COVER SHEET

Title: Pattern of Presenting Validity Evidence of Extended Matching Questions (R-type) in Health Professions Education: A systematic Review

Review group:

Hosam Eldeen Elsadig Gasmalla (Lead Reviewer)

Assistant Professor, Faculty of Medicine, Al-Neelain University-Sudan.

Research and Education Development, Sudan International University -Sudan.

hosamalwakeel@hotmail.com

Majed M. Wadi

Medical Education Department, Assessment Unit, College of Medicine, Qassim University-Kingdome of Saudi Arabia

m.wadi@qu.edu.sa

Mohamed H. Taha

Assistant professor of Medical Education, University of Sharjah, University of Sharjah-United Arab Emirates

mtaha@sharjah.ac.ae

Mohamed Elhassan Elsayed

Senior Lecturer in Medical Education, University of Limerick, Ireland

MElhassan.Elsayed@ul.ie

Muhamad Saiful Bahri Yusoff

Associate Professor at Medical Education Department, School of Medical Sciences, Universiti Sains, Malaysia

msaiful_bahri@usm.my

Mohd Kamal Mohd Napiah

Perpustakaan Hamzah Sendut,Universiti Sains Malaysia

mohdkamal@usm.my

Abstract

Background: Extended matching questions (or the R-type questions) were introduced in the 1990s; R-type MCQs offers the benefits of both the objectivity of the A-type MCQs and the domain coverage of the free-response questions. Furthermore, R-type MCQs overcome the disadvantages of A-type MCQs by providing way more options for each item (avoiding the possibility of guessing). Also, the objectivity of R-type in scoring overcome the disadvantages of the subjective and exhausting scoring process of free-response (or open-ended) questions.

Aim: This systematic review presents a comprehensive summary regarding the pattern of reporting and presenting the sources of validity evidence of Extended Matching Questions (R type) in health professions education.

Methods: Following the guidelines of BEME (Best Evidence in Medical Education) Collaboration review, a systematic search in the electronic databases, including MEDLINE via PubMed, Scopus, Web of Science, EMBASE, CINAHL, PsychINFO and ERIC will be conducted to extract studies that utilize R-type MCQs. In addition, hand searching and contacting experts in the field. The study selection criteria include original articles written in English language and published from 1990. The article must report at least one source of validity evidence. Identified studies will be screened by all authors. The quality of the included studies will be assessed using the Medical Education Research Study Quality Instrument (MERSQI) (for quantitative studies).

Systematic review registration: The topic was successfully registered with BEME on the10th July 2020 with Reg No: 0143

Keywords: Extended matching questions, Extended matching items, R-type items, R-type questions, EMQs, EMIs, Health professions education, Validity, Reliability

Contents

Background

Extended matching questions were introduced in the 1990s as a response to the criticisms of A-type MCQs, although A-type MCQs are considered valid (due to its wide coverage of the contents), reliable (scoring is objective), easy to score, and can be introduced to a large number of students at a time, it mainly assesses recall (Wood 2003; Duthie et al. 2006), and many assessments' constructors fail in the trap of making MCQs that assess only recall, beside the chance of correct guessing is there due to its number of options (4 to 5 in most literature). Thus, the doubts raised about the ability of A-type to assess the application of knowledge and clinical reasoning (Wood 2003) besides concerns about the "correct guessing." Concerns about the number of options for each question have always been there. The debate of increasing the number of options takes into consideration that introduction of a large number of options per item will mimic the situation in clinical scenarios. Also, in the same assessment, there is no need to make the same number of options for each question for the number of options is different according to the content of the question. Thus the variation in the number of options within the same assessment creates flexibility that makes assessment construction easier (Tweed 2019). In the other hand, free-response (or open-ended) questions can assess the higher levels of cognitive functions and clinical reasoning, but their limited coverage of the contents makes them less valid, the ambiguity of the intended answer of the student and/or the intentions of the examiner make them reduces the reliability, besides the logistic difficulties come with the scoring process and its subjectivity (Wilson and Case 1993).

Thus, R-type MCQs come at midway between both A-type MCQs and free-response questions. Extended matching questions have been introduced by (Case and Swanson 1993). A large number of options reduces the chance of correct guessing and allows for assessment of clinical reasoning, thus overcoming the disadvantages of A-type MCQs. Also, the objectivity of R-type in scoring overcome the disadvantages of free-response (or open-ended) questions. Moreover, scoring itself is easy. Furthermore, many other assessment tools were considered for assessment of clinical reasoning, such as script concordance test and true/false questions. However, the script concordance test is mainly used for postgraduate candidates (van Bruggen et al. 2012), and it is laborious to be constructed. In addition to that its suitability to be constructed as a computer-based assessment is of doubts, in another hand true/false questions are considered of low validity and is less used recently, which makes extended matching questions one of the best assessment tools recommended for computer-based assessment (van Bruggen et al. 2012). There

is a movement toward R-type questions in the 2000s. In the UK, it has been introduced in the examinations of the royal college of obstetricians and gynaecologists since 2006 (Duthie et al. 2006; Burton 2009), and in Australia in which it was introduced in the examinations of the Royal Australian and New Zealand College of psychiatrists in 2004 (Samuels 2006).

Validity holds a sacred place since it gives the assessment its meaning. It is the degree to which evidence supports the appropriateness of the interpretations of test scores. Validity is a unitary concept, and validation is an ongoing process of collecting evidence. The old classical framework divides validity into content, criterion (concurrent and predictive), and construct validity. This has been replaced by the current standard framework developed in the 1990s that introduces validity as a unitary concept that has to be supported by five sources of validity evidence (AERA et al. 1999) (content, response process, internal structure, relationship to other variables and consequences) (Downing 2003).

