

REVIEW ARTICLE

 OPEN ACCESS Check for updates

The pattern of reporting and presenting validity evidence of extended matching questions (EMQs) in health professions education: a systematic review

Mohamed H. Taha ^a, Hosam Eldeen Elsadig Gasmalla Mohammed ^{b*}, Mohamed Elhassan Abdalla ^c,
Muhamad Saiful Bahri Yusoff ^d, Mohd Kamal Mohd Napiah^e and Majed M. Wadi ^{ff}

^aCollege of Medicine and Medical Education Centre, University of Sharjah, Sharjah, United Arab Emirates; ^bAnatomy Department, Al-Neelain University, Khartoum, Sudan; ^cSchool of Medicine, University of Limerick, Limerick, Ireland; ^dMedical Education Department, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, Malaysia; ^eLibrary Hamzah Sendut, Universiti Sains Malaysia, Penang, Malaysia; ^fMedical Education Department, College of Medicine, Qassim University, Qassim, Saudi Arabia

ABSTRACT

The Extended matching Questions (EMQs), or R-type questions, are format of selected-response. The validity evidence for this format is crucial, but there have been reports of misunderstandings about validity. It is unclear what kinds of evidence should be presented and how to present them to support their educational impact. This review explores the pattern and quality of reporting the sources of validity evidence of EMQs in health professions education, encompassing content, response process, internal structure, relationship to other variables, and consequences. A systematic search in the electronic databases including MEDLINE via PubMed, Scopus, Web of Science, CINAHL, and ERIC was conducted to extract studies that utilize EMQs. The framework for a unitary concept of validity was applied to extract data. A total of 218 titles were initially selected, the final number of titles was 19. The most reported pieces of evidence were the reliability coefficient, followed by the relationship to another variable. Additionally, the adopted definition of validity is mostly the old tripartite concept. This study found that reporting and presenting validity evidence appeared to be deficient. The available evidence can hardly provide a strong validity argument that supports the educational impact of EMQs. This review calls for more work on developing a tool to measure the reporting and presenting validity evidence.

ARTICLE HISTORY

Received 5 November 2022
Revised 16 October 2023
Accepted 30 September 2024

KEYWORDS

Extended matching questions; extended matching items; R-type MCQ; EMQs; EMIs; health professions education; validity; reliability

Introduction

Extended matching questions (EMQs) were introduced in the 1990s as an additional form of multiple-choice questions. It was known as R-Type MCQs [1]. EMQs possess distinctive characteristics that make them a valuable tool in medical education and assessment [2]. They present clinical scenarios in a standardized patient history format, compelling examinees to make decisions and select the most appropriate option(s) from a list that encompasses a spectrum of possible diagnoses, treatments, or management strategies [1]. Unlike traditional multiple-choice questions, EMQs offer a wider array of choices, often ranging from 5 to 15 options, enabling a more comprehensive evaluation of the examinee's clinical knowledge and reasoning abilities [1]. Furthermore, EMQs require matching the clinical scenarios with the suitable option(s), with the possibility of multiple correct answers [1]. These questions are designed to replicate

real-life clinical situations, fostering the application of theoretical knowledge to practical scenarios [1,2]. They also assess an examinee's skill in prioritizing actions based on the urgency and clinical significance of various factors [2]. Particularly, EMQs excel in evaluating an individual's aptitude for generating and refining a differential diagnosis, a pivotal skill in medical decision-making [2]. Additionally, by demanding the consideration of multiple options and the application of clinical reasoning, EMQs effectively gauge critical thinking skills, a fundamental asset for healthcare professionals [1]. Example EMQ is shown in box 1.

The benefits of utilizing EMQs in medical education can be better comprehended by juxtaposing them with two other common question formats: traditional MCQs (A-Type MCQs) and short answer questions. A-Type MCQs, featuring four or five options, are widely recognized for their reliability, objectivity in scoring, and

CONTACT Majed M. Wadi  m.wadi@qu.edu.sa; mwadi001@dundee.ac.uk  Medical Education Department, College of Medicine, Qassim University, Saudi Arabia; Muhammad Saiful Bahri Yusoff  msaiful_bahri@usm.my  Medical Education Department, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, Malaysia

*Present affiliation of Hosam Eldeen Elsadig Gasmalla Mohammed is Warwick Medical School, University of Warwick, Coventry, United Kingdom

^{ff}Present affiliation of Majed M. Wadi is Centre for Medical Education, School of Medicine, University of Dundee, Dundee, Scotland, UK

Theme: Skin disease

Options:

- (A) Acne vulgaris
- (B) Herpes zoster
- (C) Infective endocarditis
- (D) Lupus erythematosus
- (E) Measles
- (F) Vitiligo

Lead-in question: Based on the list of diagnoses above, select the most appropriate one for each of the given scenarios below:

- (1) A 15-year-old boy presented with fever and skin lesions. Examination revealed fluid-filled skin lesions less than 0.5 cm in diameter, distributed on his left side of the chest along the 3rd rib. The patient also mentioned a burning or tingling sensation in the affected area before the appearance of the lesions.

Answer: B) Herpes zoster

- (2) An 18-year-old girl presented with painful skin lesions on her face. Examination revealed pus-filled lesions less than 0.5 cm in diameter. The patient mentioned a recent history of increased oiliness of her facial skin and the presence of comedones (blackheads and whiteheads) on her nose and cheeks.

Answer: A) Acne vulgaris

Box (1) example of extended matching questions (Adopted with permission from Gasmalla et al. (2023) [3].

scalability, making them suitable for assessing a large number of students simultaneously. However, doubts have been raised regarding their ability to effectively assess higher-order cognitive skills and clinical reasoning [4,5], in addition to concerns regarding the potential for correct guessing [6]. On the other hand, free-response or open-ended questions have the capacity to evaluate higher cognitive functions and clinical reasoning in-depth. Nevertheless, their limited content coverage diminishes the validity of test score interpretations, and logistical challenges associated with subjective scoring further complicate their use [7]. EMQs, in the form of R-Type MCQs, address the drawbacks of traditional MCQs by minimizing the guessing effect through the incorporation of a larger number of options. Moreover, the objectivity of scoring in R-Type MCQs overcomes the limitations of the subjective and time-consuming scoring process involved with open-ended questions [7,8]. In a study by van Bruggen, Manrique-van Woudenberg [9], various question formats, including EMQs (R-Type MCQs), Script Concordance Test (SCT), Comprehensive Integrative Puzzle (CIP), Modified Essay Question (MEQ), Short Answer Question (SAQ), Long Menu Question (LMQ), traditional MCQs (A-Type MCQs), Multiple True/False Question (MTFQ), and Virtual Patient (VP), were explored to recommend the most effective assessment tool for clinical reasoning. The authors concluded that a combination of CIP and EMQs (R-Type MCQs) exhibited favorable validity, reliability, and discriminatory effects [9].

Since the 2000s, EMQs (R-Type MCQs) have gained prominence in high-stakes exams, notably being implemented in the Royal Australian and New Zealand College of Psychiatrists examinations in 2004 [10] and in the Royal College of Obstetricians and Gynaecologists examinations in the United Kingdom since 2006 [4,11]. However, this rapid paradigm shift has raised questions

regarding the validity of these MCQ formats in medical assessment.

Understanding the concept of assessment validity is paramount for educators, as it confers meaning to the assessment process [12]. The classical validity framework, encompassing content, criterion (concurrent and predictive), and construct validity, has evolved into the contemporary standard framework introduced in the 1990s [12]. This updated model posits validity as a unitary concept that must be supported by five sources of validity evidence: content, response process, internal structure, relationship to other variables, and consequences [13,14]. Validity, thus, refers to the extent to which evidence supports the appropriateness of test score interpretations, and validation is an ongoing process of collecting and accumulating such evidence [14]. This holistic perspective of validity holds a sacred place in the realm of assessment, underpinning the credibility and defensibility of the results obtained [12].

Studies employing EMQs, and all other assessment tools, must present sufficient evidence of validity [14,15]. To the best of our knowledge, and based on a piloting study conducted, no systematic reviews exploring the reporting pattern and quality of validity evidence in studies utilizing EMQs have been reported to date. Hence, this systematic review aims to explore the pattern and quality of reporting the sources of validity evidence of EMQs in health professions education, encompassing content, response process, internal structure, relationship to other variables, and consequences, as delineated by Downing [14].

This systematic review aims to address the following questions:

- (1) What is the pattern of validity evidence reported for using EMQs in the assessment

of undergraduate and postgraduate candidates in health professions education?

- (2) What is the quality of reporting validity evidence for using EMQs in the assessment of undergraduate and postgraduate candidates in health professions education?

Methods

In this review, we followed Best Evidence in Medical Education (BEME) guidance [16]. The study protocol was approved, registered and posted to the BEME website. Authors reported the findings in alignment with the STORIES (STructured apprOach to the Reporting In healthcare education of Evidence Synthesis) statement [17] and BEME guidance [16].

Study eligibility/selection criteria

The inclusion criteria are summarized in Table 1. Our search span started from 1990 up to the Sep 2023. The articles' scope was defined as the articles that utilize EMQs introduced to health professions candidates at the undergraduate and postgraduate levels. It must report at least one validity evidence according to the framework explained by Downing [14]. We defined health professions as those of medicine, dentistry, pharmacy, nursing, and allied health care professionals. We excluded any paper that does not meet the criteria.

Search strategy and study identification

We performed a systematic search in five databases, namely, PubMed, Scopus, Web of Science, CINAHL, and ERIC. Our search was performed to extract titles and abstracts (Figure 1). The search strategy included forward and backward search and search of grey literature as well. The search strategy also included a hand-search: The authors have applied the search terms in the search bar of the journals including medical teacher, academic medicine, BMC medical education, Advances in Health Sciences Education, teaching and learning in medicine, and medical education journals.

