recommendations from the Ottawa 2010 conference [16]

Criterion	Explanation
Validity or Coherence	The results of an assessment are appropriate for a particular purpose as demonstrated by a coherent body of evidence
Reproducibility, Reliability or Consistency	The results of the assessment would be the same if repeated under similar circumstances
Equivalence	The same assessment yields equivalent scores or decisions when administered across different institutions or cycles of testing
Feasibility	The assessment is practical, realistic, and sensible, given the circumstances and context
Educational Effect	The assessment motivates those who take it to prepare in a fashion that has educational benefit
Catalytic effect	The assessment provides results and feedback in a fashion that motivates all stakeholders to create, enhance, and support education; it drives future learning forward and improves overall program quality
Acceptability	Stakeholders find the assessment process and results to be credible

or family practice predominates (covering 41 out of the 55 papers Hutchinson et al. identified). Hutchinson et al. found studies from only six countries, of which five were located in the Anglosphere [19]. Given their widespread use globally, the quality of most medical specialty exams remained to be scientifically studied according to either of the first two Ottawa Criteria (validity and reliability). To expand upon this research and address this gap in the literature, this systematic review focuses on collating up-to-date practices which have been analyzed according to any of the Ottawa Criteria from as many different countries and specialties as possible.

Goal of this review

In this systematic literature review we aim to give an overview of the current evidence regarding specialty certification exams as studied according to any of the Ottawa Criteria of Good Assessment globally. We show which medical specialties, countries and examination formats have been analyzed regarding which of the Ottawa Criteria. This provides a point of reference for future researchers or medical specialty societies looking to study or further develop their exams.

The following research questions guide this systematic review:

(1) Which medical specialty certification exams have been scientifically studied regarding the Ottawa Quality Criteria?
(2) Which Ottawa Criteria were analyzed in these exams?
(3) Which specialty certification exam has been studied most extensively in regard to the Ottawa Criteria?

Methods and analysis

Search strategy

Studies were compiled using the following seven databases: MEDLINE(R) ALL, EMBASE, APA PsycINFO and ERIC via Ovid, SCOPUS, the Cochrane Trial Library and Web of Science.

To reflect contemporary practice and continue from the timeframe used in Hutchinson et al.'s study, a search of the literature published between January 2000 and August 2020 was performed. The Population, Intervention, Comparison, Outcomes and Study (PICOS) design framework was used to establish the search strategy (see Table 2).

Because of the varying nomenclature of "specialty certification exams", we expanded our search terms to cover over 20 variations and included the medical subject headings "Specialty Boards" and "Educational Measurement". In addition, papers must reference the concept of medicine (e.g. "medic*") and a form of evaluation criteria (e.g. "valid*"). Beyond this, we only include papers written in English and published between 2000 and 2020. The search terms described in Additional file 1 were used and adapted to the seven individual databases (see Additional file 1).

Literature selection

The title, abstract and citation information of all results were retrieved from Ovid.com, Scopus.com, Webofknowledge.com and Cochranelibrary.com using the ris and Excel or csv format. They were imported into EndNote X9 and manually merged into an Excel file. The following information was made available separately to two researchers (DS and NW) for an initial round of screening: title, authors, year of publication and abstract. In this first round of screening all potentially valuable studies were included even if the fulfilment of certain criteria was questioned by one or both researchers. For instance, rather than examinations *of* physicians, many studies look at examinations *by* physicians. Others focus not on medical specialty exams but medical student exams, re-certification, maintenance of certification or formative workplace-based assessments. Papers describing assessments of other

Table 2 PICOS framework

PICOS Elements	Characteristics
P—Population	Postgraduate medical trainees, physicians post completion of university studies
I—Intervention	High-stakes summative assessment of a medical specialist candidate which takes place after completion of postgraduate practical training (e.g. residency), which is necessary for career progression and typically allows the candidate to treat patients as an independent medical specialist upon successful completion Examples are the Swiss "Facharztprüfungen", the British "Royal College Fellowship / Membership Exams" and the American "Board Certification"
C—Comparison	-
O—Outcome	Exam evaluation as measured by at least one of the Ottawa Criteria (validity, reproducibility, feasibility, equivalence, educational effect, catalytic effect, acceptability)
S—Study design	All study types

professions were similarly excluded (e.g. physician's assistants, nurses and pharmacists). Manuscripts that weren't published as complete scientific studies (e.g. conference papers, letters, editorials, reviews) were also excluded in this round, as were papers unavailable in English. In the second round the full text papers were assessed individually by DS and NW to determine whether the pre-selected papers fit the research questions. Here, papers were more likely to be excluded due to a lack of clarification of which examination method or methods were analyzed, because they assess exams which are administered almost immediately after candidates leave university or do not allow candidates to treat patients independently upon successful completion. Cases in which independent reviewers came to different conclusions were discussed bilaterally. If no agreement was found, a third reviewer (AL) was consulted for final judgement.

Analysis

All results and interpretations pertaining to the Ottawa Quality Criteria were extracted from the included papers and categorized according to the seven criteria. DS and NW initially performed this step collaboratively to ensure they were able to reach consistent results, later extractions were then performed independently. Papers in which the relevant data or categorization was complex or unclear were discussed until agreed upon by DS, NW, and AL (see Additional file 2).

For further analysis, information about the country or countries studied, medical specialty or specialties, examination method or combination of methods as well as further relevant details about the examination were retrieved from the full text of all included papers. Where papers lacked such details, they were supplemented where possible by searching online. Details published e.g. by the medical society in question were added to the final overview, and the source of this additional information was included for reference. Papers were also shortly summarized for the convenience of the reader.

The methodological quality and relevance to this review's research questions was evaluated for each study using the appraisal criteria adapted from the Medical Education Research Study Quality Instrument (MERSQI) and the "Criteria for the qualitative assessment of scientific publications" [20, 21]. The evaluation consists of the following six criteria: (1) Is the study design suited to answering the question studied? (2) Is the method described so that replication is possible without further information? (3) Is the interpretation coherent? (4) Does the study analyze at least 50 exam candidates? (5) Does it analyze more than one exam? (6) Does it analyze the entire exam(s)? Papers received one point if yes, zero if no or unclear.

Lastly, the data extracted was used to search for the specialty certification examination or examinations that were most extensively studied regarding the Ottawa Criteria by counting the number of separate Ottawa Criteria investigated as well as number of individual studies in cases where multiple papers were published that analyze the same exam.

Results

The inclusion and exclusion process is visualized according to the PRISMA flow diagram (see Fig. 1). Out of 4420 hits, 66 studies were included for data analysis.

In the 66 papers, over 43 different exams and more than 28 specialties from 18 jurisdictions are assessed.

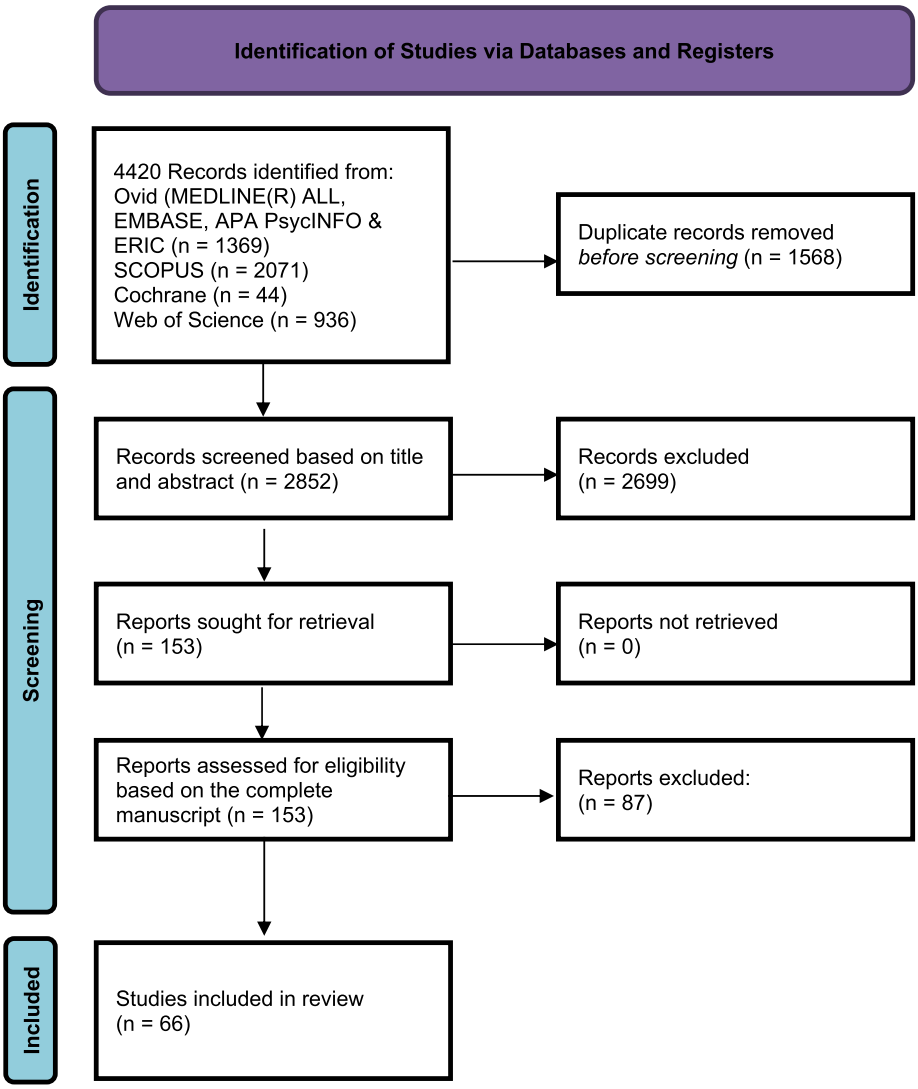

Fig. 1 PRISMA flow diagram

Overview of studies

All included studies are sorted by medical specialty and country in Table 3. The Ottawa Criteria analyzed therein are marked (+), those not analyzed (0). The right-most column shows how many of the six metrics studies fulfilled on the quality assessment tool. The examination formats used are listed, with the focus of the studies marked in bold. A complete overview including short summaries of all studies and the relevant findings can be found in the additional Excel spreadsheet (see Additional file 2).