Similar reviews

1. Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Advances in Health Sciences Education, 19(2), 233-250.
2. Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Academic Medicine, 88(6), 872-883.

Those two reviews (Cook David A et al. 2013; Cook D. A. et al. 2014) are, to some extend close to ours; however, they focused on simulation-based assessment.

Justification

1. R-type MCQs offer the benefits of both the objectivity of the A-type MCQs and the domain coverage of the free-response questions (including clinical reasoning) (Wilson and Case 1993).
2. Furthermore, R-type MCQs overcome both the disadvantages of A-type MCQs by minimizing the guessing effect as it contains more options. Also, the objectivity of R-type in scoring overcome the disadvantages of the subjective and exhausting scoring process of free-response (or open-ended) questions (Wilson and Case 1993; Tweed 2019).

3. There is a movement toward R-type questions since the 2000s. In the UK, it has been introduced in the examinations of the royal college of obstetricians and gynaecologists since 2006 (Duthie et al. 2006; Burton 2009), and in Australia in which it was introduced in the examinations of the Royal Australian and New Zealand College of psychiatrists in 2004 (Samuels 2006).
4. Studies that utilize R-type questions need to present enough evidence of validity (Downing 2003). This review aims to explore the current status of reporting those evidence and suggest reporting standards.
5. The outcome of this study will positively impact the future research work involving r-type in terms of quality of research, in which the quality of research using r-type MCQs should utilize "reporting the sources of validity" as a core part of the research.
6. This, in turn, will increase the uses of r-type questions as a tool for assessing clinical reasoning in written or computer-based assessments in both undergraduates and graduates' levels of health professions education.
7. To the best of our knowledge, and according to a scoping review we've conducted (see appendix), there are no systematic reviews reported up to date that explore the reporting pattern and quality of the validity evidence in studies utilizing r-type questions.

Objective

This systematic review presents a comprehensive summary regarding the pattern and quality of reporting and presenting the sources of validity evidence of Extended Matching Questions (R type) in health professions education, in accordance to the five sources of validity, response process, internal structure, relationship to other variables and consequences (see appendix 2), as explained by (Downing 2003)

Research question

What is the pattern of validity evidence reported of using R-type MCQs in the assessment of undergraduate and postgraduate candidates of health professions?

What is the quality of reporting validity evidence reported of using R-type MCQs in the assessment of undergraduate and postgraduate candidates of health professions?

Methods

Study eligibility/ selection criteria

Inclusion criteria are as follows:

- Original articles (study types included are controlled trials (randomized and non-randomized), quasi-experimental designs, cohort studies, including cross-sectional studies and case-control studies)
- Language: English language
- Timeframe: From 1990 until now.
- Scope: articles that utilize R-type MCQs introduced to health professions candidates at both undergraduate and postgraduate levels.
- The article must report at least one validity evidence according to the framework explained by (Downing 2003)
- Definition of health professions includes medicine, dentistry, pharmacy, and nursing (see appendix 1).

We excluded any paper that does not meet the criteria.

Search strategy and study identification

A scoping review was done on PubMed, Scopus, and Web of science. We selected these databases after a brief review of the literature (Haig and Dozier 2003; Bramer et al. 2017). In addition, to EMBASE, CINAHL, PsychINFO and ERIC will be included, as well as hand searching and contacting experts in the field. Our search was performed to extract the abstracts. The search was conducted during April 2020, updated on the 15th of June 2020. The search limit was from 1990 until the last date of search in 2020.

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”.

A full search strategy for each of the included databases is provided (see **appendix 3**). **Appendix 4** shows a list of included studies, while **appendix 5** shows a list of all studies. **Figure 1** shows the process of data identification and screening.

Search strategy and study identification for the systematic review will be the same.

Study selection process

One author (HEG) will perform the title selection.

Then two authors (MW and MT) will independently review the titles and abstracts according to the form (**Form 1 - Data appraisal**) attached as PDF (shorturl.at/dlxTZ). The same form used for data appraisal of scoping review) – if they agreed to exclude the article, it will be excluded, if they had a disagreement, then the full article will be read for further processing. Kappa method will be used to examine the degree of agreement between coders in the process of screening articles or extracting data

Data extraction

In the scoping review phase, all reviewers conducted a meeting to design a Google form to be used for data extraction (**Form 1 - Data appraisal**) attached as PDF; a link to google form is provided here (shorturl.at/dlxTZ). The form was based on the five validity sources (Downing 2003). This form includes the following variables: study titles, authors, year of publication, study purpose, the used methods, intervention (if there), results, outcomes, the existence 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). The form was pretested by three authors (HEG, MW and MT) 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 particular database to ensure accuracy. Then, each author checked the extracted data by 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 form Zoom Video Communications, Inc ©2020 version 5.0

In the systematic review, three authors (ME, MSBY, and AF) will review the full texts of the included articles (according to the "study selection process" mentioned above). Then the final list of articles will be extracted by all authors according to the form (**Form 2 - Reporting Validity Evidence of Extended Matching Questions (R-type) in Health Professions Education**) the form is attached as PDF and a link is provided here (https://forms.gle/6y8R3bMWKDQUfWV36).

Data appraisal

The quality of the included studies will be evaluated using the Medical Education Research Study Quality Instrument (MERSQI) (Reed et al. 2007) (see **Appendix 6**). The three authors will appraise the quality of each one of the included studies. Disagreements will be resolved by discussion to reach a consensus.

Data synthesis

We will provide a description of the five sources of validity (see **appendix 2**). This descriptive synthesis will be utilized based on synthesis evidence to address the review questions and objectives. In attempting to answer the original review questions, the authors are going to present their findings according to the five sources of validity (content, response process, internal structure, relation to other variables, and related consequences) (Downing 2003).