The search was initially carried out in January 2021 and then again in September 2023,

with both searches spanning from 1990 to the final day of September 2023. No new titles were identified.

Search terms

The search terms were based on 'MeSH Terms' related to the assessment tool under inspection 'Extended matching questions' OR 'Extended matching items' OR 'R-type items' OR 'R-type questions' OR 'EMQs' OR 'EMIs,' and terms related to the population 'health professions education' OR 'medical education' as well as terms related to 'validity' OR 'reliability'. The terms were defined according to the definition of terms (see appendix 1).

Study selection process

Due to the variability of accessibility to databases, this task was divided between authors according to their accessibility and institutional access. Each author performed a systematic search in one database, all titles were retrieved from citation managers and sent to our librarian (MK) in which they were collected and indexed using Mendeley, an excel spreadsheet was created including all the retrieved titles and abstracts.

Three authors (HEG, MW and MT) independently screened the titles and abstracts for all the pooled titles. Disagreements were solved by discussion and consensus was reached for all titles. To this end, the three authors met online several times using ZOOM software for an estimated total of two hours at this stage.

Data extraction

In the piloting phase, all reviewers conducted a meeting to design a template to be used for data extraction. The form was based on the five validity sources [14]. This template included the following variables: study titles, authors, year of publication, study purpose, the used methods, intervention (if there), results, outcomes, and the existence of one or more of the five sources of validity: i) content, ii) response process, iii) internal structure, iv) relation to other variables and v) related consequences, and the decision about the article (to be included or not based

Table 1. Inclusion criteria.

Criteria	Inclusion criteria
Type of paper	original articles (study designs included are controlled trials (randomized and non-randomized), quasi-experimental designs, cohort studies, cross-sectional studies and case-control studies, as well as mixed-method).
Language	included articles published in the English language
Period	1990 – 2023
Focus of study	Focus on extended matching question
Context	Health profession education (Medical, dental, pharmacy, nursing, and other allied health care professionals)
Elements of validity	Clear report of any category of validity as described by Downing [14]

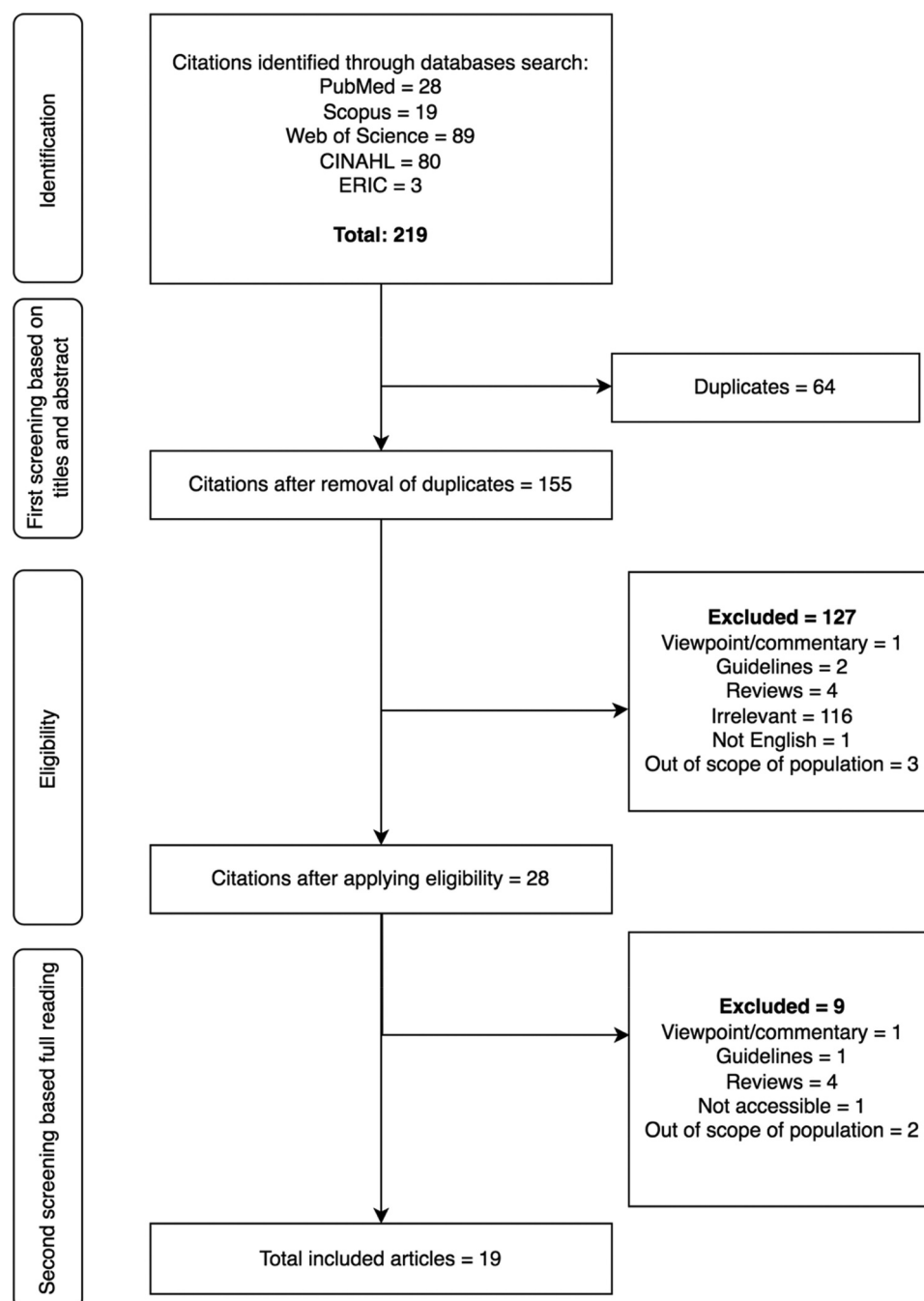


Figure 1. The process of identification and screening of titles.

on full-text screening). The template was pretested by three authors (HEG, MW and MT) independently on nine studies before commencing the review. The authors performed initial data extraction for the selected nine papers to ensure consistency. Then, each author was assigned to extract data from a database to ensure accuracy. Then, each author checked the extracted data from the other two authors. When there was any discrepancy between the authors, the final decision was made by consensus among the three authors in an online meeting using software from Zoom Video Communications, Inc ©2020 version 5.0

In the systematic review, three authors (HEG, MW and MT) independently reviewed the full texts of the included articles (according to the 'study selection process' mentioned above). Then the final list of articles was extracted by the three authors.

Quality assessment

The quality of the included studies was evaluated using the Medical Education Research Study Quality Instrument (MERSQI) [18]. The three authors appraised the quality of each one of the included

studies. The MERSQI was developed to evaluate the quality of methods utilized in experimental, and observational studies based on six domains. Those domains are study design, study sampling 'institutes and response rate', type of data, validity evidence for evaluation instrument scores, data analysis: 'sophistication and appropriateness', and outcome). The validity in the context of educational assessment is a property of the interpretation of test scores, it is not a property of the test (as a data collection tool) itself. This is a common misconception about validity [12].

We also performed a quality assessment using an approach that determines the quality of studies based on the number of reported validity evidence. To the best of our knowledge, all the operational cues definitions of validity evidence were nine as indicated in Table 3. We identified the number of validity evidence reported in each manuscript and used it as an indication of the quality of the manuscript.

Evidence synthesis

This review aimed to describe the pattern of reporting and presenting validity evidence of EMQs in health professions education, and to provide a comprehensive account of the quality of reporting validity evidence, following the five sources of validity that are content, response process, internal structure, relationship to other variables and consequences (see appendix 2), as explained by Downing [14]. A narrative description was produced after summarizing the data from the extraction form and reporting data under the following headings: status of reporting validity evidence, the quality of studies based on reporting validity evidence, and patterns of presenting the reported validity evidence. A visual infographic – figure 2 – was used to present key data, the summary of the patterns of presenting validity evidence in the included studies. Direct quotations were extracted from papers including verbatim examples of patterns of presenting the reported validity evidence are provided in appendix 4. A narrative synthesis [37] with a content analysis approach was used to identify dominant issues across studies by categorizing data, both quantitative and qualitative, into themes [38]. The narrative synthesis method entails 'collating study findings into a coherent textual narrative, with descriptions of the differences in characteristics,' and capturing the validity and relevance to the context where possible [39,40], using qualitative synthesis reporting.

Results

Our search retrieved 219 titles initially, after removal of duplicates ($n = 64$) and screening by title and

abstract the number of titles was 28. The main reasons for excluding titles were: irrelevant to the scope of the review ($n = 116$), the title is a review ($n = 4$), the titles represent a population out of the scope of the review ($n = 3$), the title represents a guideline ($n = 2$), viewpoint/commentary ($n = 1$), and the manuscript is not in English language ($n = 1$). This number was further narrowed into 19 titles following screening by full text. The justifications for exclusion by full-text screening were the study is a viewpoint/commentary ($n = 1$); a guideline ($n = 1$); a review ($n = 4$), out of scope ($n = 2$) or not accessible ($n = 1$) (Figure 1).

Kappa test

Inter-rater reliability analysis was performed between the dependent samples of Rater 1, Rater 2 and Rater 3. For this purpose, the Fleiss Kappa was calculated, which is a measure of the agreement between more than two dependent categorical samples. The Fleiss Kappa showed that there was a moderate agreement between samples Rater 1, Rater 2 and Rater 3 with $\kappa = 0.54$. Fleiss Kappa: 0.54, 95% CI (0.36–0.73) and p -value < 0.001

Included studies

Most of the studies were conducted during the period from 2001 to 2010 ($n = 12$, 63.16%). There were no articles published during the period (1991 – 2000). While in the period from 2011 to 2022 ($n = 7$, 36.84%). Most of the articles were cross-sectional studies ($n = 12$, 63.16%), followed by Cohort studies ($n = 2$, 10.52%), randomised controlled trial ($n = 2$, 10.52%). There was one article ($n = 1$, 5.26%) representing each of the following designs: quasi-experimental, retrospective study design on archival records of a medical undergraduate assessment program and one think aloud protocol.