Location

A large majority of our search results examine the specialty certification exams used in English-speaking countries, with 77% of papers focusing on the UK, USA, Australia, Ireland, Canada, or South Africa. By far the two most frequently studied countries are the United Kingdom and the United States, together comprising 55% of eligible papers (20 publications each, see Fig. 2). Other locations studied include Israel, China Hong Kong, Argentina, Egypt, Iran, Japan, Mexico, the Philippines, Portugal, and Thailand. A minority of exams are not specific to only one country: three papers look at European exams, two at Australia and New Zealand, one at the USA and Canada, and one at the UK and Ireland. A single study compares the exams across multiple countries [88].

Specialty

The certification exams used in the specialty of General Practice are the most frequently studied, with 11 studies

Table 3 Overview of studies

Author, year	Medical specialty certification exam	Examination methods	Validity	Reproducibility	Equivalence	Feasibility	Educ. effect	Catal. effect	Acceptability	Quality metrics
Berkenstadt et al., 2006 [22]	Anesthesiology—Israel	Oral and **OSCE**	+	+	0	0	0	0	+	5
Sun et al., 2019 [23]	Anesthesiology—USA	MCQ, oral (**SOE** & OSCE)	+	+	0	0	0	0	0	4
Warner et al., 2020 [24]	Anesthesiology—USA	MCQ, oral (SOE & **OSCE**)	0	+	0	+	0	0	+	4
Warner et al., 2020 [25]	Anesthesiology—USA	MCQ, oral (SOE & **OSCE**)	0	0	0	+	0	0	0	4
Gali et al., 2011 [26]	Cardiology—Argentina	**MCQ (SBA)**, oral (with real patients)	+	0	0	0	0	0	+	3
Tan et al., 2008 [27]	Clinical Oncology—United Kingdom	**MCQ (SBA)**, oral (structured & clinical)	+	0	0	0	0	0	0	5
O'Leary, 2015 [28]	Emergency Medicine—Australia, New Zealand	Written & **Clinical (OSCE)**	+	+	0	0	0	0	0	3
Bianchi et al., 2003 [29]	Emergency Medicine—USA	MCQ, **oral (simulated cases)**	0	+	0	0	0	0	0	4
Slovut et al., 2015 [30]	Endovascular Medicine—USA	**MCQ**	+	+	0	0	0	0	+	4
Khafagy et al., 2016 [31]	Family Medicine—Egypt	**MCQ, SBA**, clinical assessment	+	+	0	0	0	0	0	5
Weingarten, et al., 2000 [32]	Family Medicine—Israel	MCQ, **oral** (structured)	0	+	0	0	0	0	0	5
O'Neill et al., 2011 [33]	Family Medicine—USA	**Written (MCQ, SBA)**, oral (case discussion)	+	0	0	0	0	0	0	4
O'Neill et al., 2019 [34]	Family Medicine—USA	**Written, oral (case discussion)**	+	0	0	0	0	0	0	4
Greco et al., 2002 [35]	General Practice—Australia	Applied Knowledge Test (MCQ), Key Feature Problem (KFP), **Objective Structured Clinical Examination (OSCE)**	+	0	0	0	0	0	0	4
Munro et al., 2005 [36]	General Practice—United Kingdom	**Written (free text answers** and MCQ), Oral, consultation skills	+	+	0	0	0	0	0	5
Simpson et al., 2005 [37]	General Practice—United Kingdom	Written (free text answers and MCQ), **Oral**, consultation skills	+	0	0	0	0	0	0	4
Sandars et al., 2004 [38]	General Practice—United Kingdom	Written **(free text answers** and MCQ), Oral, consultation skills	+	+	+	0	0	0	0	2

Table 3 (continued)

Author, year	Medical specialty certification exam	Examination methods	Validity	Reproducibility	Equivalence	Feasibility	Educ. effect	Catal. effect	Acceptability	Quality metrics
Dixon, 2005 [39]	General Practice—United Kingdom	Written (free text answers and **MCQ**), Oral, consultation skills	+	0	0	0	+	+	+	3
Siriwardena et al., 2006 [40]	General Practice—United Kingdom	Written (free text answers and MCQ), Oral, **consultation skills**	+	0	0	0	0	0	0	4
Dixon, 2003 [41]	General Practice—United Kingdom	**Written (free text answers and MCQ), Oral, consultation skills**	0	0	0	0	+	+	+	4
Wass et al., 2003 [42]	General Practice—United Kingdom	Written (free text answers and MCQ), **Oral (structured)**, consultation skills	+	+	0	0	0	0	0	4
Dixon et al., 2015 [43]	General Practice—United Kingdom	**Applied Knowledge Test (MCQ);** Clinical Skills Assessment (OSCE); Workplace Based Assessment (WBA)	+	0	0	0	+	0	+	4
Partridge, 2008 [44]	General Practice—United Kingdom	**Written (free text answers and MCQ)**, Oral (structured), consultation skills	+	0	0	0	0	0	+	5
Dixon et al., 2007 [45]	General Practice—United Kingdom	Written (free text answers and **MCQ**), Oral (structured), consultation skills	+	+	0	0	0	0	+	3
Bourque et al., 2020 [46]	Internal Medicine—Canada	**MCQ,** OSCE	0	+	0	0	0	0	0	4
Chierakul et al., 2010 [47]	Internal Medicine—Thailand	Written and **clinical** (real patients)	+	+	0	0	0	0	0	4
McManus et al., 2003 [48]	Internal Medicine—United Kingdom	**Part 1: MCQ (true–false)**, Part 2: unclear	0	+	0	0	0	0	0	4
McManus et al., 2013 [49]	Internal Medicine—United Kingdom	Part 1: MCQ (SAQ); Part 2: Written**; Part 2 Clinical (PACES)**	+	0	0	0	0	0	0	5
McManus et al., 2006 [50]	Internal Medicine—United Kingdom	Part 1: MCQ (SAQ); Part 2: Written**; Part 2 Clinical (PACES)**	0	+	0	0	0	0	0	5
Atsawarungruangkit, 2015 [51]	Internal Medicine—USA	**MCQ (SBA)**	+	0	+	0	0	0	0	5

Table 3 (continued)

Author, year	Medical specialty certification exam	Examination methods	Validity	Reproducibility	Equivalence	Feasibility	Educ. effect	Catal. effect	Acceptability	Quality metrics
Marques et al., 2018 [52]	Multiple—Portugal	**Curriculum analysis, practical (real patient exam, discussion) and theoretical tests (Oral or MCQ)**	+	+	0	0	0	0	0	6
Burch et al., 2009 [53]	Multiple—South Africa	**Short-answer question test (SAQT), Data interpretation test (DIT), real patient encounters (PE)**	+	0	0	0	0	0	0	4
Burch et al., 2008 [54]	Multiple—South Africa	Written: **MCQ, short-answer question tests (SAQT)** Oral: **Real patient encounters followed by oral test**, Data interpretation Test (DIT), PE (real patient encounters and oral exam)	0	+	0	0	0	0	0	6
Cookson, 2010 [55]	Multiple—United Kingdom	**MCQ (SBA)**	+	+	0	+	+	0	+	4
Mucklow, 2011 [56]	Multiple—United Kingdom	**MCQ (SBA)**	+	+	0	0	0	0	+	6
Raddatz, et al., 2012 [57]	Not specified—USA	Not specified	+	0	0	0	0	0	0	3
Lunz et al., 2008 [58]	Not specified—USA	**Oral (real or realistic patient cases)**	0	+	0	0	0	0	0	5
Houston et al., 2009 [59]	Not specified—USA	**Oral (structured)**	+	+	0	0	0	0	0	3
Mathysen et al., 2013 [60]	Ophthalmology—Europe	written: **MCQ (true–false)**, oral (structured)	+	+	0	0	0	0	0	5
Mathysen et al., 2013 [61]	Ophthalmology—Europe	written: **MCQ (true–false)**, oral (structured)	+	+	0	+	0	0	0	5
Chow et al., 2017 [62]	Palliative Medicine—China, Hong Kong	Dissertation Appraisal Examination, **Oral Examination**	+	+	0	0	0	+	0	3
Althouse et al., 2009 [63]	Pediatrics—USA	**MCQ (SBA)**	+	0	0	0	0	0	0	5
Emadzadeh et al., 2017 [64]	Pediatrics & Gynecology—Iran	MCQ, **OSCE**	0	0	0	0	0	0	+	4
Raddatz et al., 2017 [65]	Physical Medicine and Rehabilitation—USA	**MCQ**, oral (structured)	+	+	0	0	0	0	0	4
Tibbo et al., 2004 [66]	Psychiatry—Canada	MCQ, **oral/clinical (real patients)**	0	0	0	+	0	0	+	4
Tong et al., 2018 [67]	Radiology—Europe	**Written** (SBA, MAQ, Order), **Oral** (unclear)	+	0	0	0	+	0	+	5