The data will be presented in four tables (see **appendix 7**). **Table (1)**: descriptive data of the included studies. **Table (2):** summary of the included studies. **Table (3)**: pattern of reporting validity evidence, and **table (4)** quality of reporting validity evidence**.**

Potential expectations and implications of this study

This study is going to identify, summarize, and evaluate the findings of all relevant individual studies about validity of R-type MCQs thereby making the available evidence more accessible to researchers and decision makers

This study will pave the way for developing a valid tool to evaluate the quality of research in the field of assessment.

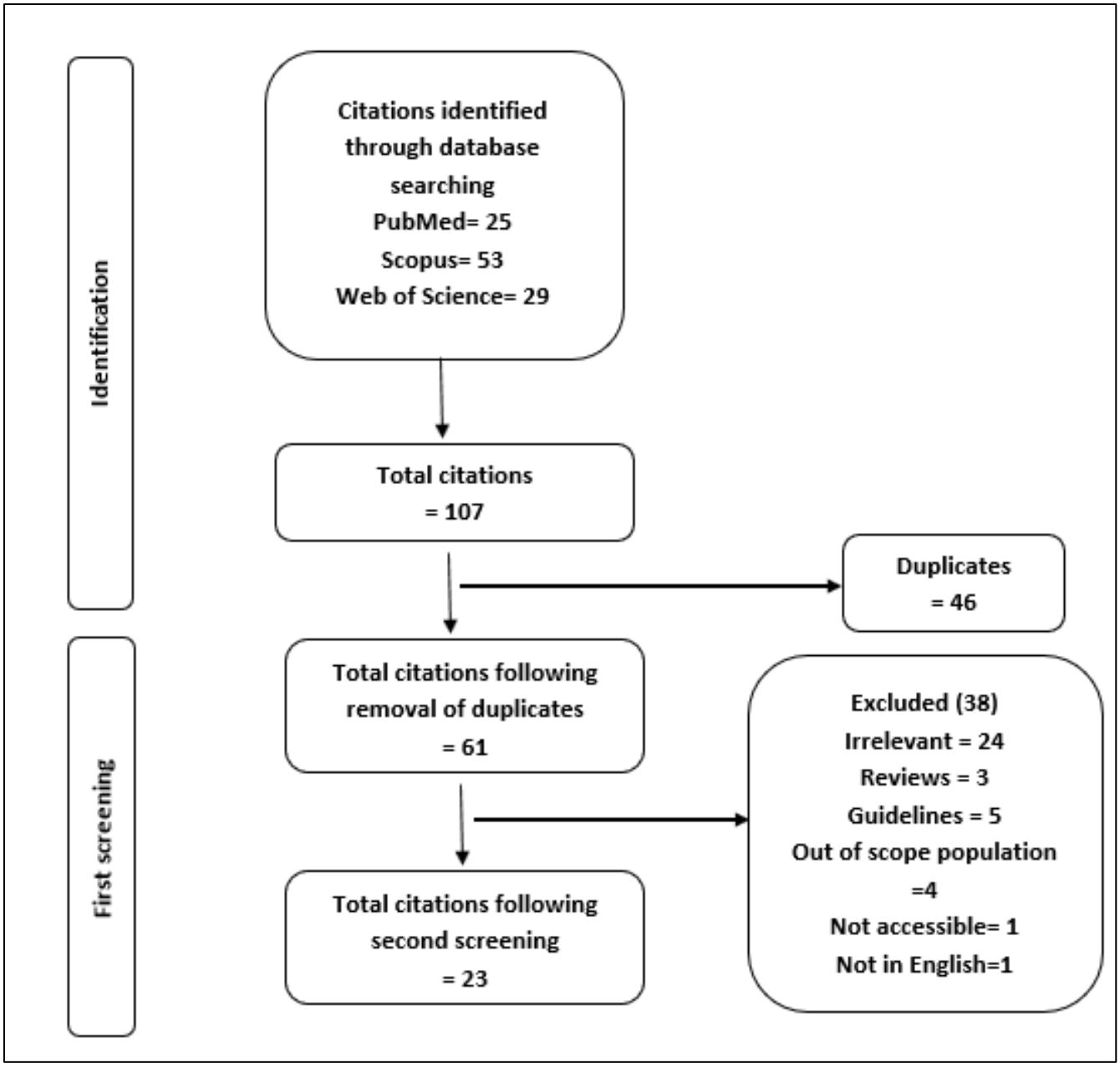

Figure 1. The process of data identification and screening in the scoping review

Project Timetable

Activity	Time limit
Protocol development	July - October 2020
BICC review	November-December 2020
Final/updated search	January 2021
Selection and screening of articles	February-March 2021
Review of full-text articles	April - May 2021
Synthesis of findings	June 2021
Evaluation, review, and dissemination	July 2021

Conflict of interest statement

The authors have no financial, personal, political, institutional, or other conflicts of interest to Report.

Plans for Updating the Review

This review will be updated to track any changes in patterns and quality of reporting validity.

Changes to The Protocol

Minor amendments to the protocol will be reported and communicated with a rationale to BEME.