The validity was defined in most of them, all the manuscripts that defined validity ($n = 12$, 63.16%) adopted the old tripartite concept (i.e., dividing validity into construct, content, criterion, etc.). It was notable that some papers presented the validity evidence without defining the validity itself.

The results are shown in three categories: the first category is the status of reporting validity evidence, here the authors indicate whether evidence is reported or not. The second category is the quality of studies based on reporting validity evidence. The third one is concerned with the patterns of presenting the reported validity evidence, here the already reported evidence is displayed according to the pattern of their presentation.

Table 2. List of included studies and the quality of studies based on the reported validity evidence.

Score of findings related to Validity															
No.	Author(s) (date)	Study title	Study scope	Study design	Population	Content			Internal structure				Relationship to other variables	Consequence	Total of reported operational cues
						Blueprint	Item quality	Qualifications of the developers	Response process	Difficulty index	Discrimination index	Internal consistency			
	Wass et al. (2011) [26]	Composite undergraduate clinical examinations: how should the components be combined to maximize reliability?	A composite examination's overall reliability, correlations between the tests, and effects of different test lengths, items, and weighting on reliability are assessed.	Cross-sectional	Undergraduate medical students	+	-	-	-	+	+	+	+	5	
	Lukić et al.(2001) [34]	Weekly quizzes in extended-matching format as a means of monitoring students' progress in gross anatomy	Monitoring students' progress in gross anatomy by comparing extended-matching quizzes, multiple-choice questions, and oral examinations	Quasi experimental	Undergraduate medical students	-	+	-	-	-	+	+	+	4	
	Beullens et al.(2002) [2]	Are extended-matching multiple-choice items appropriate for a final test in medical education?	Faculty-developed 900 EMQs (R-type MCQs) were evaluated	Cohort study	Undergraduate medical students	+	+	-	+	-	+	+	-	5	
	Basu et al. (2004a) [20]	Competence in the musculoskeletal system: assessing the progression of knowledge through an undergraduate medical course	Assessing the progress in musculoskeletal knowledge attained by medical students during the course by designing and evaluating an assessment procedure.	Cross-sectional	Undergraduate medical students	+	+	+	+	-	+	+	+	7	
	Basu et al. (2004b) [19]	Comparing and contrasting undergraduate competence in musculoskeletal medicine with cardiovascular medicine and neurology	A comparison of knowledge of the musculoskeletal system among final-year Sheffield medical students was conducted.	Cross-sectional	Undergraduate medical students	+	-	-	+	+	+	-	+	5	
	Coderre et al. (2004) [21]	The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts	Investigating whether two different types of papers were capable of assessing problem-solving abilities (standard five-option questions and extended-matching questions).	Cross-sectional	Experts clinicians and undergraduate medical students	+	+	-	-	+	+	+	+	6	
	Bhakta et al. (2005) [24]	Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education	An item response theory analysis was conducted on the results of the exam taken by fourth-year undergraduate medical students in 2011–2002	Cross-sectional	Undergraduate medical students	+	-	+	-	+	+	+	+	6	
	Beullens et al. (2005) [27]	Do extended matching multiple-choice questions measure clinical reasoning?	Researchers compared residents' and students' processes for answering EMQs (R-type MCQs).	Cross-sectional	Undergraduate medical students and postgraduates (residents)	+	+	-	-	-	-	+	+	4	
	Swanson et al. (2008) [31]	Measurement Characteristics of Content-Parallel Single-Best-Answer and Extended-Matching Questions in Relation to Number and Source of Options	A comparison of content-parallel single-best-answer and extended-matching questions concerning options number and source	Cross-sectional	Record of USMLE	-	+	+	+	+	+	-	+	7	
	Beullens et al. (2006) [33]	Diagnostic ability in relation to clinical seminars and extended-matching questions examinations	Determining if clinical reasoning can be improved with problem-solving clinical seminars and if clinical decision-making can be measured by the EMQ examination.	Cross-sectional	Undergraduate medical students	-	-	-	-	-	+	+	-	2	
	Dory et al. (2010) [23]	Is case-specificity content-specific? An analysis of data from extended-matching questions	Content-specificity linked to the variability of performance, was tested using extended-matching items (EMIs)	Cohort study	Postgraduates students	+	+	-	-	-	+	-	+	4	
	Harding et al. (2010) [22]	The performance of junior doctors in applying clinical pharmacology knowledge and prescribing skills to standardized clinical cases	Clinical pharmacology and therapeutics (CPT) knowledge and prescribing skills of foundation year 1 (FY1) doctors are assessed using standardized clinical cases.	Cross-sectional	Undergraduate medical students and postgraduates (residents)	+	+	+	+	-	+	+	+	7	
	Gibson and Leinster (2011) [35]	How do students with dyslexia perform in extended matching questions, short answer questions, and observed structured clinical examinations?	An analysis of how students with dyslexia perform in extended matching questions (EMQs), short answer questions (SAQs), and observed structured clinical examinations (OSCEs) and the implications of those differences.	Randomized controlled trail	Undergraduate medical students	-	-	-	-	-	-	+	-	1	
	D'Antoni et al. (2012) [30]	Psychometric properties and podiatric medical student perceptions of USMLE-style items in a general anatomy course	In the general anatomy course assessment at the New York College of Podiatric Medicine, traditional A-type MCQs and EMQs (R-type MCQs) were investigated in terms of psychometric properties.	Cross-sectional	Undergraduate medical students	-	+	-	-	-	+	+	+	5	
	Fong et al. (2012) [36]	The Validity of the Assessment Methods in an Undergraduate Psychiatry Module Examination among a Sample of Fourth Year Malaysian Medical Students	Clinical, theory and continuous assessment components of an undergraduate Psychiatry Module examination were tested for convergence, discriminant validity, and predictive validity regarding the overall examination performance of fourth-year medical students.	Observational retrospective	Undergraduate medical students	-	-	-	-	-	-	+	-	1	
	Vorstenbosch et al. (2014) [32]	Exploring the Validity of Assessment in Anatomy: Do Images Influence Cognitive Processes Used in Answering Extended Matching Questions?	The use of extended matching questions (EMQs) was investigated. During answering questions with and without images, students' cognitive processes were examined to gain insight into how images affect test item validity.	"think aloud" protocol	Undergraduate medical students	-	+	-	+	-	-	-	-	2	
	Tan et al. (2018) [29]	Evaluating a Bedside Tool for Neuroanatomical Localization with Extended-Matching Questions	EMQ was used as an outcome measure for an experimental study evaluating neuroanatomical localization 14k among second-year students	Randomized controlled trial	Undergraduate medical students	-	+	+	-	-	+	+	-	4	
	Nazim et al. (2019) [28]	Assessing clinical reasoning skills using Script Concordance Test (SCT) and extended matching questions (EMQs): A pilot for urology trainees	Tests of Script Concordance and EMQs (R-type MCQs) were used to assess clinical reasoning ability among urology residents, interns and medical students.	Cross-sectional	Undergraduate medical students and postgraduates (residents)	-	+	+	+	-	+	+	-	5	
	AlShamlan et al. (2020) [25]	Evaluation of Multifaceted Assessment of the Fifth-Year Medical Students in Family Medicine Clerkship, Saudi Arabia Experience	An evaluation of the multifaceted assessment of the family medicine clerkship at Imam Abdulrahman Bin Faisal University and the correlation between them.	Cross-sectional	Undergraduate medical students	+	-	-	+	-	-	+	+	4	

The operational cues listed in Table 2 were adapted from the original framework introduced by Downing [14] following a thorough discussion among the authors.

Status of reporting validity evidence

Based on the operational cues developed by the authors, the number of reported validity evidence in the included studies was denoted (Table 2). The reliability coefficient as a description of item analysis was the most frequent evidence of validity ($n = 14$, 73.68%) to be reported [2,19–25,28–30,33,34]. Similarly, the description and/or mentioning of correlations to another measure were reported with 73.68%. [2,20–22,25–28,30,31,33–36]. Followed by the description and/or mentioning of revision and/or meeting for item quality was the third most frequent evidence of validity to be reported ($n = 16$, 63.16%) [2,19–24,27–31]. Difficulty index [19,24,26,31] was the least frequent evidence ($n = 4$, 21.05%) to be reported (Table 2).

The quality of studies based on reporting validity evidence

The quality of included articles is described with two approaches. The first approach was the quality assessment using Medical Education Research Study Quality Instrument (MERSQI) [18]. The average score based on (MERSQI) was 10.7 out of the maximum score of the instrument which is 18. The quality of included studies ranged from 13.8 (Maximum) to 9.2 (Minimum).

The second approach is determining the quality of studies based on reporting validity evidence. The number of validity evidence reported in each manuscript (Table 2) was used as an indication of the quality of the manuscript. According to this approach, the best quality was presented by about three studies (15.8%), reporting seven out of nine evidence. About half of them reported less than half of the evidence (four and less than four reported evidence) (Table 2).

Patterns of presenting the reported validity evidence

The reported validity evidence among the included studies are categorized according to their five sources [14] into the following headings; a) Content-related sources of validity including blueprinting, quality of contents, and qualification of the developers, b) Response process-related sources of validity, c) Internal structure-related sources of validity including difficulty index estimation, discrimination index estimation, and internal consistency estimation, d) Relationship to other variables and e) consequence. A summary is presented in Table 3 with verbatim examples for each category. The whole thematic analysis and verbatim text are presented in appendix 3.

Content-related sources of validity evidence

Blueprinting

Blueprinting as a source of validity evidence was reported in ten manuscripts ($n = 10$, 52.63%) in which two patterns were presented. The term 'blueprint' was stated clearly in six studies [2,19,20,22,25,26]. In the other pattern the term 'blueprint' wasn't stated, instead, it was inferred in four studies [21,23,24,27] (Table 3).