Table 3 (continued)

Author, year	Medical specialty certification exam	Examination methods	Validity	Reproducibility	Equivalence	Feasibility	Educ. effect	Catal. effect	Acceptability	Quality metrics
Yeung et al., 2013 [68]	Radiology—United Kingdom	MCQ, rapid reporting session, long-cases reporting session, **oral (structured)**	0	+	0	+	0	0	0	4
Yeung et al., 2011 [69]	Radiology—United Kingdom	MCQ, SBA, **reporting session, rapid reporting session, oral (structured)**	+	+	0	0	+	0	+	4
Yang et al., 2013 [70]	Radiology—USA	Written (unclear), **oral (structured)**	0	+	0	0	0	0	0	4
Kerridge et al., 2016 [71]	Radiology—USA	**MCQ,** oral (structured)	+	0	0	+	+	0	+	1
Yang et al., 2010 [72]	Radiology—USA	**MCQ**, oral	+	0	+	0	0	0	0	4
Pascual-Ramos et al., 2018 [73]	Rheumatology—Mexico	MCQ, **OSCE**	+	0	0	+	0	0	0	4
Smith et al., 2007 [74]	Rural and Remote Medicine—Australia	**Written (MCQ, SBA), StAMPS,**	+	+	0	+	+	0	+	4
Beasley et al., 2013 [75]	Surgery (9 different surgical specialties)—Australia, New Zealand	**Written and clinical/viva**	+	+	0	0	0	0	0	6
De Montbrun, 2016 [76]	Surgery (colon and rectal)—USA	MCQ, oral, **COSATS (technical skill tasks)**	+	+	0	+	0	0	0	4
Lineberry et al., 2020 [77]	Surgery (endoscopic)—USA	MCQ, **manual skills test**	+	+	0	+	0	0	0	4
Motoyama, et al., 2020 [78]	Surgery (esophageal)—Japan	**Clinical experience**	+	0	0	0	0	0	0	5
De Montbrun, 2017 [79]	Surgery (general and colorectal)—USA, Canada	**Multiple**	+	+	0	+	0	0	0	5
Crisostomo, 2011 [80]	Surgery (general)—Philippines	MCQ, **oral (structured)**	+	+	0	0	0	0	0	4
Rhodes et al., 2007 [81]	Surgery (general)—USA	**MCQ, oral**	+	0	0	0	+	0	0	3
Cundy, 2012 [82]	Surgery (orthopaedic)—Australia	Written and **clinical/viva** (real patients)	0	+	0	0	0	0	+	1
Hohmann et al., 2018 [83]	Surgery (orthopaedic)—Australia, UK, South Africa & Canada	**Multiple: Written (MCQ, essays, short questions), oral (viva voce), operative sessions, clinical sessions or OSCEs**	+	0	+	0	0	0	0	5
Gillis et al., 2020 [84]	Surgery (orthopaedic)—Canada	**S-OSCE**	+	0	0	+	0	+	0	3
Ullmann et al., 2006 [85]	Surgery (plastic)—Israel	MCQ, **oral** (unstructured)	+	0	0	0	0	0	+	4

Table 3 (continued)

Author, year	Medical specialty certification exam	Examination methods	Validity	Reproducibility	Equivalence	Feasibility	Educ. effect	Catal. effect	Acceptability	Quality metrics
Dwyer et al., 2020 [86]	Surgery (sports medicine)—Canada	MCQ, **OSCE**, in-training evaluation, surgical logbook, **intraoperative and cadaveric assessment**	+	+	0	0	+	0	0	4
Payne et al., 2011 [87]	Urology—United Kingdom, Ireland	**MCQ (SBA, EMI), oral (structured)**	+	0	0	0	0	0	+	6

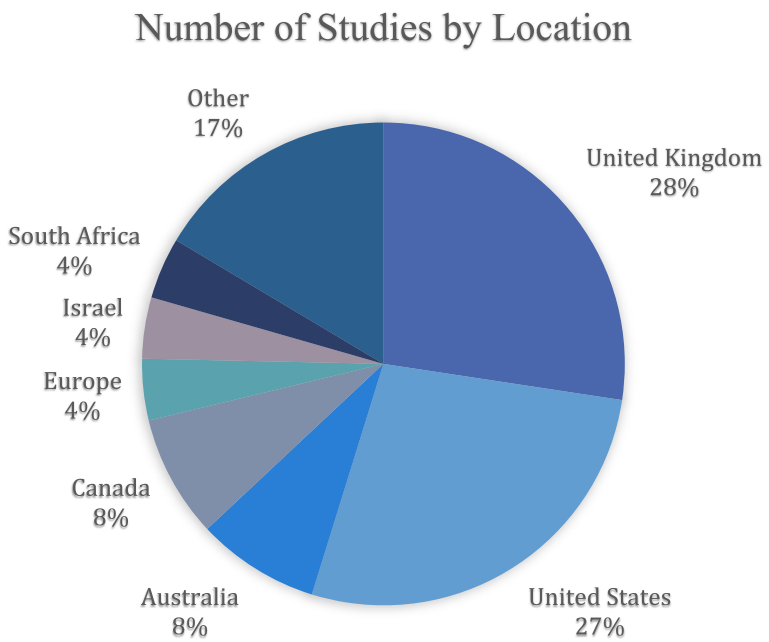

Fig. 2 Number of studies by location

focusing on this domain. Various kinds of surgical specialties are studied in 12 different publications. Internal Medicine is analyzed in six different studies. A further six studies assess the exams used in Radiology. Anesthesiology and Family Medicine are the medical specialties under consideration in four studies each. Emergency Medicine, Ophthalmology, and Pediatrics are each analyzed in two studies. One study has been published about each of the following medical specialty certification exams: Cardiology, Clinical Oncology, Endovascular Medicine, Gynecology, Palliative Medicine, Physical Medicine and Rehabilitation, Psychiatry, Rheumatology, Rural and Remote Medicine, and Urology (see Fig. 3). Five studies evaluate multiple medical specialties. Three studies fail to specify which specialty exam is under analysis.

Methodological quality and relevance assessment

Eligible papers receive an average of 4.15 out of the six possible points for relevance and methodological quality. 94% (62/66, see Fig. 4) of studies receive a point for criterion 1 ("Is the study design suited to answering the question studied?"), 89% (59/66) for criterion 2 ("Is the method described so that replication is possible without further information?"), and 95% (63/66) for criterion 3 ("Is the interpretation coherent?"). The number of candidates analyzed is at least 50 in 73% (48/66) of studies (criterion 4). 36% (24/66) of studies compare multiple exams (criterion 5). Finally, 28% (18/66) of the included studies analyze the entire exam(s) (criterion 6). Many focus on only a subset of the specialty certification exam, though some studies also receive zero points on this metric since it is unclear what the entire specialty certification exam under consideration consists of.

Examination methods

The nomenclature varies widely across different examination modalities. The most common methods used include multiple choice questions (MCQ), structured oral exams with expert discussions and objective structured clinical examinations (OSCE). Essay questions, dissertation appraisal or clinical experience are less frequently evaluated in the studies. Few medical specialty certification exams only use a single examination method. A large majority of studies published therefore focus on exams using a combination of different modalities, comprising of at least one written and one oral method.

Ottawa criteria

On average, studies examine 2.1 of the 7 Ottawa Criteria. The most frequently studied criterion is validity (51/66 studies), followed by reliability (37/66) and acceptability (20/66). Feasibility is a topic of analysis in 13 papers. Equivalence and catalytic effect are least commonly researched, with 4 studies mentioning results belonging to those categories each (see Fig. 5).

No medical specialty certification exam has been analyzed in respect to all seven Ottawa Criteria. Even when collating evidence from multiple studies, only 16 out of 46 exams have been analyzed in respect to three or more.