References

AERA, APA, NCME. 1999. Standards for Educational and Psychological Testing. American Educational Research Association.
Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. 2017. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 6(1):245.
Burton JL. 2009. How to write and how to answer EMQs. Obstetrics, Gynaecology & Reproductive Medicine. 19(12):359-361.
Case SM, Swanson DB. 1993. Extended-matching items: a practical alternative to free-response questions. Teaching and Learning in Medicine: An International Journal. 5(2):107-115.
Case SM, Swanson DB. 2000. Constructing written test questions for the basic and clinical sciences. 3 ed. National Board of Medical Examiners Philadelphia.
Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. 2013. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Academic Medicine. 88(6):872-883.
Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. 2014. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 19(2):233-250.
Downing SM. 2003. Validity: on the meaningful interpretation of assessment data. Medical education. 37(9):830-837.
Duthie S, Hodges P, Ramsay I, Reid W. 2006. EMQs: a new component of the MRCOG Part 2 exam. The Obstetrician & Gynaecologist. 8(3):181-185.
Haig A, Dozier M. 2003. BEME Guide No 3: Systematic searching for evidence in medical education - Part 1: Sources of information. Medical Teacher. 25(4):352-363. English.
Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. 2007. Association between funding and quality of published medical education research. Jama-J Am Med Assoc. 298(9):1002-1009. English.
Samuels A. 2006. Extended matching questions and the Royal Australian and New Zealand College of Psychiatrists written examination: an overview. Australasian Psychiatry. 14(1):63-66.
Tweed M. 2019. 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. 19(1):354.
van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, Vos J. 2012. Preferred question types for computer-based assessment of clinical reasoning: a literature study. Perspect Med Educ. 1(4):162-171.
Wilson R, Case S. 1993. Extended matching questions: an alternative to multiple-choice or free-response questions. J Vet Med Educ. 20(3).
Wood EJ. 2003. What are Extended Matching Sets Questions? Bioscience Education. 1(1):1-8.

Appendix

Appendix 1: Definition of terms

Participants: Health professions education at the undergraduate or postgraduate level (trainee) in medicine, dentistry, pharmacy and nursing.

Intervention: extended matching questions "r-type": a type of MCQs composed of theme, options list, lead-in statement, and stems (Case and Swanson 2000)

Outcome: validity comprises the evidence provided to endorse or disprove the interpretation of assessment scores. The documentation falls under five categories as sources of validity: content, response process, internal structure, relationship to other variables and consequences.

Appendix 2: Operational definitions of validity scources

Source of validity evidence		Operational cue(s)
Content	Blueprint	Description and/or mentioning of test development using a blueprint
	Items quality	Description and/or mentioning of revision and/or meeting for item quality
	Qualifications of test developers	Description and/or mentioning of the specialties and/or qualifications of test developers
Response process		Description and/or mentioning of quality control measures
Internal structure		Description and/or mentioning of item analysis (item difficulties)
		Description and/or mentioning of item analysis (discrimination)
		Description and/or mentioning of item analysis (reliability coefficient)
Relationship to other variables		Description and/or mentioning of correlations (negative or positive) to another measure
Consequences		Description and/or mentioning the impact of the test scores on learners, teachers, or policy

Appendix 3: Full search strategy (in the scoping review)

PubMed: ((((((("Extended matching questions"[All Fields] OR "Extended matching items"[All Fields]) OR (R-type[All Fields] AND ("Items"[Journal] OR "items"[All Fields]))) OR (R-type[All Fields] AND questions[All Fields])) OR "EMQs"[All Fields]) OR "EMIs"[All Fields]) AND ((("validity"[All Fields] OR "reliability"[All Fields]) OR validity[All Fields]) OR ("IEEE Trans Reliab"[Journal] OR "reliability"[All Fields]))) AND ((("health occupations"[MeSH Terms] OR ("health"[All Fields] AND "occupations"[All Fields]) OR "health occupations"[All Fields] OR ("health"[All Fields] AND "professions"[All Fields]) OR "health professions"[All Fields]) AND ("education"[Subheading] OR "education"[All Fields] OR "educational status"[MeSH Terms] OR ("educational"[All Fields] AND "status"[All Fields]) OR "educational status"[All Fields] OR "education"[All Fields] OR "education"[MeSH Terms])) OR ("education, medical"[MeSH Terms] OR ("education"[All Fields] AND "medical"[All Fields]) OR "medical education"[All Fields] OR ("medical"[All Fields] AND "education"[All Fields])))

Web of Sciences

("extended matching questions") Refined by: LANGUAGES: (ENGLISH) AND DOCUMENT TYPES: (ARTICLE OR REVIEW)

("extended matching items") Refined by: LANGUAGES: (ENGLISH) AND DOCUMENT TYPES: (ARTICLE OR REVIEW

Scopus

TITLE-ABS-KEY("Extended Matching Question") AND DOCTYPE(ar OR re) AND PUBYEAR > 1989 AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))

TITLE-ABS-KEY("Extended Matching item") AND DOCTYPE(ar OR re) AND PUBYEAR > 1990 AND (LIMIT-TO (LANGUAGE,"English"))

Appendix 4: List of Included Studies (in the scoping review)

No.	Citation	Sources of Validity:				
		Content	Response process	Internal structure	Relationship to other variables	Consequences
1.	Blackwell, T. A. (1991). A comparison of short-answer and extended-matching question scores in an objective structured clinical exam. *Academic Medicine*, *66*(9).	+	Not clear	Not clear	+	Not clear
2.	Solomon, D. J., Speer, A. J., Perkowski, L. C., & DiPette, D. J. (1994). Evaluating problem solving based on the use of history findings in a standardized-patient examination. *Academic Medicine*	+	+	+	Not clear	Not clear
3.	Fenderson, B. A., Damjanov, I., Robeson, M. R., Veloski, J. J., & Rubin, E. (1997). The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Human pathology*, *28*(5), 526-532.	Not clear	None	+	Not clear	Not clear
4.	Lukić, I. K., Glunčić, V., Katavić, V., Petanjek, Z., Jalšovec, D., & Marušić, A. (2001). Weekly quizzes in extended-matching format as a means of monitoring students' progress in gross anatomy. *Annals of Anatomy-Anatomischer Anzeiger*, *183*(6), 575-579.	+		+	+	