Quality of contents

The quality of the contents as a source of validity evidence was reported in eight manuscripts. There were five identified patterns of presentation among the reported evidence. The first pattern was reporting a panel meeting [22,24,28]. The second pattern was mentioning that the test items were retrieved from a Q-bank [20,23,27]. The third pattern was stating that the items were developed based on guidelines and textbooks [19,30]. The fourth pattern was indicating that the training of developers was the base items' quality [2]. Moreover, the fifth group was assembled in which there was no certain pattern, [21,29–31], so we named it (other patterns). The indication of the quality of items in this last group was faint.

Qualification of the developers

The reporting of the qualification of the developers as a source of validity evidence was presented in two patterns. In the first pattern, the developers were highlighted as content experts [2,22,28,29]. While in the second pattern the test developers were highlighted as education experts [24,31] (in which the test development was supervised by experts in medical education). However, none of the manuscripts made it clear if the content experts were education experts as well, and vice versa.

Response process-related sources of validity evidence

The reporting of the response process as a source of validity evidence was presented in three patterns. The first one is when procedures related to test organization were stated [2,19,20,22,31]. The second pattern is when procedures related to making candidates familiar with the format were declared [28,29]. Vorstenbosch, Bouter [32] stated the procedures that were meant to assess the response process, and that was considered as a third pattern.

Difficulty index estimation

The reported difficulty index estimation as an internal structure-related source of validity was presented in two patterns. The first one was the estimation of

Table 3. Summary of validity evidence.

Source of validity evidence	Operational cue(s)	Number of papers (%)	Example of verbatim of presenting the reported validity evidence
Content	Blueprint	10 (52.63%) [2,19-27]	"The test was blueprinted against internal and external guidelines. To ensure representative sampling of curriculum content, selected themes were transferred into a blueprinting framework" [19] "This was achieved by using a 2-dimensional blueprint design which matched the derived subject matter against relevant basic and clinical science scenarios important to the PHMO." [20] "The examination consisted of 98 EMQs distributed across eight specialties and 27 themes, each with two to four EMQs. There were 12 Oncology, 12 A&E, 12 Infectious Diseases, 16 Orthopaedics, 8 Rheumatology and 12 Rehabilitation EMQs" [24] "The assessment tools were constructed by a panel of subject matter experts (SME)" [22]
	Items	13 (68.42%) [2,19-26,27-30,30,31]	"All questions were devised by the lead educational supervisor within each specialty. Training in EMQ writing was provided to the medical specialty supervisors." [24] "Twenty EMQs were selected from the examination question bank; these had been used in a previous examination and had known psychometric quality." [27] "The items were retrieved from an item bank used for the regional examination of general practice certification in the Dutch-speaking part of the country and translated by a professional translator." [23] "We qualitatively analyzed the item-writing manuals of three medical professions (pediatric, osteopathic, and allopathic medicine) and felt that the one created by the National Board of Medical Examiners provided the most in-depth information." [30]
	Qualifications of test developers	6 (31.58%) [2,22,24,28,29,31]	"The development of the EMQs started with an introduction session for 'writing parties' led by Cox and Swanson of the National Board of Medical Examiners (NBME). The disciplines were autonomous in organizing writing groups and composing their part of the test. The degree of compliance with the recommended procedure was not controlled." [2] "The stems within each clinical presentation were designed to be as similar as possible in length, difficulty, and the presence of distracters." [21] "After reviewing the EMQ to ensure that all parts of the neuromus system were assessed, 10 items in which the localization was repeatedly tested were discarded." [29] "All the vignettes and questions were constructed from the clinical situation representative of urological practice by 5 content experts (urologists) and assessed by a medical education expert in the first round prior to the administration of the test" [28] "Consensus on the content was reached by expert discussion and agreement with a panel of 15 attending neurologists from the Department of Neurology, National Neuroscience Institute, Singapore." [29] "Members of a United States Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge (CK) Test Material Development Committee served as content experts" [31]
Response process	Description and/or mentioning of quality control measures	8 (42.11%) [2,19,20,22,28,29,31,32]	"The test was organized simultaneously in three lecture rooms. The instructions were presented audibly by one 'voice'. The booklets were presented in a random order. One co-worker per 20 students guaranteed the distribution of the booklets. The student had to hand in the filled-out booklet before taking a break. The students were not prepared for content, or for question format." [2] "Pre-test familiarization of candidates with the testing and scoring system. Since the learners were not fully aware of these methods of assessment, a voluntary practice session was held with them to make them familiar with the format." [28] "The aim of this study is to explore the effect of images on the validity of test items by investigating the cognitive processes students use to answer them. Methods: At the start of the experiment, the participants were informed about the procedure" [32] "We calculated the item difficulty (p value) of questions in each system to provide further evidence beyond the blueprinting process for their similarity in standard, thus allowing the cross-comparison of scores between disciplines. This is calculated by dividing the number of correct responses for each question by the total number of responses. Values range between 0 and 1, a high value indicating more correct responses. The average item difficulty of a musculoskeletal question was 40.71, for a cardiovascular question it was 0.71 and for a neurological question it was 0.69." [19]
	Internal structure	Difficulty index estimation	4 (21.05%) [19,24,26,31]
Internal structure	Discrimination index estimation	5 (26.32%) [21,24,29-31] 14 (73.68) [2,19-26,28-30,33,34]	Internal consistency estimated by Cronbach's alpha "The 55-item EMQ showed good reliability (Cronbach's α 0.85) and good generalizability (a G-coefficient 0.85 and a dependability coefficient of 0.83)" [29] Internal consistency estimated by Kuder – Richardson 20 (KR-20) "EMQs reliability was calculated using Kuder – Richardson 20 (KR-20) showed a mean of 0.84 and standard deviation of 0.11 for batch 2018, and a mean of 0.87 and standard deviation of 0.02 for batch 2019." [25] Other methods (not specified) "Increasing the number of topics with fewer items each would increase reliability but also testing time. The results of the generalisability analysis G study are reported in Table 1. They indicate that the variance component attributed to subjects was small (645 and 601 for 2005 and 2006, respectively), accounting for only 2.64 and 2.50% of score variance. Variance attributed to items was over twice that of variance attributed to topics. The variance attributed to interaction between subjects and items was the largest component explaining around two-thirds of the total variance." [23] Person-separation index "An estimate of the internal consistency reliability of the examination is based on the Person Separation Index where the estimates on the logit scale for each person are used to calculate reliability, the Person Separation Index was low, 0.50, indicating a low reliability. This, however, can be ascribed to the intrinsic homogeneity of the students who are selected under rigorous criteria and who are all studying for the same exam." [24]
	Internal consistency estimation	Description and/or mentioning of item analysis (reliability coefficient)	14 (73.68%) [26,28], [2], [20-22,25,27,30,31,33-36]
Relationship to other variables	Description and/or mentioning of correlations to another measure	17 (73.68%) [26,28], [2], [20-22,25,27,30,31,33-36]	
Consequences	Description and/or mentioning the impact of the test scores on learners, teachers, or policy	11 (57.89%) [19-24,26,27,30,31,34]	

the difficulty index by calculating the percentage of correct responses in the item [19,26,31]. Bhakta, Tennant [24] estimation of difficulty index by other more complex formulae was considered as another pattern.

Discrimination index estimation

The reported discrimination index estimation as an internal structure-related source of validity was presented in three patterns. One was the estimation of the discrimination index by calculating the point biserial index [30]. Bhakta, Tennant [24] presented another pattern by using complex formulae/comparisons for the estimation of the discrimination index. A third pattern was mentioning the discrimination index without stating how it was estimated [21,29,31].

Internal consistency estimation

The reported internal consistency estimation as an internal structure-related source of validity was presented in four patterns. Estimation of internal consistency by Cronbach's alpha [19,21,22,28,29,33]. Another pattern was the estimation of internal consistency using Kuder – Richardson 20 (KR-20) [25,30,34]. A third pattern was grouped into other methods, in which the approach of estimation was not specified [2,19,23,26]. Bhakta, Tennant [24] mentioned the Pearson-separation index and it was considered a fourth pattern.

Relationship to other variables-related sources of validity evidence

The relationship to other variables as a source of validity evidence was reported in three patterns. One pattern is presenting the relationship between two or more assessment tools [21,22,25,26,28,31,33–36,41]. Another pattern is presenting the relationship with similar assessments in different years [2,20]. The third pattern by Beullens, Struyf [27] was presenting the relationship between the assessment tool and the assessed skill.

Consequence-related sources of validity evidence

The consequence as a source of validity evidence was reported in two patterns. One pattern is presenting the consequence of curriculum development [19,20,22]. The other pattern was presenting the consequence of the practices of test construction [21,23,24,26,27,30,31,34].

Discussion

Assessment-related educational research is expected to have enough evidence to build a strong validity argument. However, misconceptions about validity

and malpractices in implementing this concept in assessment-related educational research have been reported [42] and it remains unclear what sorts of evidence to be presented [12,15] and how to present them. This review aimed to explore the current status of reporting and presenting those evidence in studies that include EMQs.

In this review, the authors made a differentiation between reporting and presenting validity evidence. In which reporting means whether a certain source of evidence was stated or not, so it's a yes or no classification. While presenting the evidence refers to the way it was expressed, for instance, if the source of evidence is content, then what kind of evidence was presented? Was it a blueprint? A proof of the quality of the contents? If it is a blueprint, was it stated or inferred? However, according to the inclusion criteria, all the included studies had to report some kind of validity evidence. In this study, the authors indicate that (reporting some kind of validity evidence) is not enough. The number of the reported evidence is important as well as the source of the reported evidence in which the consistency between the evidence can help build a strong validity argument (e.g., arguing that a certain assessment tool can provide wider coverage of the curricular contents requires providing evidence such as -at least- blueprinting and reliability coefficient together, and not one of them alone, or other evidence of different source).