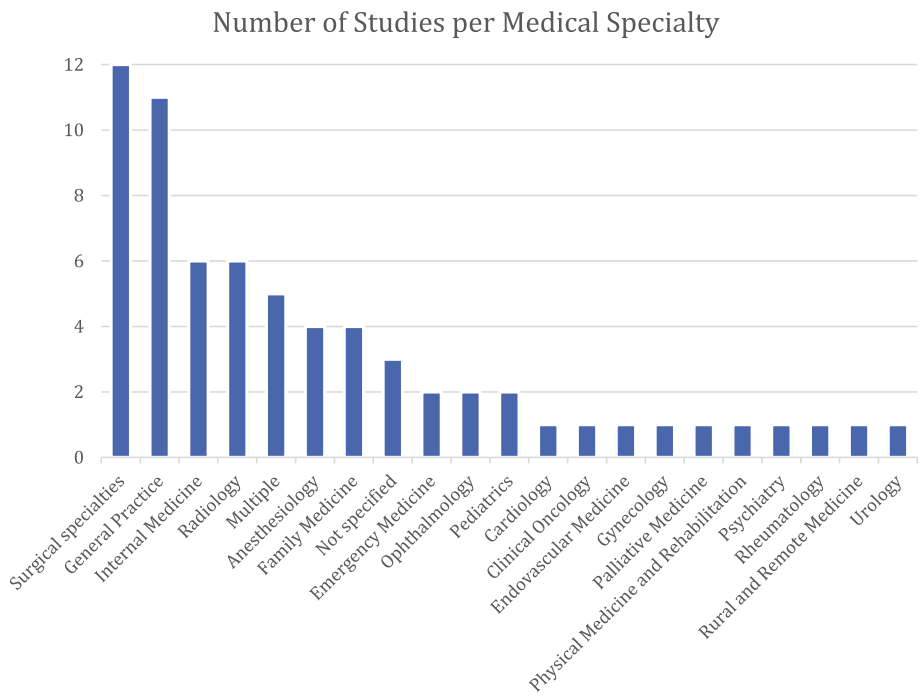

Fig. 3 Number of studies by medical specialty

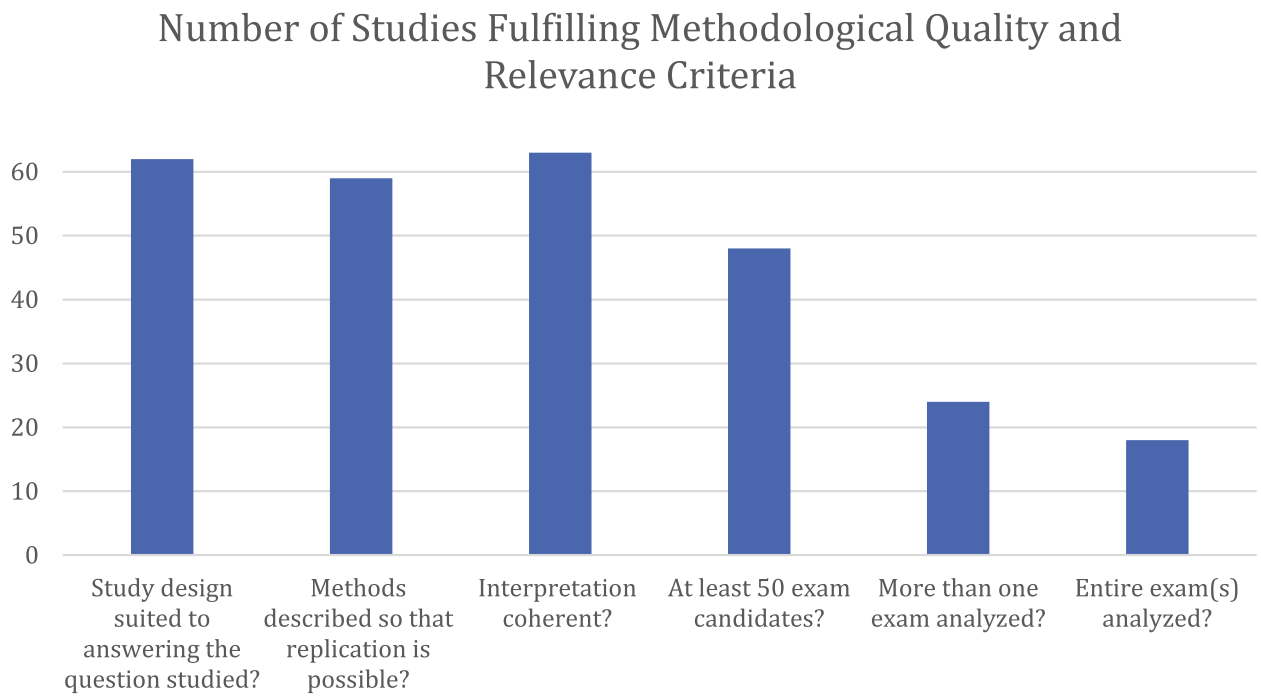

Fig. 4 Number of studies fulfilling methodological quality and relevance criteria

Three exams have been analyzed in respect to five, and two exams in respect to six of the seven Ottawa Criteria (see Fig. 6).

The most extensively studied exam – The MRCGP

The Membership of the Royal College of General Practitioners (MRCGP) exam is the most extensively

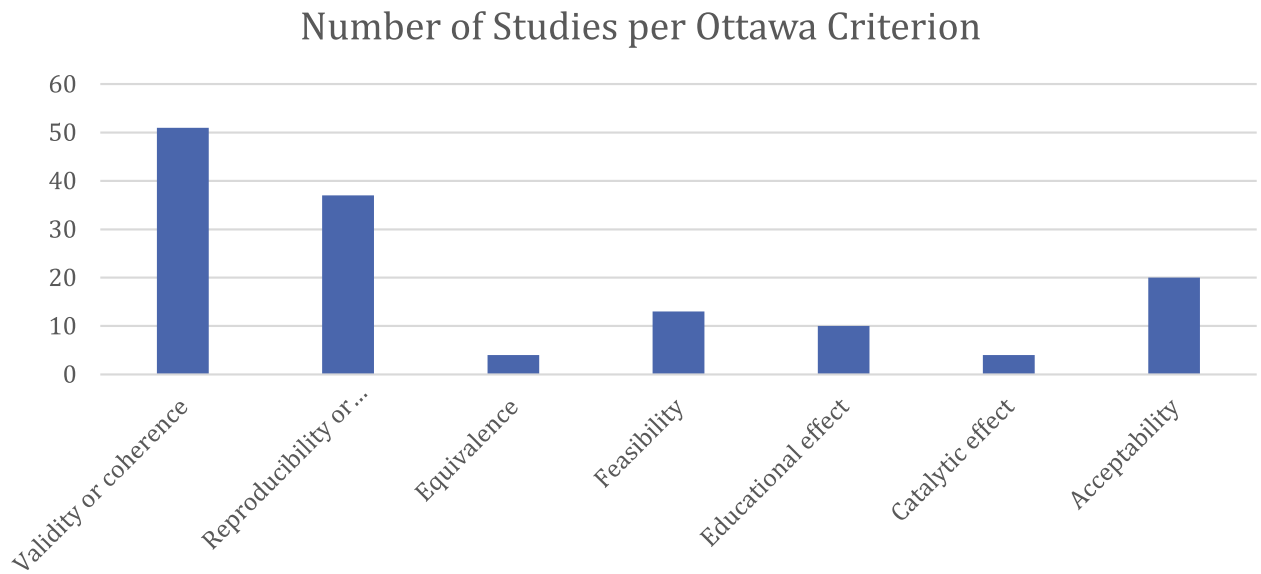

Fig. 5 Number of Studies per Ottawa Criterion

studied specialty certification exam regarding the Ottawa Criteria, with 10 different papers published between 2000 and 2020. Apart from the feasibility criterion, all Ottawa Criteria are covered by the literature.

The MRCGP examination aims to test the skill and knowledge of a doctor who "has satisfactorily completed specialty training for general practice and is competent to enter independent practice in the UK without further supervision" [89]. Although changes continue to be made to the exam's format, in nearly all the included studies it is described as consisting of the following four parts: a written exam (called "written paper") made up of free text answers, a multiple-choice question exam, an oral exam and a video section examining consultation skills [37–39, 41, 42]. The individual research studies published about the MRCGP are presented in more detail below.

Dixon [41] surveyed registrars about their views on the various MRCGP modules and their effects on learning. He found that candidates perceived study groups of fellow registrars as particularly helpful to prepare for the written and oral components, and feedback from trainers as especially useful for the consultation skills video component. Many said they had read more review articles but not original articles as preparation. Most candidates believed that preparing for the oral module increased their understanding of moral and ethical principles.