5.	Wass, V., McGibbon, D., & Van der Vleuten, C. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability?. *Medical Education*, *35*(4), 326-330.	+	+	+	+	Not clear
6.	Beullens, J., Damme, B. V., Jaspaert, H., & Janssen, P. J. (2002). Are extended-matching multiple-choice items appropriate for a final test in medical education?. *Medical teacher*, *24*(4), 390-395.	+	Not clear	+	Not clear	Not clear
7.	Basu, S., Roberts, C., Newble, D. I., & Snaith, M. L. (2004). Comparing and contrasting undergraduate competence in musculoskeletal medicine with cardiovascular medicine and neurology. *Rheumatology*, *43*(11), 1398-1401.	+	+	+	+	Not clear
8.	Basu, S., Roberts, C., Newble, D. I., & Snaith, M. (2004). Competence in the musculoskeletal system: assessing the progression of knowledge through an undergraduate medical course. *Medical education*, *38*(12), 1253-1260.	+	+	+	Not clear	+
9.	Coderre, S. P., Harasym, P., Mandin, H., & Fick, G. (2004). The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Medical Education*, *4*(1), 23.	+	Not clear	+	Not clear	None

10.	Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education*, *5*(1), 9.	+	Not Clear	+	Not clear	Not clear
11.	Beullens, J., Struyf, E., & Van Damme, B. (2006). Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. *Medical education*, *40*(12), 1173-1179.	+	Not clear	+	Not clear	Not clear
12.	Swanson, D. B., Holtzman, K. Z., Allbee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine*, *81*(10), S52-S55.	+	+	+	+	Not clear
13.	Dory, V., Gagnon, R., & Charlin, B. (2010). Is case-specificity content-specificity? An analysis of data from extended-matching questions. *Advances in health sciences education*, *15*(1), 55-63.	+	Not clear	+	Not clear	Not clear
14.	Harding, S., Britten, N., & Bristow, D. (2010). The performance of junior doctors in applying clinical pharmacology knowledge and prescribing skills to	+	+	+	+	Not clear

	standardized clinical cases. *British journal of clinical pharmacology*, *69*(6), 598-606.					
15.	D'Antoni, A. V., DiLandro, A. C., Chusid, E. D., & Trepal, M. J. (2012). Psychometric properties and podiatric medical student perceptions of USMLE-style items in a general anatomy course. *Journal of the American Podiatric Medical Association*, *102*(6), 517-528.	+	+	+	+	+
16.	Metcalfe, N. H. (2012). Testing the test: an analysis of the MRCGP Applied Knowledge Test as an assessment tool. *Education for Primary Care*, *23*(1), 13-18.	+		+		+
17.	Nabishah, M., Nasri, A. B., Salam, A., Harlina, H. S., & Ima Nirwana, S. (2012). Setting the Standard of Student Performance: An Alternative Borderline Method. *International Medical Journal*, *19*(2).	+			+	+
18.	NAZIM, S. M., TALATI, J. J., PINJANI, S., BIYABANI, S. R., ATHER, M. H., & NORCINI, J. J. (2019). Assessing clinical reasoning skills using Script Concordance Test (SCT) and extended matching questions (EMQs): A pilot for urology trainees. *Journal of Advances in Medical Education & Professionalism*, *7*(1), 7.	+	+	Not clear	+	Not clear

19.	AlShamlan, N. A., Al Shammari, M. A., Darwish, M. A., Sebiany, A. M., Sabra, A. A., & Alalmaie, S. M. (2020). Evaluation of Multifaceted Assessment of the Fifth-Year Medical Students in Family Medicine Clerkship, Saudi Arabia Experience. *Journal of Multidisciplinary Healthcare, 13*, 321.	Not clear	Not clear	+	+	Not clear
20.	Sturmberg, J.P. and Farmer, E.A., 2008. Assessing general practice knowledge base: The applied knowledge test. *Australian Journal of General Practice*, *37*(8), p.659.	+	+	+	+	+
21.	Beullens, J., Struyf, E. and Van Damme, B., 2005. Do extended matching multiple-choice questions measure clinical reasoning?. *Medical education*, *39*(4), pp.410-417.	+	+	+	Not clear	Not clear
22.	Eijsvogels, T.M., van den Brand, T.L. and Hopman, M.T., 2013. Multiple choice questions are superior to extended matching questions to identify medicine and biomedical sciences students who perform poorly. *Perspectives on medical education*, *2*(5-6), pp.252-263.	Not clear	Not clear	+	+	+
23.	Swanson, D.B., Holtzman, K.Z. and Allbee, K., 2008. Measurement characteristics of content-parallel single-	+	+	+	+	Not clear

	best-answer and extended-matching questions in relation to number and source of options. *Academic Medicine*, *83*(10), pp.S21-S24.					

Appendix 5: List of All Studies (in the scoping review)

No.	Citation	Sources of Validity:					Decision
		Content	Response process	Internal structure	Relationship to other variables	Consequences	
1.	Blackwell, T. A. (1991). A comparison of short-answer and extended-matching question scores in an objective structured clinical exam. *Academic Medicine*, *66*(9).	+	Not clear	Not clear	+	Not clear	Included
2.	Solomon, D. J., Speer, A. J., Perkowski, L. C., & DiPette, D. J. (1994). Evaluating problem solving based on the use of history findings in a standardized-patient examination. *Academic Medicine*	+	+	+	Not clear	Not clear	Included
3.	Beck, D. E., Boh, L. E., & O'Sullivan, P. S. (1995). Evaluating student performance in the experiential setting with confidence. *American journal of pharmaceutical education*, *59*, 236-236.						Excluded, Irrelevant
4.	Mooney, G. A., Bligh, J. G., Leinster, S. J., & Warenius, H. M. (1995). An electronic						Excluded, Irrelevant