Definition of validity

The adopted definition of validity in the included studies is the old tripartite concept as described by the American Psychological Association [43] (i.e., dividing validity into construct, content, criterion, etc.) instead of implementing the current definition of validity as a unitary concept [13]. It was adopted by the majority of the included studies ($n = 12$, 63.16%). The rest of the studies didn't clarify the adopted framework. Gasmalla and Tahir [12] reported that the emphasis on the 'types of validity' rather than the 'sources of validity evidence' is a common misconception among researchers. The misidentification of the implemented validity framework, not stating it, and/or embracing an outdated framework has been previously reported [15,44]. This misconception negatively impacted the practice of reporting and presenting validity evidence.

Reporting and presenting validity evidence

Regarding reporting of validity evidence, the results of this review indicated that validity evidence is suffering under-reporting. Driving inferences about the educational impact of the EMQs require better

reporting and a clear presentation of the reported evidence. The most-reported evidence in the included studies was the reliability coefficient and its relationship to another variable. These specific two pieces of evidence support some features of the educational impact of EMQs such as the ability of these questions to encourage deep learning, assess clinical reasoning, and differentiate experts from non-experts as well as being more capable of assessing clinical reasoning than A-type MCQs [2,27–29,33]. However, these inferences about the educational impact of the EMQs are jeopardized by the under-reporting of significant evidence such as item difficulties ($n = 4$, 21.05%) and item discrimination ($n = 5$, 26.32%).

The presentation pattern of the reported validity evidence affects the value of the evidence and imposes a challenge when the weighting of the evidence is considered to determine its value. For example, presenting content-related sources of validity in the included studies had its shortcoming as well. Blueprint was only reported in ten studies (52.63%). About half of them indicated blueprinting by inferring to process rather than clearly stating a blueprint. Blueprint is the crown jewel of the evidence related to content. It shields against threats of validity [45,46]. Failing to report it can substantially weaken the validity argument by introducing construct underrepresentation [6]. Regarding the presentation patterns of the quality of the contents (Table 3), the different four patterns add to the variability of the weight of the evidence, i.e., is reporting a panel meeting more valuable than retrieving items from a Q-bank? Items retrieved from the question banks are usually reviewed extensively by two different committees [47] and were tested previously with their difficulty index and point biserial coefficient measured [48]. This makes presenting a question bank as a source of items superior to presenting a panel meeting when reporting the quality of the contents. This example denotes that not only the number of reported items is important, but also the presentation pattern of the reported evidence matters in weighting the value of the evidence.

Quality of included studies

For the assessment of the quality of the included studies, two different approaches were used. In the first approach, the quality of the included studies - based on MERSQI- gives an average score of 10.7 out of the maximum score of the instrument which is 18. The MERSQI was developed to evaluate the quality of methods utilized in experimental and observational studies based on six domains. However, none of these domains entails the quality of reporting and presenting validity evidence. Even the section (Validity evidence for evaluation instrument scores) of the

MERSQI, involves validating the data collection tool in the manuscript under scrutiny. This is quite different from reporting and presenting validity evidence, for the validity in the context of educational assessment is a property of the interpretation of test scores, it is not a property of the test (as a data collection tool) itself, this is a common misconception about the validity [12]. To the best of our knowledge, no tool involves this measure. Although there are not many reports of the quality of assessment-related educational research in reporting validity evidence, the 'less than ideal' pattern is backdated since Cook, Zendejas [49] set the alarm, followed by [44]. However, there is no clear guide on how to report and present validity evidence in assessment-related research.

Thus, the authors sought to assess the quality of the included studies based on another approach, which is by determining the number of reported validity evidence. In this review, no study reported all the aspects of validity evidence operational cues. Among the total of nine defined operational cues in this review, none of the included studies reported the whole nine. The best record was reporting seven, and even this was reported by only three studies (15.78%). About half of the included studies failed to report more than half of the operational cues (Table 2).

Implications

The implication of this study for educators is how they should rely on assessment-related educational research to build a strong validity argument in favor of the educational impact of the EMQs. The outcome of this study will positively impact future research work involving assessment in terms of quality of research, in which the quality of research should utilize 'reporting and presenting the sources of validity' as a core part of the research. For considering such research: a) the definition of validity as a unitary concept has to be advocated and adopted by researchers in this field, and b) the validity evidence in assessment-related educational research must be fairly reported and presented. Another implication is that quality assessment of such studies requires the development of a suitable tool that evolves around reporting and presenting validity evidence. This as well implies all assessment-related educational research.

Strengths

The authors presented a primary measure of the quality of presenting validity evidence based on the number of operational cues included in each study. To this end, the second strong point is that we included all studies utilizing EMQs even if they didn't present evidence of validity, this reflects the

real situation of the quality of the studies and provided wider inclusion of studies. A third strength is that we extended our findings to not only report the validity evidence but to describe the way they were presented.

Limitations

Since there is no clear guidance on how to report and present evidence of validity in assessment-related research, it was difficult to extract and identify patterns with the ambiguity of the authors' reporting validity evidence in the included studies. However, the authors believe these limitations have little impact on our conclusion since the pilot study provided the authors with the necessary insights to identify and evaluate the presented evidence of validity. Furthermore, we managed the difficulty of the extraction by extracting the data independently and if there were any discrepancies we solved them by consensus. Another limitation is that our review included articles published in the realms of Medicine, Dentistry, Pharmacy and Nursing, this didn't cover other specialties such as physiotherapy. Lastly, the search included articles published only in the English language.

Conclusions

This study concludes that reporting and presenting validity evidence appeared to be deficient. Thus, the available evidence can hardly provide a strong validity argument that supports the educational impact of EMQs. The most-reported pieces of evidence were the reliability coefficient, as well as the relationship to another variable. The least reported evidence were item difficulties and discrimination. These pieces of evidence are important indicators of the educational impact. However, the discrepancy in the reporting frequency between the most-reported two first two pieces of evidence and the least-reported last two pieces of evidence make it hard to build a strong validity argument. Blueprint, although crucial, was explicitly mentioned in less than a third of the studies, which in turn added to the challenge.

The lack of awareness of the existence of hypothetical frameworks that considers validity as a unitary concept seems to be a contributing factor. However, and most importantly, the lack of prescriptive guidance on how to apply that framework to assessment-related research is likely the main issue. This work also calls for more work on developing a tool to measure the quality of both the reporting and the presenting of validity evidence.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Mohamed H. Taha  <http://orcid.org/0000-0003-0808-5590>

Hosam Eldeen Elsadig Gasmalla Mohammed  <http://orcid.org/0000-0003-2590-8587>

Mohamed Elhassan Abdalla  <http://orcid.org/0000-0002-9241-1370>

Muhamad Saiful Bahri Yusoff  <http://orcid.org/0000-0002-4969-9217>

Majed M. Wadi  <http://orcid.org/0000-0002-8117-770X>

References

- [1] Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. *Teach And Learn In Med: An Int J.* 1993;5 (2):107–115. doi: [10.1080/10401339309539601](https://doi.org/10.1080/10401339309539601)
- [2] Beullens J, Van Damme B, Jaspert H, et al. Are extended-matching multiple-choice items appropriate for a final test in medical education? *Med Teach.* 2002;24(4):390–395. doi: [10.1080/0142159021000000843](https://doi.org/10.1080/0142159021000000843)
- [3] Gasmalla HEE, Tahir MEMM. R-Type MCQs (extended matching questions). In: Gasmalla H, Ibrahim A, Wadi M, Taha M, editors. *Written assessment in medical education.* Cham, Switzerland: Springer; 2023. p. 91–99.
- [4] Duthie S, Hodges P, Ramsay I, et al. Emqs: a new component of the MRCOG part 2 exam. *The Obstetrician & Gynaecologist.* 2006;8(3):181–185. doi: [10.1576/toag.8.3.181.27255](https://doi.org/10.1576/toag.8.3.181.27255)
- [5] Wood EJ. What are extended matching sets questions? *Biosci Educ.* 2015;1(1):1–8. doi: [10.3108/beej.2003.01010002](https://doi.org/10.3108/beej.2003.01010002)
- [6] Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Adv Health Sci Educ Theory Pract.* 2002;7(3):235–241. doi: [10.1023/a:1021112514626](https://doi.org/10.1023/a:1021112514626)
- [7] Wilson R, Case S. Extended matching questions: an alternative to multiple-choice or free-response questions. *J Vet Med Educ.* 1993;20(3). [accessed 2023 June]. <https://scholar.lib.vt.edu/ejournals/JVME/V20-3/wilson.html>
- [8] Tweed M. Adding to the debate on the numbers of options for MCQs: the case for not being limited to MCQs with three, four or five options. *BMC Med Educ.* 2019;19(1):354. doi: [10.1186/s12909-019-1801-x](https://doi.org/10.1186/s12909-019-1801-x)
- [9] van Bruggen L, Manrique-van Woudenberg M, Spierenburg E, et al. Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspect Med Educ.* 2012;1 (4):162–171. doi: [10.1007/s40037-012-0024-1](https://doi.org/10.1007/s40037-012-0024-1)
- [10] Samuels A. Extended matching questions and the Royal Australian and New Zealand college of