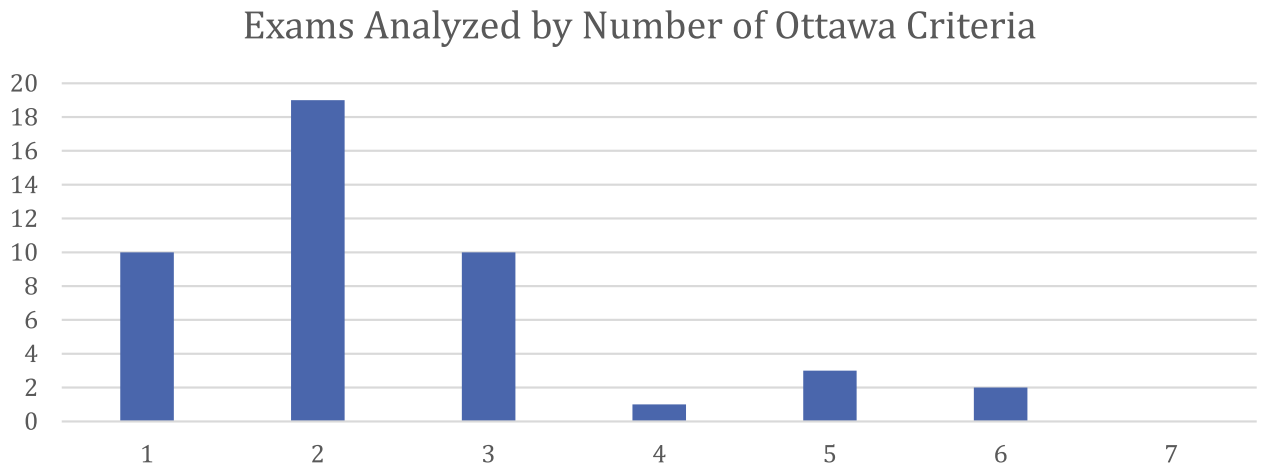

Fig. 6 Exams Analyzed by Number of Ottawa Criteria

The "written paper" in the MRCGP examination aims to test candidates' problem-solving skills, knowledge of current literature and critical appraisal skills of study methodology. As described by Sandars et al. [38], candidates are awarded three and a half hours to read three literature extracts and write concise "notes" answering 12 questions, usually about the studies' methodology and how this relates to a given scenario relevant to general practitioners (GPs). Munro et al. [36] show these examiner-marked "Free Text Answers" achieve relatively high measures of reliability with Cronbach's alpha consistently lying between 0.85 and 0.88. Dixon [39] asked candidates to rate their impression of the question formats, finding that the single best answer (SBA) format was rated easiest, and the treatment algorithm completion, extended matching and summary completion questions as more difficult. Summary completion questions were also criticized for testing language ability. Overall, the acceptability of the written paper was high among candidates, although a majority believe this module also contained inappropriate questions. The written paper module of the MRCGP exam seemed particularly helpful in encouraging candidates to regularly read journal articles. Partridge [44] supports this view and further emphasizes the importance of critical appraisal skills for the written paper. A large majority of the candidates were generally satisfied with this part of the exam and found the questions to be clear and relevant to General Practice.

The "Multiple Choice Paper" (MCP) uses a number of question formats including SBA, images and extended matching questions to test the breadth and depth of candidates' knowledge. Dixon et al. [45] asked GP trainers to sit a shortened version of the MCP, finding that they significantly outperform registrars even with no preparation on overall scores as well as questions specifically related to General Practice and practice administration. Trainers did not manage to answer questions related to research methodology or critical appraisal significantly more often than candidates. Accordingly, despite other question topics being perceived as easy, research methodology and critical appraisal questions were rated as difficult by most trainers. Dixon et al. [43] also summarize candidates' views on this part of the exam, finding that it was perceived to succeed in its aim of being a fair test of candidates' knowledge and relevant to General Practice. The topics research and statistics were found to be most difficult by those candidates. However, they did not achieve lower mean scores in these fields. Small adjustments such as adding a calculator and allowing ten minutes extra time were made in response to this feedback.

The "Oral Exam" aims to test candidates' decision-making skills and professional values using two 20-min oral exams with a pair of examiners and five topics each. The examiners chose their own questions. Wass et al. [42] find the reliability coefficients to be lower than required (intercase 0.65, pass/fail 0.85) and recommend increasing the testing time and number of topics covered, suggesting five oral exams with one examiner each. This would increase intercase reliability to 0.78 and pass/fail reliability to 0.92. Simpson et al. [37] also looked at the oral exam, arguing that "the assessment of professional values was largely examined at the level of knowledge and comprehension, with few examiners encouraging candidates to justify their expressed viewpoint or allowing them to demonstrate how they might use these values to support their decision making."

For the "Consulting Skills Assessment", candidates are asked to submit videos of themselves interacting with seven real patients. They can choose those seven consultations to best demonstrate 15 performance criteria and are then rated by seven independent GPs trained for this assessment. Siriwardena et al. [40] compared this module with the 'observing patient involvement' (OPTION) scale – an independently validated scale for shared decision making – finding that it predicts both the performance criterion 'sharing of management options' as well as overall MRCGP results.

Discussion

This systematic review tackles an important question in current medical education research: How can we credibly test and certify physicians' competence? Specialty certification exams are crucial for patient safety, candidates' career progression and accountability to the public, yet evidence to their quality has thus far been lacking. By searching seven different databases and using a wide variety of possible variations in search terms, we collate a comprehensive outline of the research regarding studied Ottawa Quality Criteria in specialty certification exams published in the past twenty years. 66 studies were included. Reliability, validity, and acceptability are the criteria most frequently studied in respect to specialist exams in this literature. As was the case in the previous literature review by Hutchinson et al., the largest body of evidence is centered on the UK and USA as well as the General Practice specialty [19]. However, we document a large increase in the number of different countries, medical specialties and Ottawa Criteria studied during the past twenty years.

The exact nomenclature used to describe examination quality indicators in the literature and the relative emphasis of the authors may vary, yet there exists widespread agreement as to which qualities good examinations must fulfill. Medical specialty certification especially must be valid, reliable, and objective. When repeated, they ought to give similar results and therefore be reproducible,

independently of factors such as examiner bias [3]. Furthermore, such assessments must be feasible, remain as cost-efficient as possible, and provide adequate feedback *for* and *of* learning [90]. All these aspects are covered through the seven criteria for good assessment from the Ottawa Conference chosen for this review: validity, reliability, equivalence, feasibility, acceptability, catalytic and educational effect. Due to the high-stakes summative nature of specialty certification exams, the focus often lies on ensuring validity and reliability rather than educational or catalytic effect. This trend is reflected in the number of studies found analyzing each criterion.

We can see how Ottawa Criteria sometimes conflict with each other. For instance, although acceptability among candidates may suffer if an examination program neglects the provision of constructive feedback, the priority for an institution organizing the exam may lie on credibly signaling to the public, healthcare institutions and patients that a passing candidate is ready for independent practice. How such tradeoffs among quality criteria may be improved with limited resources can be studied with the feasibility criterion. Feasibility, including financial cost associated with different examination methods, is thus a major concern regarding high-quality specialty certification exams organized in resource-constrained contexts. Despite its relevance, we observe a relative scarcity of studies concerning this criterion [24, 25, 55, 61, 66, 68, 71, 73, 74, 76, 77, 79, 84, 91].

According to Miller's framework for assessing clinical skills, competence and performance, clinical assessment can be conceptualized in four progressive levels: the learner proceeds through "knows," "knows how" and "shows how" to "does" [92]. It is the first that is easiest to test reliably on a written exam, yet proficiency at the highest level must be reached before a candidate can be certified for independent practice. Unless they can demonstrate their knowledge, skills, as well as attitudes, we cannot be sure this is the case: "No single assessment method can provide all the data required for judgment of anything so complex as the delivery of professional services by a successful physician" [92]. All examination methods face limitations on at least one Ottawa Quality Criterion and cannot be expected to cover all levels of Miller's framework [93]. Well-designed specialty certification exams manage to also check the higher levels of Miller's pyramid, and thereby make the exam conditions match the reality of working as a certified physician more closely [3, 92, 94].

The necessity of combining different assessment methods in specialty certification exams was highlighted for US internal medicine residents specifically in a 1998 non-systematic review article by Holmboe et al. [95]. They summarize studies published between 1966 and 1998 and argue that since the written American Board of Internal Medicine (ABIM) certification exam alone is insufficient to adequately assess clinical competence, it should be supplemented by other examination methods in the clinic such as rating scales of interpersonal skills and attitudes, medical record audits, clinical evaluation exercises (CEX) and standardized patient exams. More recent work has further recommended expanding assessment to include competencies such as teamwork and population care [96, 97]. It is therefore encouraging that a majority of medical specialty certification exams analyzed in this review use triangulation methods and e.g. complement multiple choice questions ("knows" and "knows how") with OSCEs ("shows how") and workplace based assessments ("does") [22–29, 31, 35–46].

Single exams only at the end of an educational period can lead candidates to ignore the feedback given [93]. Further development of the medical specialty certification process may therefore consist of additional longitudinally administered assessments (e.g. workplace-based assessments).

This approach has seen increasing support in so-called programmatic assessments of competency-based medical education (CBME). In this system-based approach to assessment design, pass-fail decisions are based on a portfolio containing datapoints created by multiple assessors and assessments [90, 98]. The time of examination gets decoupled from the time of a high-stake decision such as promotion or graduation [99]. The aim is for candidates to gain valuable information from both a mentor's critical feedback, support, and self-reflection without such programs becoming overly bureaucratic or time-consuming [100, 101]. A combination of longitudinally repeated workplace-based assessments and structured examinations as summarized in this article seems most promising in supporting this goal as well as providing crucial data points for the high-stakes decision on qualification for unsupervised practice.

Since most papers analyzed in this review focus exclusively on one aspect of the exam, it is often not possible to comprehensively evaluate the entire specialty certification exam. Few studies look at multiple examination formats and compare them [24, 27, 35, 41, 54, 69, 86–88]. Strengths and weaknesses identified with just one assessment method may therefore be compensated for in another part of the exam without this effect being accounted for in the literature.

It is possible that, due to the time span under consideration, examination formats have changed in the time since the included studies have been conducted, and certain critiques expressed in a paper may have already been incorporated into practice. This is the case with the Membership of the Royal College of General

Practitioners (MRCGP UK) exam, which was used to illustrate the literature on a particularly well researched medical specialty certification exam. At the time of study, this consisted of a written exam made up of free text answers, a multiple-choice question exam, an oral exam and a video section examining consultation skills.