	study guide for problem-based learning. *Medical education*, *29*(6), 397-402.						
5.	Fenderson, B. A., Fishback, J., & Damjanov, I. (1996). Weekly mini-examinations (quizzes) based on extended-matching questions as a means for monitoring medical student performance. *CMJ*, *37*(4).						Excluded, Not accessible
6.	Fenderson, B. A., Damjanov, I., Robeson, M. R., Veloski, J. J., & Rubin, E. (1997). The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Human pathology*, *28*(5), 526-532.	Not clear	None	+	Not clear	Not clear	Included
7.	Dugdale, A. E. (1998). The pathway MCQ: a method for teaching and testing deeper knowledge. *Medical Teacher*, *20*(3), 250-253.						Excluded, Irrelevant
8.	Lukić, I. K., Glunčić, V., Katavić, V., Petanjek, Z., Jalšovec, D., & Marušić, A. (2001). Weekly quizzes in extended-matching format as a means of monitoring	+		+	+		Included

	students' progress in gross anatomy. *Annals of Anatomy-Anatomischer Anzeiger*, *183*(6), 575-579.						
9.	Wass, V., McGibbon, D., & Van der Vleuten, C. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability?. *Medical Education*, *35*(4), 326-330.	+	+	+	+	Not clear	Included
10.	Beullens, J., Damme, B. V., Jaspaert, H., & Janssen, P. J. (2002). Are extended-matching multiple-choice items appropriate for a final test in medical education?. *Medical teacher*, *24*(4), 390-395.	+	Not clear	+	Not clear	Not clear	Included
11.	George, S. (2003). Extended matching items (EMIs): solving the conundrum. *Psychiatric Bulletin*, *27*(6), 230-232.						Excluded, Irrelevant
12.	Henly, D. C. (2003). Use of Web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education*, *7*(3), 116-122.						Excluded, Irrelevant
13.	Basu, S., Roberts, C., Newble, D. I., & Snaith, M. L. (2004). Comparing and	+	+	+	+	Not clear	Included

	contrasting undergraduate competence in musculoskeletal medicine with cardiovascular medicine and neurology. *Rheumatology*, *43*(11), 1398-1401.						
14.	Basu, S., Roberts, C., Newble, D. I., & Snaith, M. (2004). Competence in the musculoskeletal system: assessing the progression of knowledge through an undergraduate medical course. *Medical education*, *38*(12), 1253-1260.	+	+	+	Not clear	+	Included
15.	Coderre, S. P., Harasym, P., Mandin, H., & Fick, G. (2004). The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Medical Education*, *4*(1), 23.	+	Not clear	+	Not clear	None	Included
16.	McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical teacher*, *26*(8), 709-712.						Excluded, Review
17.	Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using	+	Not Clear	+	Not clear	Not clear	Included

	item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education*, *5*(1), 9.						
18.	Beullens, J., Struyf, E., & Van Damme, B. (2005). Do extended matching multiple-choice questions measure clinical reasoning?. *Medical education*, *39*(4), 410-417.	+	+	+			Included
19.	Beullens, J., Struyf, E., & Van Damme, B. (2006). Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. *Medical education*, *40*(12), 1173-1179.	+	Not clear	+	Not clear	Not clear	Included
20.	Samuels, A. (2006). Extended matching questions and the Royal Australian and New Zealand College of Psychiatrists written examination: an overview. *Australasian Psychiatry*, *14*(1), 63-66.						Excluded, Guideline
21.	Swanson, D. B., Holtzman, K. Z., Allbee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and	+	+	+	+	Not clear	Included.

	one-best-answer items in relation to number of options. *Academic Medicine*, *81*(10), S52-S55.						
22.	Burton, J. L. (2007). EMQS–how to write and how to answer. *Obstetrics, Gynaecology & Reproductive Medicine*, *17*(1), 25-27.						Exclude. Guideline
23.	Jamkar, A. V., Burdick, W., Morahan, P., Yemul, V. Y., & Singh, G. (2007). Proposed model of case based learning for training undergraduate medical student in surgery. *Indian Journal of Surgery*, *69*(5), 176-183.						Excluded, Irrelevant
24.	Lim, E. C. H., Seet, R. C. S., Oh, V. M., Chia, B. L., Aw, M., Quak, S. H., & Ong, B. K. (2007). Computer-based testing of the modified essay question: the Singapore experience. *Medical Teacher*, *29*(9-10), e261-e268.						Excluded, Irrelevant
25.	Fowell, S. L., Fewtrell, R., & McLaughlin, P. J. (2008). Estimating the minimum number of judges required for test-centred						Excluded, Irrelevant

	standard setting on written assessments. Do discussion and iteration have an influence?. *Advances in health sciences education*, *13*(1), 11-24.						
26.	McMahon, R. F., & Benbow, E. W. (2008). Designing assessment of pathology in the undergraduate curriculum. *Diagnostic Histopathology*, *14*(9), 453-458.						Excluded, Guideline
27.	O'Flynn, K. J. (2008). Introduction to the extended matching questions (EMQs). *British Journal of Medical and Surgical Urology*, *1*(1), 45-45.						Excluded, irrelevant
28.	Sturmberg, J. P., & Farmer, E. A. (2008). Assessing general practice knowledge base: The applied knowledge test. *Australian Journal of General Practice*, *37*(8), 659.	+	+	+	+	+	Included
29.	Swanson, D. B., Holtzman, K. Z., & Allbee, K. (2008). Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Academic Medicine*, *83*(10), S21-S24.	+	+	+	+		Included
30.	Tomlin, J. L., Pead, M. J., & May, S. A. (2008). Attitudes of veterinary faculty to the assessment of clinical reasoning using extended matching questions. *Journal of*						Excluded. Out of scope population