- psychiatrists written examination: an overview. *Australas Psychiatry*. 2006;14(1):63–66. doi: [10.1080/j.1440-1665.2006.02247.x](https://doi.org/10.1080/j.1440-1665.2006.02247.x)
- [11] Burton JL. How to write and how to answer EMQs. *Obstet, Gynaecol & Reprod Med*. 2009;19(12):359–361. doi: [10.1016/j.ogrm.2009.09.002](https://doi.org/10.1016/j.ogrm.2009.09.002)
 - [12] Gasmalla HEE, Tahir ME. The validity argument: addressing the misconceptions. *Med Teach*. 2021;43(12):1453–1455. doi: [10.1080/0142159X.2020.1856802](https://doi.org/10.1080/0142159X.2020.1856802)
 - [13] Aera APA, NCME. Standards for educational and psychological testing. American Psychological Association; 1999.
 - [14] Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837. doi: [10.1046/j.1365-2923.2003.01594.x](https://doi.org/10.1046/j.1365-2923.2003.01594.x)
 - [15] Cook DA, Zendejas B, Hamstra SJ, et al. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract*. 2014;19(2):233–250. doi: [10.1007/s10459-013-9458-4](https://doi.org/10.1007/s10459-013-9458-4)
 - [16] Hammick M, Dornan T, Steinert Y. Conducting a best evidence systematic review. Part 1: from idea to data coding. *BEME Guide No 13 Med Teach*. 2010;32(1):3–15. doi: [10.3109/01421590903414245](https://doi.org/10.3109/01421590903414245)
 - [17] Gordon M, Gibbs T. STORIES statement: publication standards for healthcare education evidence synthesis. *BMC Med*. 2014;12(1):1–9. doi: [10.1186/s12916-014-0143-0](https://doi.org/10.1186/s12916-014-0143-0)
 - [18] Reed DA, Cook DA, Beckman TJ, et al. Association between funding and quality of published medical education research. *JAMA*. 2007;298(9):1002–1009. doi: [10.1001/jama.298.9.1002](https://doi.org/10.1001/jama.298.9.1002)
 - [19] Basu S, Roberts C, Newble DI, et al. Comparing and contrasting undergraduate competence in musculoskeletal medicine with cardiovascular medicine and neurology. *Rheumatology (Oxford)*. 2004;43(11):1398–1401. doi: [10.1093/rheumatology/keh343](https://doi.org/10.1093/rheumatology/keh343)
 - [20] Basu S, Roberts C, Newble DI, et al. Competence in the musculoskeletal system: assessing the progression of knowledge through an undergraduate medical course. *Med Educ*. 2004;38(12):1253–1260. doi: [10.1111/j.1365-2929.2004.02017.x](https://doi.org/10.1111/j.1365-2929.2004.02017.x)
 - [21] Coderre SP, Harasym P, Mandin H, et al. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Med Educ*. 2004;4(1):23. doi: [10.1186/1472-6920-4-23](https://doi.org/10.1186/1472-6920-4-23)
 - [22] Harding S, Britten N, Bristow D. The performance of junior doctors in applying clinical pharmacology knowledge and prescribing skills to standardized clinical cases. *Br J Clin Pharmacol*. 2010;69(6):598–606. doi: [10.1111/j.1365-2125.2010.03645.x](https://doi.org/10.1111/j.1365-2125.2010.03645.x)
 - [23] Dory V, Gagnon R, Charlin B. Is case-specificity content-specificity? An analysis of data from extended-matching questions. *Adv Health Sci Educ Theory Pract*. 2010;15(1):55–63. doi: [10.1007/s10459-009-9169-z](https://doi.org/10.1007/s10459-009-9169-z)
 - [24] Bhakta B, Tennant A, Horton M, et al. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Med Educ*. 2005;5(1):9. doi: [10.1186/1472-6920-5-9](https://doi.org/10.1186/1472-6920-5-9)
 - [25] AlShamlan NA, Al Shammari MA, Darwish MA, et al. Evaluation of multifaceted assessment of the fifth-year medical students in family medicine clerkship, Saudi Arabia experience, *J Multidiscip Healthc*. 2020;13:321–328. [10.2147/JMDH.S241586](https://doi.org/10.2147/JMDH.S241586)
 - [26] Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Med Educ*. 2001;35(4):326–330. doi: [10.1046/j.1365-2923.2001.00929.x](https://doi.org/10.1046/j.1365-2923.2001.00929.x)
 - [27] Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? *Med Educ*. 2005;39(4):410–417. doi: [10.1111/j.1365-2929.2005.02089.x](https://doi.org/10.1111/j.1365-2929.2005.02089.x)
 - [28] Nazim SM, Talati JJ, Pinjani S, et al. Assessing clinical reasoning skills using script concordance test (SCT) and extended matching questions (EMQs): a pilot for urology trainees. *J Adv Med Educ Prof*. 2019;7(1):7–13. doi: [10.30476/jamp.2019.41038](https://doi.org/10.30476/jamp.2019.41038)
 - [29] Tan K, Chin HX, Yau CWL, et al. Evaluating a bedside tool for neuroanatomical localization with extended-matching questions. *Anat Sci Educ*. 2018;11(3):262–269. doi: [10.1002/ase.1741](https://doi.org/10.1002/ase.1741)
 - [30] D'Antoni AV, DiLandro AC, Chusid ED, et al. Psychometric properties and podiatric medical student perceptions of USMLE-style items in a general anatomy course. *J Am Podiatr Med Assoc*. 2012;102(6):517–528. doi: [10.7547/1020517](https://doi.org/10.7547/1020517)
 - [31] Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Acad Med*. 2008;83(10 Suppl):S21–4. doi: [10.1097/ACM.0b013e318183e5bb](https://doi.org/10.1097/ACM.0b013e318183e5bb)
 - [32] Vorstenbosch MA, Bouter ST, van den Hurk MM, et al. Exploring the validity of assessment in anatomy: do images influence cognitive processes used in answering extended matching questions? *Anat Sci Educ*. 2014;7(2):107–116.
 - [33] Beullens J, Struyf E, Van Damme B. Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. *Med Educ*. 2006;40(12):1173–1179. doi: [10.1111/j.1365-2929.2006.02627.x](https://doi.org/10.1111/j.1365-2929.2006.02627.x)
 - [34] Lukić IK, Glunčić V, Katavić V, et al. Weekly quizzes in extended-matching format as a means of monitoring students' progress in gross anatomy. *Ann Anat*. 2001;183(6):575–579. doi: [10.1016/s0940-9602\(01\)80073-x](https://doi.org/10.1016/s0940-9602(01)80073-x)
 - [35] Gibson S, Leinster S. How do students with dyslexia perform in extended matching questions, short answer questions and observed structured clinical examinations? *Adv Health Sci Educ*. 2011;16(3):395–404. doi: [10.1007/s10459-011-9273-8](https://doi.org/10.1007/s10459-011-9273-8)
 - [36] Fong CL, Sidi H, Mahady ZA, et al. The validity of the assessment methods in an undergraduate psychiatry module examination among a sample of Fourth Year Malaysian medical students. *Int Med J*. 2012;19(4):347–351.
 - [37] Harden A, Garcia J, Oliver S, et al. Applying systematic review methods to studies of people's views: an example from public health research. *J Epidemiol & Community Health*. 2004;58(9):794–800. doi: [10.1136/jech.2003.014829](https://doi.org/10.1136/jech.2003.014829)
 - [38] Pope C, Mays N, Popay J. How can we synthesize qualitative and quantitative evidence for healthcare policy-makers and managers?. *Healthc Manage Forum*. 2006;19(1):27–31. doi: [10.1016/S0840-4704\(10\)60079-8](https://doi.org/10.1016/S0840-4704(10)60079-8). PMID: 17330642.
 - [39] Ryan R, Cochrane Consumers and Communication Review Group. Cochrane consumers and communication review group: data synthesis and analysis. 2013. <https://cccr.cochrane.org/>
 - [40] Campbell M, Katikireddi SV, Sowden A, et al. Improving conduct and reporting of narrative

- synthesis of quantitative data (ICONS-Quant): protocol for a mixed methods study to develop a reporting guideline. *BMJ Open*. 2018;8(2):e020064. doi: [10.1136/bmjopen-2017-020064](https://doi.org/10.1136/bmjopen-2017-020064)
- [41] Johnson EO, Charchanti AV, Troupis TG. Modernization of an anatomy class: from conceptualization to implementation. A case for integrated multimodal-multidisciplinary teaching. *Anat Sci Educ*. 2012;5(6):354–366. doi: [10.1002/ase.1296](https://doi.org/10.1002/ase.1296)
- [42] St-Onge C, Young M. Evolving conceptualisations of validity: impact on the process and outcome of assessment. *Med Educ*. 2015;49(6):548–550. doi: [10.1111/medu.12734](https://doi.org/10.1111/medu.12734)
- [43] Association AP. Standards for educational and psychological tests and manuals. Washington (DC): American Psychological Association; 1966.
- [44] Young M, St-Onge C, Xiao J, et al. Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspect Med Educ*. 2018;7(3):182–191. doi: [10.1007/s40037-018-0433-x](https://doi.org/10.1007/s40037-018-0433-x)
- [45] Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38(3):327–333. doi: [10.1046/j.1365-2923.2004.01777.x](https://doi.org/10.1046/j.1365-2923.2004.01777.x)
- [46] Raymond MR, Grande JP. A practical guide to test blueprinting. *Med Teach*. 2019;8(41):854–861. doi: [10.1080/0142159X.2019.1595556](https://doi.org/10.1080/0142159X.2019.1595556)
- [47] Freeman A, Nicholls A, Ricketts C, et al. Can we share questions? Performance of questions from different question banks in a single medical school. *Med Teach*. 2010;32(6):464–466. doi: [10.3109/0142159X.2010.486056](https://doi.org/10.3109/0142159X.2010.486056)
- [48] Sales D, Sturrock A, Boursicot K, et al. Blueprinting for clinical performance deficiencies—lessons and principles from the general medical Council’s fitness to practise procedures. *Med Teach*. 2010;32(3):e111–4. doi: [10.3109/01421590903386781](https://doi.org/10.3109/01421590903386781)
- [49] Cook DA, Zendejas B, Hamstra SJ, et al. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ*. 2014;19(2):233–250. doi: [10.1007/s10459-013-9458-4](https://doi.org/10.1007/s10459-013-9458-4)

Appendices

Appendix1: Full search strategy

PubMed: BT_001 ((((((Extended matching questions) OR (Extended matching items)) OR (R-type items)) OR (R-type questions)) OR (EMQs)) OR (EMIs)) AND ((health professions education) OR (medical education)) AND ((validity) OR (reliability)) AND (NULL)).