The RCGP has since decided to change the formats to further improve the examination. Due to potential problems pertaining to validity and reliability – particularly inter-rater reliability – the use of oral examinations has been discontinued in many countries including the UK in favor of more clearly structured examination formats. The written exam has been complemented with an OSCE-based Clinical Skills Assessment (CSA) and a Workplace Based Assessment (WPBA). In the CSA, patients are played by trained and calibrated actors which allows for the simulation of real-life consultations [102]. The goal of the WPBA is to evaluate candidates in their day-to-day practice and provide constructive feedback as well as specifically assess aspects of professional behavior that are difficult to measure using only the written exam and the CSA [103]. Further adaptations were introduced on a temporary basis due to the Covid pandemic [104]. Exactly how the current version of the MRCGP exam compares to the previous examination format on various assessment criteria has not been shown in the existing literature.

Although the proven validity and reliability of OSCE-style examination formats has increased their attractiveness to institutions around the world, a possible downside may relate to their acceptability. "Examiners generally do not like structured assessments" due to the lack of spontaneity and flexibility to adapt the assessment to the abilities of the candidate [69]. Nevertheless, the current combination of examination formats arguably allows for a comparatively comprehensive candidate assessment: written exams test knowledge, can be highly standardized and are easily feasible, OSCEs fulfill standardization requirements while allowing for an assessment of the "shows how" level of test learning, and WPBAs complement these methods by providing more realistic, personalized assessment data.

Strengths & limitations

With search terms covering a variety of possible synonyms of medical specialty certification exams, this review provides the most extensive and up-to-date overview thus far, allowing for an accurate picture of current medical specialty certification exams that have been scientifically evaluated in regard to any of the Ottawa Criteria globally. Together, these seven criteria cover vital aspects of assessment quality.

However, we find many exams have yet to be scientifically analyzed according to any of the Ottawa Quality Criteria. This means some countries and medical specialties are not included in this review. We find it is common for specialty certification exams or different examination formats to be scientifically studied across only a select few criteria or only pertaining to part of the examination. An overall quality ranking leading to a clear recommendation regarding which exam best achieves all seven Ottawa Criteria of Good Assessment could not be supported by the current literature regarding specialty certification exams.

Another limitation of this review is that literature published in languages other than English or exclusively in databases not included in our search is not included in this review. This disadvantages countries where the primary language is not English and may partially explain the predominance of literature about exams based in the UK, USA and other anglophone countries in our findings.

Implications for practice

This systematic literature review provides an overview of medical specialty certification exams, the respective examination methods used, and their evaluation in respect to the Ottawa Criteria. It can thus assist those looking to improve the current specialty certification exams by showcasing the strengths and weaknesses of existing exams. Based on the findings of the papers presented in this systematic review, we can build upon the research most relevant to our medical specialty and learn from the strengths and weaknesses highlighted in examination formats studied in other countries. Certifying bodies looking to expand their current set of examination methods can find tried and tested methods in the research presented here. By collating the published research, this review can also guide readers deciding which specialty certification exams to accept in their jurisdiction. Finally, it offers an index of the leading researchers in this area, serving those looking to further collaborate or study a specific certification exam in respect to the seven Ottawa Criteria.

Implications for future research

Further research should summarize how well exams fulfill all Ottawa Criteria and compare them accordingly. What is the best examination method to use in resource-constrained settings? Which medical specialty manages to test its candidates most reliably? And overall, regarding all seven Ottawa Criteria, what's the best way to organize a medical certification exam? These kinds of research inquiries seem promising as they reflect the literature gaps highlighted in this review. Numerous studies comparing certified to non-certified doctors exist

[105–107], yet studies linking the examination formats to the subsequent performance of certified compared to non-certified physicians would be more useful in deciding how best to structure a specialty certification exam. For instance, natural experiments when certifying bodies update their practices or cohort studies following doctors certified using different examination methods could look at varying outcomes in patient safety. Case reports highlighting the use of innovative new examination formats may also offer potential improvements to the established techniques. Further research should fill the gaps highlighted in this review regarding the examinations, countries and the Ottawa Criteria not yet studied, to allow for a holistic comparison across examinations.

Further research should use the seven Ottawa Criteria to focus on medical specialty certification exams in more non-English-speaking countries and a wider variety of specialties. They should make use of additional sources such as grey literature internal to certifying institutions and expert interviews to shed insight into less frequently studied Criteria such as feasibility. Establishing a common nomenclature which covers the pre-requisites, assessment methods and consequences of medical specialty certification exams would make future comparisons more straight-forward. Although the general quality of the studies we found was good, most of the current research analyzed only a fraction of the entire exam and did not compare different examinations. These approaches should be pursued to allow for a more comprehensive evaluation and better guide recommendations for future practice.

Overall, despite the increased interest over the past few decades outlined above, there continues to be an urgent need for more publicly available research to return the trust which the public places in the certification process of medical doctors.

Conclusion

The past twenty years have seen a growing interest in the topic of patient safety and effective medical specialty certification exams. This is reflected in a growing number of studies analyzing medical specialty certification exams covering a larger variety of medical specialties, countries, and Ottawa Criteria. Medical specialty certification exams vary significantly between countries and are constantly adapted to changing circumstances through new examination formats. Due to their implications for patient safety, rising public scrutiny over medical self-regulation and their impact on candidate's career opportunities, it is of paramount importance they be supported by a large body of evidence which demonstrates fulfillment of all seven Ottawa Criteria of good assessment. Due to their reliance on multiple assessment methods and data-points, aspects of programmatic assessment suggest a promising way forward in the development of effective medical specialty certification exams. To confirm and expand on these results, future research should focus on examinations held outside the Anglosphere, analyses of entire certification exams, and comparisons across examination methods.

Abbreviations

ABIM	American Board of Internal Medicine
CBME	Competency-Based Medical Education
CEX	Clinical Evaluation Exercise
CSA	Clinical Skills Assessment
GP	General Practitioner
MCP	Multiple Choice Paper
MCQ	Multiple Choice Question
MERSQI	Medical Education Research Study Quality Instrument
MRCGP exam	Membership of the Royal College of General Practitioners exam
OPTION scale	'Observing patient involvement' scale
OSCE	Objective Structured Clinical Examination
PICOS	Population, Intervention, Comparison, Outcomes and Study
RCGP	Royal College of General Practitioners
SBA	Single Best Answer
WPBA	Workplace Based Assessment

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-023-04600-x.

Additional file 1. Search terms.

Additional file 2. Overview of studies

Acknowledgements

We would like to thank Sina Riz à Porta for her excellent work editing the manuscript. We would also like to acknowledge both anonymous reviewers and Editorial Board Member Giampiera Bulfone for their constructive comments.

Authors' contributions

Original idea and study design by SH. DS created and implemented the search strategy, was responsible for graphs and formatting and wrote the present systematic review. Screening via inclusion and exclusion criteria was done by DS and NW in two rounds, with AL serving as a final judge in cases of disagreement. AL and DS designed the methodological quality and relevance criteria. DS and NW scored all included studies on these six criteria, categorized them by medical specialty, country, and examination method, summarized them and extracted relevant content regarding the seven Ottawa Criteria from eligible studies in Additional file 2. AL and SH both provided valuable inputs on the entire manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is included within the article (and its additional files).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
SH is an Editorial Board Member of BMC Medical Education. Otherwise, the authors declare that they have no competing interests.

Author details
[1]University of Bern, Institute for Medical Education, Bern, Switzerland.