	veterinary medical education, 35(4), 622-630.						
31.	Tomlin, J. L., Pead, M. J., & May, S. A. (2008). Veterinary students' attitudes toward the assessment of clinical reasoning using extended matching questions. *Journal of veterinary medical education, 35*(4), 612-621.						Excluded. Out of scope population
32.	Baird, A. S. (2010). The new Extended Matching Question (EMQ) paper of the MFSRH examination. *J Fam Plann Reprod Health Care, 36*(3), 171-173.						Exclude. Guideline
33.	Dory, V., Gagnon, R., & Charlin, B. (2010). Is case-specificity content-specificity? An analysis of data from extended-matching questions. *Advances in health sciences education, 15*(1), 55-63.	+	Not clear	+	Not clear	Not clear	Included
34.	Harding, S., Britten, N., & Bristow, D. (2010). The performance of junior doctors in applying clinical pharmacology knowledge and prescribing skills to	+	+	+	+	Not clear	Included

	standardized clinical cases. *British journal of clinical pharmacology*, *69*(6), 598-606.						
35.	Chandratilake, M., Davis, M., & Ponnamperuma, G. (2011). Assessment of medical knowledge: the pros and cons of using true/false multiple choice questions. *Natl Med J India*, *24*(4), 225-8.						Excluded, irrelevant
36.	Gibson, S., & Leinster, S. (2011). How do students with dyslexia perform in extended matching questions, short answer questions and observed structured clinical examinations?. Advances in health sciences education, 16(3), 395-404.						Excluded, irrelevant
37.	McKendree, J., & Snowling, M. J. (2011). Examination results of medical students with dyslexia. *Medical education*, *45*(2), 176-182.						Excluded. Irrelevant
38.	D'Antoni, A. V., DiLandro, A. C., Chusid, E. D., & Trepal, M. J. (2012). Psychometric properties and podiatric medical student perceptions of USMLE-style items in a general anatomy course. *Journal of the*	+	+	+	+	+	Included

	American Podiatric Medical Association, *102*(6), 517-528.						
39.	Metcalfe, N. H. (2012). Testing the test: an analysis of the MRCGP Applied Knowledge Test as an assessment tool. *Education for Primary Care*, *23*(1), 13-18.	+		+		+	Included
40.	Nabishah, M., Nasri, A. B., Salam, A., Harlina, H. S., & Ima Nirwana, S. (2012). Setting the Standard of Student Performance: An Alternative Borderline Method. *International Medical Journal*, *19*(2).						Included
41.	Park, J. C., Kim, K. S., Park, J. C., & Kim, K. S. (2012). A comparison between discrimination indices and item-response theory using the Rasch Model in a clinical course written examination of a medical school. *Korean Journal of Medical Education*, *24*(1), 15-21.						Included
42.	van Bruggen, L., Manrique-van Woudenbergh, M., Spierenburg, E., & Vos, J. (2012). Preferred question types for						Excluded. Review

	computer-based assessment of clinical reasoning: a literature study. *Perspectives on medical education*, *1*(4), 162-171.						
43.	Eijsvogels, T. M., van den Brand, T. L., & Hopman, M. T. (2013). Multiple choice questions are superior to extended matching questions to identify medicine and biomedical sciences students who perform poorly. *Perspectives on medical education*, *2*(5-6), 252-263.			+	+	+	Included
44.	Kelly, M., Bennett, D., Bruce-Brand, R., O'Flynn, S., & Fleming, P. (2014). One week with the experts: a short course improves musculoskeletal undergraduate medical education. *JBJS*, *96*(5), e39.						Excluded. Irrelevant
45.	Jan, H., Guimicheva, B., Gosh, S., Hamid, R., Penna, L., & Sarris, I. (2014). Evaluation of healthcare professionals' understanding of eponymous maneuvers and mnemonics in emergency obstetric care provision. *International Journal of Gynecology & Obstetrics*, *125*(3), 228-231.						Excluded. Irrelevant
46.	Vorstenbosch, M. A., Bouter, S. T., van den Hurk, M. M., Kooloos, J. G., Bolhuis, S.						Included

	M., & Laan, R. F. (2014). Exploring the validity of assessment in anatomy: Do images influence cognitive processes used in answering extended matching questions?. *Anatomical sciences education*, *7*(2), 107-116.						
47.	Brenner, E., Chirculescu, A. R., Reblet, C., & Smith, C. (2015). Assessment in anatomy. *European Journal of Anatomy*, *19*(1), 105-124.						Excluded. Irrelevant
48.	Guraya, S. Y. (2016). The pedagogy of teaching and assessing clinical reasoning for enhancing the professional competence: a systematic review. *Biosciences Biotechnology Research Asia*, *13*(3), 1859-1866.						Excluded. Irrelevant
49.	Hippisley-Cox, J., & Coupland, C. (2017). Development and validation of risk prediction equations to estimate survival in patients with colorectal cancer: cohort study. *bmj*, *357*, j2497.						Excluded. Irrelevant

50.	Ortega-Morán, J. F., Pagador, J. B., Sánchez-Peralta, L. F., Sánchez-González, P., Noguera, J., Burgos, D., ... & Sánchez-Margallo, F. M. (2017). Validation of the three web quality dimensions of a minimally invasive surgery e-learning platform. *International journal of medical informatics*, *107*, 1-10.						Excluded, Irrelevant
51.	Humm, K. R., & May, S. A. (2018). Clinical Reasoning by Veterinary Students in the First-Opinion Setting: Is It Encouraged? Is It Practiced?. *Journal of veterinary medical education*, *45*(2), 156-162.						Excluded. Out of scope population
52.	Fox, M., Blake, D., & Jacobs, D. (2018). Veterinary parasitology teaching at London–Meeting the 'Day-One Competency'needs of new veterinarians. *Veterinary parasitology*, *254*, 131-134.						Excluded. Out of scope population
53.	Tan, K., Chin, H. X., Yau, C. W., Lim, E. C., Samarasekera, D., Ponnamperuma, G., & Tan, N. C. (2018). Evaluating a bedside tool for neuroanatomical localization with						Excluded, Guideline