Scopus

(TITLE-ABS-KEY (extended AND matching AND questions) OR TITLE-ABS-KEY (r-type AND items) OR TITLE-ABS-KEY (extended AND matching AND items) OR TITLE-ABS-KEY (r-type AND questions) OR TITLE-ABS-KEY (emq) OR TITLE-ABS-KEY (emi) AND TITLE-ABS-KEY (health AND professions AND education) OR TITLE-ABS-KEY (medical AND education) AND TITLE-ABS-KEY (validity) OR TITLE-ABS-KEY (reliability)).

Web of Sciences

('extended matching questions') OR ('extended matching items') OR ('EMQs') OR ('EMIs').

CINAHL

((('Extended matching questions' OR 'Extended matching items' OR 'R-type items' OR 'R-type questions' OR 'EMQs' OR 'EMIs') AND ('health professions education' OR 'medical education') AND ('validity' OR 'reliability'))).

ERIC

((('Extended matching questions' OR 'Extended matching items' OR 'R-type items' OR 'R-type questions' OR 'EMQs' OR 'EMIs') AND ('health professions education' OR 'medical education') AND ('validity' OR 'reliability'))).

Appendix2: Thematic analysis of patterns of presenting the reported validity evidence

Theme 1

Content-related sources of validity

a) blueprinting

As shown in [Table 1](#), there were two patterns. Either the term 'blueprint' was stated clearly or inferred. Verbatim examples of the first pattern are shown below:

The test was blueprinted against internal and external guidelines. To ensure representative sampling of curriculum content, selected themes were transferred into a blueprinting framework [19].

This was achieved by using a 2-dimensional blueprint design which matched the derived subject matter against relevant basic and clinical science scenarios important to the PRHO [20].

The examination was blueprinted from the clinical core curriculum [26].

For each assessment tools, there is also a specific content blueprint to cover suitable course contents [25].

A subject matter expert (SME) panel constructed a blueprint [22].

A list of common clinical problems was made per medical discipline. From these problems, a list of corresponding diseases was derived. For each disease, the aspects relevant for the 7th-year student were selected (e.g., diagnosis, therapy). For some aspects, the required level of knowledge had to be specified (e.g., a thorough knowledge concerning differential diagnosis or notions of therapy). Each discipline had to start from a table – called blueprint – formed by the dimensions “diseases” and “aspects”. [2]

Verbatim examples of the second pattern 'where the blueprint was inferred' are shown below:

Only questions concerning diagnosis or pathogenesis were included. Half the questions concerned internal medicine problems and the other half related to other disciplines. The main complaints involved in the 20 EMQs are shown in [Table 1](#) [27].

The test covered four domains (chest, urogenital system, locomotor system, and dermatology) with 50 topics and 159 items nested within these topics [23].

An examination for four clinical presentations, each representing a different domain in gastrointestinal medicine, was constructed: dysphagia, chronic diarrhea, nausea and vomiting, and elevated liver enzymes [21].

The examination consisted of 98 EMQs distributed across eight specialties and 27 themes, each with two to four EMQs. There were 12 Oncology, 14 Anaesthetics, 12 Dermatology, 12 A&E, 12 Infectious Diseases, 16 Orthopaedics, 8 Rheumatology and 12 Rehabilitation EMQs [24].

b) quality of contents

There were four identified patterns, reporting a panel meeting, items retrieved from a Q-bank, items' quality based on following guidelines and textbooks, and items' quality based on the training of developers. However, a fifth group was assembled in which there was no certain pattern, so we named it (other patterns). Below are verbatim examples for each:

Reporting a panel meeting.

The assessment tools were constructed by a panel of subject matter experts (SME) [22].

All questions were devised by the lead educational supervisor within each specialty. Training in EMQ writing was provided to the medical specialty supervisors [24].

All the vignettes and questions were constructed from the clinical situation representative of urological practice by 5 content experts (urologists) and assessed by a medical education expert in the first round prior to the administration of the test [28].

Items retrieved from a Q-bank.

Twenty EMQs were selected from the examination question bank; these had been used in a previous examination and had known psychometric quality [27].

Following item analysis and review, the test was modified extensively. Unsatisfactory items were revised or replaced by similar questions from a predesigned bank and additional items were added to improve the reliability of the test [20].

The items were retrieved from an item bank used for the regional examination of general practice certification in the Dutch-speaking part of the country and translated by a professional translator [23].

Items' quality based on following guidelines and textbooks.

The researchers constructed each question using recommended published guidelines [19].

We qualitatively analyzed the item-writing manuals of three medical professions (podiatric, osteopathic, and allopathic medicine) and felt that the one created by the National Board of Medical Examiners provided the most in-depth information [30].

Items' quality based on the training of developers.

The development of the EMQs started with an introduction session for "writing parties" led by Case and Swanson of the National Board of Medical Examiners (NBME). The disciplines were autonomous in organizing writing groups and composing their part of the test. The degree of compliance with the recommended procedure was not controlled [2].

Other patterns of presenting Items' quality.

The items were predominantly clinical vignettes and either single-best-answer or extended-matching items [30].

The stems within each clinical presentation were designed to be as similar as possible in length, difficulty, and the presence of distracters [21].

After reviewing the EMQ to ensure that all parts of the neuraxis were assessed, 10 items in which the localization was repeatedly tested were discarded [29].

To select options statistically for three-, four-, five-, and eight-option A-type versions of the items, an option "priority" was assigned to each incorrect response on the basis of response statistics in the previous study. Both the proportion of examinees choosing an option and their performance on the test as a whole were used in selection, with options frequently chosen by poorly performing examinees given a higher priority. The options assigned the highest priority were then selected for the four A-type versions of the items used in the study. A similar procedure was used to select options statistically for the five- and eight-option EMQ versions, with priorities combined across both items in an EMQ set in choosing responses for inclusion [31].

Content-related sources of validity: qualification of the developers.

There were two patterns. The developers were highlighted either as content experts or as education experts. Verbatim examples are shown below:

The developers as content experts.

The EMQs were developed by staff members of 10 medical disciplines [2].

All the vignettes and questions were constructed from the clinical situation representative of urological practice by 5 content experts (urologists) and assessed by a medical education expert in the first round prior to the administration of the test [28].

Consensus on the content was reached by expert discussion and agreement with a panel of 15 attending neurologists from the Department of Neurology, National Neuroscience Institute, Singapore [29].

The assessment tools were constructed by a panel of subject matter experts (SME). The panel consisted of five clinical pharmacists (one with a special interest in education), two medical consultants (one a professor of Clinical Pharmacology and the other with a special interest in Clinical Pharmacology), three specialist registrars (one geriatric medicine registrar with special interest in clinical pharmacology, one acute medicine registrar and one general surgeon), four FY1 doctors, and one psychologist [22].

The developers as education experts.

Members of a United States Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge (CK) Test Material Development Committee served as content experts [31].

All questions were devised by the lead educational supervisor within each specialty [24].

Theme 2

Response process-related sources of validity

There were three patterns identified, stating procedures related to test organization. Mentioning procedures related to making candidates familiar with the format, and declaring procedures meant to assess the response process. Below are verbatim examples for each:

Stating procedures related to test organization.

The test was organized simultaneously in three lecture rooms. The instructions were presented audibly by one “voice”. The booklets were presented in a random order. One co-worker per 20 students guaranteed the distribution of the booklets. The student had to hand in the filled-out booklet before taking a break. The students were not prepared for content, or for question format [2].

Participants needed a unique password to access the assessment, meaning substitutes could not take part. The researcher invigilated each student to prevent cheating [19].

Access to the test for these students was offered via the internet over a 3-day period. Whilst this strategy could undermine the security of the test, it was a formative exercise for the volunteers, who were well aware that cheating on the test would be of no value to them and would undermine a research study being undertaken by a fellow student. However, to improve test security, each final year volunteer required a unique password to access the assessment through the Networked Learning Environment, meaning that substitute students could not take part [20].

For each examinee, a small number of study items were randomly selected for presentation by test administration software, subject to the constraint that the same examinee never saw more than one version of any stem [31].

There was an allowance of 10 min for the completion of the WUSCE and 5 min for the EMQ. The same assessments were carried out at all sites within a week [22].

Mentioning procedures to make candidates familiar with the format.

Pre-test familiarization of candidates with the testing and scoring system. Since the learners were not fully aware of these methods of assessment, a voluntary practice session was held with them to make them familiar with the format [28].

All items had the same lead-in; candidates were asked to work out the site of the lesion for each clinical vignette [29].

Declaring procedures meant to assess the response process.

The aim of this study is to explore the effect of images on the validity of test items by investigating the cognitive processes students use to answer them. Methods: At the start of the experiment, the participants were informed about the procedure [32].

Theme 3: Internal structure-related sources of validity

a) difficulty index estimation

There were two patterns identified. Estimation by calculating the percentage of correct responses in the item and by more complex formulae. Below are verbatim examples for each:

Estimation by calculating the percentage of correct responses in the item.

We calculated the item difficulty (p value) of questions in each system to provide further evidence beyond the blueprinting process for their similarity in standard, thus allowing the cross-comparison of scores between disciplines. This is calculated by dividing the number of correct responses for each question by the total number of responses. Values range between 0 and 1, a high value indicating more correct responses. The average item difficulty of a musculoskeletal question was 40.71, for a cardiovascular question it was 0.71 and for a neurological question it was 0.69 [19].

Table 1 Descriptions of the individual examination components (average score 68.7%) [26].

For each version of each study item, six indices were calculated. The first was the item difficulty, proportion of examinees who responded to the item correctly [31].

Estimation by more complex formulae.