Received: 28 May 2023 Accepted: 18 August 2023
Published online: 30 August 2023

References

1. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Qual Saf Health Care. 2008;17(3):216–23.
2. Memon MA, Joughin GR, Memon B. Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. Adv Health Sci Educ Theory Pract. 2010;15(2):277–89.
3. Thiessen N, Fischer MR, Huwendiek S. Assessment methods in medical specialist assessments in the DACH region - overview, critical examination and recommendations for further development. GMS J Med Educ. 2019;36(6):Doc78–Web.
4. Wijnen-Meijer M, Burdick W, Alofs L, Burgers C, Ten Cate O. Stages and transitions in medical education around the world: Clarifying structures and terminology. Med Teach. 2013;35(4):301–7.
5. Cassel CK, Holmboe ES. Professionalism and accountability: the role of specialty board certification. Trans Am Clin Climatol Assoc. 2008;119:295–303 (discussion 303-304).
6. Sharp LK, Bashook PG, Lipsky MS, Horowitz SD, Miller SH. Specialty board certification and clinical outcomes: The missing link. Acad Med. 2002;77(6):534–42.
7. Chen J, Rathore SS, Wang Y, Radford MJ, Krumholz HM. Physician board certification and the care and outcomes of elderly patients with acute myocardial infarction. J Gen Intern Med. 2006;21(3):238–44.
8. Prystowsky JB, Bordage G, Feinglass JM. Patient outcomes for segmental colon resection according to surgeon's training, certification, and experience. Surgery. 2002;132(4):663–70 (discussion 670-662).
9. Reid RO, Friedberg MW, Adams JL, McGlynn EA, Mehrotra A. Associations between physician characteristics and quality of care. Arch Intern Med. 2010;170(16):1442–9.
10. Lipner RS, Hess BJ, Phillips RL Jr. Specialty board certification in the United States: issues and evidence. J Contin Educ Health Prof. 2013;33(Suppl 1):S20–35.
11. Institute of Medicine Committee on Quality of Health Care in A. In: To Err is Human: Building a Safer Health System. edn. Edited by Kohn LT, Corrigan JM, Donaldson MS. Washington (DC): National Academies Press (US). Copyright 2000 by the National Academy of Sciences. All rights reserved.; 2000.
12. Weiss KB. Future of board certification in a new era of public accountability. J Am Board Fam Med. 2010;23(Suppl 1):S32–39.
13. McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA. The quality of health care delivered to adults in the United States. New Engl J Med. 2003;348(26):2635–45.
14. James JT. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care. J Patient Saf. 2013;9(3):122–8.
15. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ. 2016;353:i2139.
16. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206–14.
17. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D. 2018 Consensus framework for good assessment. Med Teach. 2018;40(11):1102–9.
18. Spike N. Is medical postgraduate certification improving health outcomes? Med Educ. 2002;36:7–8.
19. Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ. 2002;36(1):73–91.
20. Herb U, Geith U. Kriterien der qualitativen Bewertung wissenschaftlicher Publikationen: Befunde aus dem Projekt visOA. Information - Wissenschaft & Praxis. 2020;71(2–3):77–85.
21. Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med. 2015;90(8):1067–76.
22. Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation-based accreditation into the anesthesiology Israeli national board exams. Israel Med Assoc J. 2006;8(10):728–33.
23. Sun H, Warner DO, Patterson AJ, Harman AE, Rathmell JP, Keegan MT, Dainer RJ, McLoughlin TM, Fahy BG, MacArio A. The American Board of Anesthesiology's Standardized Oral Examination for Initial Board Certification. Anesth Analg. 2019;129(5):1394–400.
24. Warner DO, Lien CA, Wang T, Zhou Y, Isaak RS, Peterson-Layne C, Harman AE, Macario A, Gaiser RR, Suresh S, et al. First-Year Results of the American Board of Anesthesiology's Objective Structured Clinical Examination for Initial Certification. Anesth Analg. 2020;23:1412–8.
25. Warner DO, Isaak RS, Peterson-Layne C, Lien CA, Sun H, Menzies AO, Cole DJ, Dainer RJ, Fahy BG, Macario A, et al. Development of an objective structured clinical examination as a component of assessment for initial board certification in anesthesiology. Anesth Analg. 2020;130(1):258–64.
26. Gali A, Roiter H, de Mollein D, Swieszkowskimtsac S, Atamañuk N, Guerreromtsac AA, Grancellimtsac H, Barero C. Evaluation of the quality of multiple-choice questions used in cardiology certification and recertification exams. Revista Argentina de Cardiologia. 2011;79(5):419–22.
27. Tan LT, McAleer JJA. The Introduction of Single Best Answer Questions as a Test of Knowledge in the Final Examination for the Fellowship of the Royal College of Radiologists in Clinical Oncology. Clin Oncol. 2008;20(8):571–6.
28. O'Leary F. Simulation as a high stakes assessment tool in emergency medicine. EMA - Emerg Med Australas. 2015;27(2):173–5.
29. Bianchi L, Gallagher EJ, Korte R, Ham HP. Interexaminer agreement on the American Board of Emergency Medicine oral certification examination. Ann Emerg Med. 2003;41(6):859–64.
30. Slovut DP, Saiar A, Gray BH. Endovascular medicine certification 2005–2014: report from the american board of vascular medicine. Vascular Medicine (United Kingdom). 2015;20(3):245–50.
31. Khafagy G, Ahmed M, Saad N. Stepping up of MCQs' quality through a multi-stage reviewing process. Educ Prim Care. 2016;27(4):299–303.
32. Weingarten MA, Polliack MR, Tabenkin H, Kahan E. Variations among examiners in family medicine residency board oral examinations. Med Educ. 2000;34(1):13–7.
33. O'Neill TR, Royal KD, Puffer JC. Performance on the American Board of Family Medicine (ABFM) certification examination: are superior test-taking skills alone sufficient to pass? J Am Board Fam Med. 2011;24(2):175–80.
34. O'Neill TR, Peabody MR, Stelter KL, Puffer JC, Brady JE. Validating the Test Plan Specifications for the American Board of Family Medicine's Certification Examination. J Am Board Fam Med. 2019;32(6):876–82.
35. Greco M, Spike N, Powell R, Brownlea A. Assessing communication skills of GP registrars: a comparison of patient and GP examiner ratings. Med Educ. 2002;36(4):366–76.
36. Munro N, Denney ML, Rughani A, Foulkes J, Wilson A, Tate P. Ensuring reliability in UK written tests of general practice: the MRCGP examination 1998–2003. Med Teach. 2005;27(1):37–45.
37. Simpson RG, Ballard KD. What is being assessed in the MRCGP oral examination? A qualitative study. Br J Gen Pract. 2005;55(515):430–6.
38. Sandars J, Coughlin S, Foulkes J. The assessment of skills in evidence-based medicine: The MRCGP examination approach. Educ Prim Care. 2004;15:550–63.
39. Dixon H. The multiple-choice paper of the MRCGP examination: a study of candidates' views of its content and effect on learning. Educ Prim Care. 2005;16(6):655–62.
40. Siriwardena AN, Edwards AGK, Campion P, Freeman A, Elwyn G. Involve the patient and pass the MRCGP: investigating shared decision making