	extended-matching questions. *Anatomical sciences education*, *11*(3), 262-269.						
54.	Tran, H., Ross, M. W., Diamond, P. M., Berg, R. C., Weatherburn, P., & Schmidt, A. J. (2018). Structural validation and multiple group assessment of the short internalized homonegativity scale in homosexual and bisexual men in 38 European countries: Results from the European MSM Internet Survey. *The Journal of Sex Research*, *55*(4-5), 617-629.						Excluded. Irrelevant
55.	Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., ... & Estrada, C. A. (2019). Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine*, *94*(6), 902-912.						Excluded. Review
56.	Kamath, P., Palacios, J. C., & Soares, M. R. (2019). Implementation of a Competency-based Pressure Ulcer Curriculum for Medical Students: Outcomes from an						Excluded. Irrelevant

	Educational Intervention Study. *Wound management & prevention*, *65*(4), 42-47.						
57.	NAZIM, S. M., TALATI, J. J., PINJANI, S., BIYABANI, S. R., ATHER, M. H., & NORCINI, J. J. (2019). Assessing clinical reasoning skills using Script Concordance Test (SCT) and extended matching questions (EMQs): A pilot for urology trainees. *Journal of Advances in Medical Education & Professionalism*, *7*(1), 7.	+	+	Not clear	+	Not clear	Included
58.	Sam, A. H., Peleva, E., Fung, C. Y., Cohen, N., Benbow, E. W., & Meeran, K. (2019). Very short answer questions: a novel approach to summative assessments in pathology. *Advances in Medical Education and Practice*, *10*, 943.						Excluded. Irrelevant
59.	AlShamlan, N. A., Al Shammari, M. A., Darwish, M. A., Sebiany, A. M., Sabra, A. A., & Alalmaie, S. M. (2020). Evaluation of Multifaceted Assessment of the Fifth-Year Medical Students in Family Medicine Clerkship, Saudi Arabia Experience.	Not clear	Not clear	+	+	Not clear	Included

	Journal of Multidisciplinary Healthcare, *13*, 321.						
60.	Henssen, D. J., van den Heuvel, L., De Jong, G., Vorstenbosch, M. A., van Cappellen van Walsum, A. M., Van den Hurk, M. M., ... & Bartels, R. H. (2020). Neuroanatomy Learning: Augmented Reality vs. Cross-Sections. *Anatomical sciences education*, *13*(3), 353-365.						Excluded. Irrelevant
61.	Isreb, S., Attwood, S., Hesselgreaves, H., McLachlan, J., & Illing, J. (2020). The Development of an Online Standalone Cognitive Hazard Training for Laparoscopic Cholecystectomy: A Feasibility Study. *Journal of Surgical Education*, *77*(1), 1-8.						Excluded. Irrelevant

Appendix 6: Medical Education Research Study Quality Instrument

Domain	MERSQI Item	Score	Maximum Score
Study design	Single Group Cross-sectional or single group posttest only	1	3
	Single group pretest & posttest	1.5	
	Nonrandomized, 2 groups	2	
	Randomized controlled trial	3	
Sampling	**Institutions studied:**		3
	1	0.5	
	2	1	
	3	1.5	
	Response rate, %:		
	Not applicable		
	<50 or not reported	0.5	
	50- 74	1	
	>75	1.5	
Type of data	Assessment by participants	1	3
	Objective measurement	3	
Validity of evaluation instrument	**Internal structure:**		3
	Not applicable		
	Not reported	0	
	Reported	1	
	Content:		
	Not applicable		
	Not reported	0	
	Reported	1	
	Relationships to other variables:		
	Not applicable		
	Not reported	0	
	Reported	1	
Data analysis	**Appropriateness of analysis:**		3
	Inappropriate for study design or type of data	0	

	Appropriate for study design, type of data	1	
	Complexity of analysis:		
	Descriptive analysis only	1	
	Beyond descriptive analysis	2	
Outcomes	Satisfaction, attitudes, perception, opinions, general facts	1	3
	Knowledge, skills	1.5	
	Behaviors	2	
	Patient/health care outcome	3	
Total possible score			**18**

Appendix 7: Dummy tables for presenting data

Table (1): Descriptive data of the included studies

Variable	Details	No (%)
Time period	1991 - 2000	
	2001 - 2010	
	2011 - 2020	
Country		
HPE* Discipline		

***** The defined group of health professions education

Table (2): Summary of the included studies

Author (Year)	Country	Population	Study design	Outcome*	Remarks

*** based on the framework explained by Downing**

Table (3): Pattern of reporting validity evidence

Source of validity evidence		Operational cue(s)	+/-	Exact sentence/phrase	Tool +/-	Test +/-
Content	Blueprint	Description and/or mentioning of test development using a blueprint				
	Items quality	Description and/or mentioning of revision and/or meeting for item quality				
	Qualifications of test developers	Description and/or mentioning of the specialties and/or qualifications of test developers				

Response process	Description and/or mentioning of quality control measures				
Internal structure	Description and/or mentioning of item analysis (item difficulties)				
	Description and/or mentioning of item analysis (discrimination)				
	Description and/or mentioning of item analysis (reliability coefficient)				
Relationship to other variables	Description and/or mentioning of correlations (negative or positive) to another measure				
Consequences	Description and/or mentioning the impact of the test scores on learners, teachers, or policy				

Table (4): quality of reporting validity evidence

Source of validity evidence	Number of articles (%)
One source	
Two sources	
Three sources	
Four sources	
Five sources	