It is within rash model; Location Value identifies question difficulty on logit scale [24].

b) discrimination index estimation

There were three patterns identified. Estimation by calculating point biserial index, using complex formulae/ comparison and the third patterns was not mentioning the method of estimation. Below are verbatim examples for each:

Estimation by calculating point biserial index.

The PB correlation for the extended-matching item in Figure 2 was 0.53 (GA 2011 cohort), which indicates excellent discriminative ability [30].

Estimation using complex formulae/comparisons.

A comparison between the classical approach and the Rasch approach regarding discrimination is given in Figure 7 [24].

Method of estimation was not mentioned

The test was optimized using item analysis where overly concordant (i.e., items with a discrimination index of less than 0.2) or discordant (i.e., items likely to be interpreted in different ways) questions were discarded, trimming the test to 55 items [29].

For each version of each study item, six indices were calculated. The third was an index of item discrimination: the item-total (biserial) correlation, calculated as the correlation between the item (scored 0/1 for incorrect/correct) and the reported Step 2 CK score [31].

Table 2 Extended-matching. Average discrimination index = 0.58 [21].

c) internal consistency estimation

There were four patterns identified. Estimation by Cronbach's alpha, Kuder – Richardson 20 (KR-20) and the third pattern was grouped into other methods (or not specified). Below are verbatim examples for each:

Internal consistency estimated by Cronbach's alpha.

The internal consistency (Cronbach's α) was 0.53 [28].

The reliability of this test was Cronbach's α = 0.75 [19].

Internal consistency measured with Cronbach's alpha was 0.88 in the first series, 0.87 in the second and 0.82 in the third series [33].

The 55-item EMQ showed good reliability (Cronbach's α 0.85) and good generalizability (a G-coefficient 0.85 and a dependability coefficient of 0.83) [29].

Table 2 Extended-matching – Alpha coefficient = 0.66 [21].

The Cronbach's alpha of the EMQs ranged from 0.55 to 0.78 [22].

Internal consistency estimated by Kuder – Richardson 20 (KR-20).

The KR-20 reliability estimates on 19 weekly quizzes in extended-matching format ranged from 0.44 to 0.89 with a mean of 0.64. [34].

EMQs reliability was calculated using Kuder – Richardson 20 (KR-20) showed a mean of 0.84 and standard deviation of 0.11 for batch 2018, and a mean of 0.87 and standard deviation of 0.02 for batch 2019 [25].

Kuder-Richardson 20 coefficients of 0.71 to 0.76 [30].

Other methods (not specified).

The split-half reliability was between 0.73 and 0.86 [2].

The reliability (internal consistency) of this test was 0.6 [20].

The reliability of the composite examination was estimated using multivariate generalizability theory. Weighting tests according to items per test or total test time gave improved reliabilities of 0.93 and 0.81, respectively. Double weighting of the clinical component marginally affected the reliability (0.76) [26].

Increasing the number of topics with fewer items each would increase reliability but also testing time. The results of the generalisability analysis G study are reported in Table 1. They indicate that the variance component attributed to subjects was small (645 and 601 for 2005 and 2006, respectively), accounting for only 2.64 and 2.50% of score variance. Variance attributed to items was over twice that of variance attributed to topics. The variance attributed to interaction between subjects and items was the largest component explaining around two-thirds of the total variance [23].

Person-separation index.

An estimate of the internal consistency reliability of the examination is based on the Person Separation Index where the estimates on the logit scale for each person are used to calculate reliability. the Person Separation Index was low, 0.50, indicating a low reliability. This, however, can be ascribed to the intrinsic homogeneity of the students who are selected under rigorous criteria and who are all studying for the same exam [24].

Theme 4

Relationship to other variables -related sources of validity

There were three patterns identified. Relationship between two or more assessment tools, with similar assessments in different years, and the relationship between assessment tool and assessed skill. Below are verbatim examples for each:

Relationship between two or more assessment tools.

Both SCT and EMQs are capable of discriminating between learners according to their clinical experience in urology. Both SCT and EMQs successfully differentiated residents from interns and students in one-way analysis of variance (ANOVA) [28].

Table 2 gives the disattenuated correlations, i.e., the true correlations after factors contributing to the variance between the tests have been corrected for, between the individual examination components [26].

Pearson correlations between DTI scores and EMQ examination scores were low but significant [33].

Strong positive correlations were found between total marks and MCQs, EMQs, and OSCE examinations ($r = 0.857, 0.779$ and 0.715 , respectively). Moderate correlation was found between total marks and continuous assessment ($r = 0.508$). Additionally, moderate correlations were seen between OSCE and all other assessment forms. Similarly, moderate correlations were observed between MCQs and EMQs. Objective Structured Clinical Examinations (OSCE), extended matching questions (EMQ), and multiple-choice questions exams (MCQ) are correlated [25].

Performance in SAQ and EMQ was positively correlated for students with and without dyslexia in all years except for cohort 2004 (Pearsons $r = 0.52$ $p = \text{NS}$) [35].

This result was observed both for A-types and EMQs. The effect was largest for A-types with options selected by committee, with mean item difficulties ranging from 0.728 for the base version to 0.816 for the three-option version. The standard deviation of Step 2 CK scores for U.S./Canadian first-takers on the proportion-correct scale is roughly 0.08, so the effect size should be viewed as moderate to large [31].

Students (>80%) reported that single-best-answer items were easier than extended-matching items [30].

Relationship of examination format to cognitive process. The results of the logistic regression analysis are as follows, in Table 3. There was no difference in the odds of using “expert” methods of problem solving (scheme- inductive or pattern recognition) across the two examination formats (MCQ or EMQ) [21].

The lack of correlation between the EMQs and WUSCEs is explained, at least in part, by the EMQs being less able to discriminate between the application of CPT knowledge than the WUSCEs [22].

Table 2. Correlations between individual assessment tools used in the psychiatry assessment [36].

We conclude that extended-matching quizzes were at least as effective as multiple-choice quizzes and oral examinations and may be better for acquiring synthetic understanding of anatomical concepts especially in combination with other means of knowledge assessment [34].

Relationship with similar assessments in different years.

Construct validity was checked by calculating correlations with exam results. The Pearson correlations between the total test score per series and the total examination score of each of the 7 years of medical education were modest to rather weak but statistically significant. These correlations varied from series to series and from year to year but were in general moderate, which proves criterion validity [2].

Despite the relatively small number of items in each category, there was a significant difference in basic (F 1/4 39.61, $p < 0.001$) and clinical science knowledge (F 1/4 71.43, $p < 0.001$) between years. Using an ANOVA, with a Tamhane’s T2 adjustment there was a significant difference at the 0.05 level between the mean differences in basic science between Years 3, 4 and 5 but not between Years 1 and 2, or between Years 2 and 3. Significant differences in clinical science knowledge were found between Years 2 and 3 and Years 4 and 5 [20].

Relationship between assessment tool and assessed skill.

This study showed that subjects who gave more correct answers also used more forward reasoning, the form of reasoning that is aimed for [27].

Theme 5: Consequence-related sources of validity

There were two patterns identified. Consequence on curriculum development, and the practices of test construction. Below are verbatim examples for each:

Consequence on curriculum development.

Despite the disparity of teaching between musculoskeletal medicine and other major organ systems within Sheffield’s integrated medical curriculum, the knowledge base of medical students in the basic and clinical musculoskeletal sciences appears to be similar to that for cardiovascular medicine and neurology by the time of graduation [19].

Tested final year medical students at Sheffield University did possess sufficient knowledge. Its acquisition throughout the 5 years of the course appeared to be relatively independent of specialty-specific clinical teaching [20].

This study showed that a large proportion of FY1 doctors failed to demonstrate the level of CPT knowledge and prescribing ability required at this stage of their careers. We identified areas of performance weakness that posed high risk to patient safety and suggested ways to improve the prescribing by FY1 doctors [22].

Consequence on the practices of test construction.

Further research into the best format for testing clinical competence at the end of the undergraduate curriculum is urgently needed [26].

It is possible to assess processes of clinical reasoning using EMQs [27].

Reliability of tests using EMIs would benefit from an increase in the number of topics rather than of items per topic. Nevertheless, increasing the number of topics makes sense from a content-validity standpoint, especially in general practice where the domain one wants to generalise to is so broad. Using topics which are more cross-sectional may be one way to strike a balance between broad sampling of cases and satisfactory reliability per unit of testing time [23].

Use of the extended-matching format and smaller numbers of options per item (and more items) should result in more efficient use of testing time and greater score precision per unit of testing time. From a practical perspective, this study suggests that item-writing committees are fairly effective in selection of distracters, even without access to response statistics [31].

Outcomes from students who recently took the APMLE Part I suggest that incorporation of USMLE-style items into the general anatomy course was a successful measure and prepared them well [30].

The results are important to examination construction, by demonstrating direct evidence that problem solving can be tested by pencil-and-paper formats, and specifically change some of the presented misperceptions about the standard five-option MCQ format. Secondly, demonstrating that the two formats can evoke scheme utilization is important [21].

The analysis clearly shows how students use options, and this information coupled with expert knowledge and understanding of the questions will help questions writers to create and improve the quality of EMQs. It allows for a formal test of the unidimensionality of the questions and thus the validity of the summed score. Given the metric calibration which follows fit to the model, it also allows for the establishment of items banks to facilitate continuity and equity in exam standards. Thus modern psychometric methods based upon the work of Rasch provide a useful approach to the calibration and analysis of EMQ undergraduate medical assessments [24].

In conclusion, extended-matching quizzes were educationally at least as effective as oral examinations, but needed less time to perform and evaluate, and were certainly more objective. In comparison with multiple-choice questions, they were also at least as effective. We recommend weekly quizzes based on extended-matching format for monitoring student's progress during a gross anatomy course. When combined with other means of knowledge assessment, such as extended-matching quizzes, they may promote the development of higher level of knowledge in gross anatomy [34].