in a consulting skills examination using a validated instrument. Br J Gen Pract. 2006;56(532):857–62.
41. Dixon H. Candidates' views of the MRCGP examination and its effects upon approaches to learning: a questionnaire study in the Northern Deanery. Educ Prim Care. 2003;14:146–57.
42. Wass V, Wakeford R, Neighbour R, Van Der Vleuten C. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ. 2003;37(2):126–31.
43. Dixon H, Blow C, Milne P, Siriwardena N, Milne H, Elfes C. Quality assurance of the Applied Knowledge Test (AKT) of the MRCGP examination - an immediate post-test questionnaire evaluation of the candidates' views. Educ Prim Care. 2015;26(4):223–32.
44. Partridge J. Feedback from candidates sitting the written module of the Membership of the Royal College of General Practitioners examination in Spring 2006: a satisfactory conclusion. Educ Prim Care. 2008;19(2):165–72.
45. Dixon H, Blow C, Irish B, Milne P, Siriwardena AN. Evaluation of a postgraduate examination for primary care: perceptions and performance of general practitioner trainers in the multiple choice paper of the Membership Examination of the Royal College of General Practitioners. Educ Prim Care. 2007;18(2):165–72.
46. Bourque J, Skinner H, Dupré J, Bacchus M, Ainslie M, Ma IWY, Cole G. Performance of the Ebel standard-setting method in spring 2019 royal college of physicians and surgeons of canada internal medicine certification examination consisted of multiple-choice questions. J Educ Eval Health Prof. 2020;17:12–Web.
47. Chierakul N, Danchaivijitr S, Kontee P, Naruman C. Reliability and validity of long case and short case in internal medicine board certification examination. J Med Assoc Thai. 2010;93(4):424–8.
48. McManus IC, Mooney-Somers J, Dacre JE, Vale JA. Reliability of the MRCP(UK) Part I Examination, 1984–2001. Med Educ. 2003;37(7):609–11.
49. McManus IC, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. BMC Med Educ. 2013;13:103.
50. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 2006;6:42. https://rdcu.be/dkHg6.
51. Atsawarungruangkit A: Relationship of residency program characteristics with pass rate of the American Board of Internal Medicine certifying exam. Med Educ Online. 2015;20(1):28631.
52. Marques TR, Lains I, Martins MJ, Goiana-Da-Silva F, Sampaio F, Pessanha I, Fernandes DH, Brandao M, Pinto Teixeira P, de Oliveira SM, et al. Evaluation of the medical board exam in Portugal. Acta Med Port. 2018;31(11):670–9.
53. Burch VC, Norman GR. Turning words into numbers: establishing an empirical cut score for a letter graded examination. Med Teach. 2009;31(5):442–6.
54. Burch VC, Norman GR, Schmidt HG, van der Vleuten CP. Are specialist certification examinations a reliable measure of physician competence? Adv Health Sci Educ. 2008;13(4):521–33.
55. Cookson J. A critique of the specialty certificate examinations of the Federation of Royal Colleges of Physicians of the UK. Clin Med J R Coll Phys Lond. 2010;10(2):141–4.
56. Mucklow J. Development and implementation of the specialty certificate examinations. *Clinical Medicine*. J R Coll Phys Lond. 2011;11(3):235–8.
57. Raddatz MM, Royal KD, Pennington J. Evaluating the Systematic Validity of a Medical Subspecialty Examination. Online Submission Paper presented at the Midwestern Educational Research Association Annual Meeting. 2012.
58. Lunz ME, Bashook PG. Relationship between candidate communication ability and oral certification examination scores. Med Educ. 2008;42(12):1227–33.
59. Houston JE, Myford CM. Judges' perception of candidates' organization and communication, in relation to oral certification examination ratings. Acad Med. 2009;84(11):1603–9.
60. Mathysen DG, Aclimandos W, Roelant E, Wouters K, Creuzot-Garcher C, Ringens PJ, Hawlina M, Tassignon MJ. Evaluation of adding item-response theory analysis for evaluation of the European Board of Ophthalmology Diploma examination. Acta Ophthalmol. 2013;91(7):e573–577.
61. Mathysen DGP, Aclimandos W, Roelant E, Wouters K, Creuzot-Garcher C, Ringens PJ, Hawlina M, Tassignon MJ. History and future of the European Board of Ophthalmology Diploma examination. Acta Ophthalmol. 2013;91(6):589–93.
62. Chow R, Zhang L, Soong IS, Mang OWK, Lui LCY, Wong KH, Siu SWK, Lo SH, Yuen KK, Yau YSH, et al. Inter-rater Reliability of Examiners in the Hong Kong College of Radiologists' Palliative Medicine Oral Examination. Hong Kong J Radiol. 2017;20(3):232–6.
63. Althouse LA, Du Y, Ham HP. Confirming the validity of the general pediatrics certification examinations: a practice analysis. J Pediatr. 2009;155(2):155–156.e151.
64. Emadzadeh A, Ravanshad Y, Makarem A, Azarfar A, Ravanshad S, Aval SB, Mehrad-Majd H, Alizadeh A. Challenges of OSCE national board exam in Iran from participants' perspective. Electronic Phys [Electronic Resource]. 2017;9(4):4195–201.
65. Raddatz MM, Robinson LR. Demonstrating construct validity of the american board of physical medicine and rehabilitation part i examination: an analysis of dimensionality. PM and R. 2017;9(10):985–9.
66. Tibbo P, Templeman K. The RCPSC oral examination: patient perceptions and impact on participating psychiatric patients. Can J Psychiat. 2004;49(7):480–6.
67. Tong E, Spooner M, Van Delden O, Uberoi R, Sheehan M, O'Neill DC, Lee M. The European board of interventional radiology examination: a cross-sectional web-based survey. Cardiovasc Intervent Radiol. 2018;41(1):21–6.
68. Yeung A, Booth TC, Larkin TJ, McCoubrie P, McKnight L. The FRCR 2B oral examination: is it reliable? Clin Radiol. 2013;68(5):466–71.
69. Yeung A, Booth TC, Jacob K, McCoubrie P, McKnight L. The FRCR 2B examination: a survey of candidate perceptions and experiences. Clin Radiol. 2011;66(5):412–9.
70. Yang JC, Wallner PE, Becker GJ, Bosma JL, Gerdeman AM. Reliability of oral examinations: Radiation oncology certifying examination. Pract Radiat Oncol. 2013;3(1):74–8.
71. Kerridge WD, Gunderman RB. The validity and timing of the ABR core exam. Acad Radiol. 2016;23(9):1176–9.
72. Yang JC, Gerdeman AM, Becker GJ, Bosma JL. American Board of Radiology diagnostic radiology initial qualifying (written) examinations. AJR Am J Roentgenol. 2010;195(1):10–2.
73. Pascual-Ramos V, Guilaisne Bernard-Medina A, Flores-Alvarado DE, Portela-Hernández M, Maldonado-Velázquez MDR, Jara-Quezada LJ, Amezcua-Guerra LM, Rubio-Judith López-Zepeda NE, Álvarez-Hernandez E, Saavedra MÁ, et al. The method used to set the pass mark in an objective structured clinical examination defines the performance of candidates for certification as rheumatologists. Reumatologia Clinica. 2018;14(3):137–41.
74. Smith JD, Prideaux D, Wolfe CL, Wilkinson TJ, Sen Gupta T, DeWitt DE, Worley P, Hays RB, Cowie M. Developing the accredited postgraduate assessment program for Fellowship of the Australian College of Rural and Remote Medicine. Rural Remote Health. 2007;7(4):805.
75. Beasley SW, Wannan C, Hardware N. Justification and implications of the introduction of an expanded Close Marking System for the Fellowship Examination. ANZ J Surg. 2013;83(6):444–7.
76. De Montbrun S, Roberts PL, Satterthwaite L, Macrae H. Implementing and evaluating a national certification technical skills examination. Ann Surg. 2016;264(1):1–6.
77. Lineberry M, Park YS, Hennessy SA, et al. The Fundamentals of Endoscopic Surgery (FES) skills test: factors associated with first-attempt scores and pass rate. Surg Endosc. 2020;34:3633–43. https://doi.org/10.1007/s00464-020-07690-6.
78. Motoyama S, Yamamoto H, Miyata H, Yano M, Yasuda T, Ohira M, Kajiyama Y, Toh Y, Watanabe M, Kakeji Y, et al. Impact of certification status of the institute and surgeon on short-term outcomes after surgery for thoracic esophageal cancer: evaluation using data on 16,752 patients from the National Clinical Database in Japan. Esophagus. 2020;17(1):41–9.
79. Montbrun d, Lynn S. High Stakes Technical Skill Assessments in Surgery: Development, Implementation and Predicting Performance. 2017. p. 2017.

80. Crisostomo AC. The Effect of Standardization on the Reliability of the Philippine Board of Surgery Oral Examinations. J Surg Educ. 2011;68(2):138–42.
81. Rhodes RS, Biester TW, Bell RH Jr, Lewis FR Jr. Assessing Surgical Knowledge: a Primer on the Examination Policies of the American Board of Surgery. J Surg Educ. 2007;64(3):138–42.
82. Cundy P. Examining the orthopaedic examiners: Reliability of the Part 2 Orthopaedic Clinical Fellowship Examination in Australia. ANZ J Surg. 2012;82(9):607–11.
83. Hohmann E, Tetsworth K. Fellowship exit examination in orthopaedic surgery in the commonwealth countries of Australia, UK, South Africa and Canada. Are they comparable and equivalent? A perspective on the requirements for medical migration. Med Educ Online. 2018;23(1):1537429–7.
84. Gillis ME, Scott SA, Richardson CG, Oxner WM, Gauthier L, Wilson DA, Glennie RA. Developing and assessing the feasibility of implementing a Surgical Objective Structured Clinical Skills Examination (S-OSCE). J Surg Educ. 2020;77(4):939–46.
85. Ullmann Y, Fodor L, Meilick B, Eshach H, Ramon Y, Meilick A. The oral board examination for plastic surgery: Seeking a better way. Med Teach. 2006;28(4):360–4.
86. Dwyer T, Chahal J, Murnaghan L, Theodoropoulos J, Cheung J, McParland A, Ogilvie-Harris D. Development of a certification examination for orthopedic sports medicine fellows. Can J Surgery Journal canadien de chirurgie. 2020;63(2):E110–7.
87. Payne SR, Pickard RS, O'Flynn KJ, Winton EP. Does the Intercollegiate Specialty Examination in urology (FRCS Urol) assess the breadth of the urology syllabus? Br J Med Surg Urol. 2011;4(4):139–47.
88. Hohmann E, Tetsworth K. Fellowship exit examination in orthopaedic surgery in the commonwealth countries of Australia, UK, South Africa and Canada. Are they comparable and equivalent? A perspective on the requirements for medical migration. Med Educ Online. 2018;23(1):1537429.
89. Practitioners RCoG. The RCGP Curriculum. Being a General Practitioner. 2019.
90. Lockyer J, Carraccio C, Chan M-K, Hart D, Smee S, Touchie C, Holmboe ES, Frank JR. Core principles of assessment in competency-based medical education. Med Teach. 2017;39(6):609–16.
91. de Montbrun S, Louridas M, Grantcharov T. Passing a technical skills examination in the first year of surgical residency can predict future performance. J Grad Med Educ. 2017;9(3):324–9.
92. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63–67.
93. Van Der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. Med Educ. 2005;39(3):309–17.
94. Witheridge A, Ferns G, Scott-Smith W. Revisiting Miller's pyramid in medical education: the gap between traditional assessment and diagnostic reasoning. Int J Med Educ. 2019;10:191–2.
95. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med. 1998;129(1):42–8.
96. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach. 2013;35(7):564–8.
97. Kogan JR, Holmboe E. Realizing the Promise and Importance of Performance-Based Assessment. Teach Learn Med. 2013;25(sup1):S68–74.
98. Misra S, Iobst WF, Hauer KE, Holmboe ES. The importance of Competency-Based Programmatic Assessment in Graduate Medical Education. J Grad Med Educ. 2021;13(2s):113–9.
99. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, van Tartwijk J. A model for programmatic assessment fit for purpose. Med Teach. 2012;34(3):205–14.
100. Driessen EW, van Tartwijk J, Govaerts M, Teunissen P, van der Vleuten CP. The use of programmatic assessment in the clinical workplace: a Maastricht case report. Med Teach. 2012;34(3):226–31.
101. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve tips for programmatic assessment. Med Teach. 2015;37(7):641–6.
102. MRCGP Clinical Skills Assessment (CSA) https://www.rcgp.org.uk/training-exams/mrcgp-exam/mrcgp-clinical-skills-assessment-csa.aspx.
103. WPBA assessments. https://www.rcgp.org.uk/training-exams/training/workplace-based-assessment-wpba/assessments.aspx.
104. MRCGP: Recorded Consultation Assessment (RCA). https://www.rcgp.org.uk/mrcgp-exams/recorded-consultation-assessment.
105. Silber JH, Kennedy SK, Even-Shoshan O, Chen W, Mosher RE, Showan AM, Longnecker DE. Anesthesiologist board certification and patient outcomes. Anesthesiology. 2002;96(5):1044–52.
106. Wallace A, McFarland BH, Selvam N, Sahota G. Quality of care provided by board-certified versus non-board-certified psychiatrists and neurologists. Acad Med. 2017;92(1):108–15.
107. Norcini J, Lipner R, Kimball H. The certification status of generalist physicians and the mortality of their patients after acute myocardial infarction. Acad Med. 2001;76(10 Suppl):S21–23.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC