DR. JAIME JORDAN (Orcid ID : 0000-0002-6573-7041)
DR. LAURA R HOPSON (Orcid ID : 0000-0002-1183-4751)
DR. SUZANNE BENTLEY (Orcid ID : 0000-0003-0192-3133)
DR. NICOLE M. DEIORIO (Orcid ID : 0000-0002-8123-1112)
DR. SALLY SANTEN (Orcid ID : 0000-0002-8327-8002)
DR. LALENA M YARRIS (Orcid ID : 0000-0003-1277-2852)
DR. WENDY C. COATES (Orcid ID : 0000-0002-3305-8802)
DR. MICHAEL A. GISONDI (Orcid ID : 0000-0002-6800-3932)

Article type : Original Contribution

Title: Leveling the field: Development of Reliable Scoring Rubrics for Quantitative and Qualitative Medical Education Research Abstracts

Running Title: Abstract Scoring Rubrics…

Authors: Jaime Jordan, MD, MAEd,[1,2] Laura R. Hopson, MD, [3] Caroline Molins, MD, MSMEd,[4] Suzanne K. Bentley, MD, MPH,[5] Nicole M. Deiorio, MD,[6] Sally A. Santen, MD, PhD,[6,7] Lalena M. Yarris, MD, MCR,[8] Wendy C. Coates, MD,[1] Michael A. Gisondi, MD[9]

Affiliations:

[1] Department of Emergency Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA
[2] Department of Emergency Medicine, Ronald Reagan UCLA Medical Center, Los Angeles, CA
[3]Department of Emergency Medicine, University of Michigan, Ann Arbor, MI
[4]AdventHealth Emergency Medicine Residency, Orlando, FL
[5]Icahn School of Medicine at Mount Sinai, New York, NY
[6]Virginia Commonwealth University School of Medicine, Richmond, VA
[7]University of Cincinnati College of Medicine
[8]Department of Emergency Medicine, Oregon Health & Science University, Portland, OR
[9]Department of Emergency Medicine, Stanford University, Palo Alto, CA

Please address correspondence to:
Jaime Jordan
UCLA Emergency Medicine
924 Westwood Boulevard, Suite 300
Los Angeles, CA 90024
Tel: 310-794-0585
Fax: 310-794-0599
Email: jaimejordanmd@gmail.com

Presentations: This work was presented at SAEM Annual Meeting, May 13th, 2021.
Financial support: none
Acknowledgements: We would like to acknowledge that this project originated to meet an SAEM Education Committee Objective and would like to thank all the committee members for their support of this work.
Conflicts of interest: JJ, LRH, CM, SKB, NMD, SAS, LMY, WCC, MAG report no conflict of interest.
Author contributions: JJ and MAG conceived the study. JJ, MAG, LRH, CM, and SKB contributed to the design of the study. JJ, LRH, CM, SKB, NMD, SAS, LMY, WCC, MAG

contributed to data collection. JJ analyzed the data. JJ, LRH, CM, SKB, NMD, SAS, LMY, WCC, MAG contributed to drafting of the manuscript and critical revision.

Abstract:

Background: Research abstracts are submitted for presentation at scientific conferences, however, criteria for judging abstracts are variable. We sought to develop two rigorous abstract scoring rubrics for education research submissions reporting (1) quantitative data and (2) qualitative data; and then to collect validity evidence to support score interpretation.

Methods: We used a modified Delphi method to achieve expert consensus for scoring rubric items to optimize content validity. Eight education research experts participated in two separate modified Delphi processes, one to generate quantitative research items and one for qualitative. Modifications were made between rounds based on item scores and expert feedback.

Homogeneity of ratings in the Delphi process was calculated using Cronbach's alpha, with increasing homogeneity considered an indication of consensus. Rubrics were piloted by scoring abstracts from 22 quantitative publications from *Academic Emergency Medicine Education and Training* "Critical Appraisal of Emergency Medicine Education Research" (11 highlighted for excellent methodology and 11 that were not) and 10 qualitative publications (5 highlighted for excellent methodology and 5 that were not). Intraclass correlation coefficient (ICC) estimates of reliability were calculated.

Results: Each rubric required three rounds of a modified Delphi process. The resulting quantitative rubric contained nine items: quality of objectives, appropriateness of methods, outcomes, data analysis, generalizability, importance to medical education, innovation, quality of writing, and strength of conclusions. Cronbach's alpha for the 3rd round=0.922; ICC for total scores during piloting = 0.893. The resulting qualitative rubric contained 7 items: quality of study aims, general methods, data collection, sampling, data analysis, writing quality, and strength of conclusions. Cronbach's alpha for the 3rd round = 0.913; ICC for the total scores during piloting =0.788.

Conclusion: We developed scoring rubrics to assess quality in quantitative and qualitative medical education research abstracts to aid in selection for presentation at scientific meetings. Our tools demonstrated high reliability.

Introduction

The scientific abstract is the standard method for researchers to communicate brief written summaries of their findings. The written abstract is the gatekeeper for selection for presentation at professional society meetings.[1] A research presentation serves many purposes including dissemination of new knowledge, an opportunity for feedback, and the prospect of fostering an investigator's academic reputation. Beyond the presentation, abstracts, as written evidence of scientific conference proceedings, often endure through publication in peer-reviewed journals. Because of the above, abstracts may be assessed in a number of potentially high-stakes situations.

Abstracts are selected for presentation at conferences through a competitive process based on factors such as study rigor, importance of research findings, and relevance to the sponsoring professional society. Prior literature has shown poor observer-agreement in the abstract selection process.[2] Scoring rubrics are often used to guide abstract reviewers in an attempt to standardize the process, reduce bias, support equity, and promote quality.[3] There are limited data describing the development and validity evidence of such scoring rubrics but the data available suggest that rubrics may be based on quality scoring tools for full research reports and published guidelines for abstracts.[2, 4-5] Medical conferences often apply rubrics designed for judging clinical or basic science submissions which reflect standard hypothesis-testing methods and often use a single subjective Gestalt rating for quality decisions.[6] This may result in the systematic exclusion of studies which employ alternate, but equally rigorous methods, such as research in medical education. Existing scoring systems, commonly designed for biomedical research, may not accurately assess the scope, methods, and types of results commonly reported in medical education research abstracts, which may lead to a disproportionately high rate of rejection of these abstracts. There are additional challenges in reviewing qualitative research abstracts using a standard hypothesis-testing rubric. In these qualitative studies, word-count constraints may limit the author's ability to convey the study's outcome appropriately.[7] It is

problematic for qualitative studies to be constrained to a standard quantitative abstract template, which may lead to low scores by those applying the rubric and a potential systematic bias against qualitative research.

Prior literature has described tools to assess quality in medical education research manuscripts, such as the Medical Education Research Study Quality Instrument (MERSQI) and the Newcastle-Ottawa Scale-Education (NOS-E).[8] A limited attempt to utilize the MERSQI tool to retrospectively assess internal medicine medical education abstracts achieving manuscript publication showed increased scores for the journal abstract relative to the conference abstract.[4] However, the MERSQI and similar tools were not developed specifically for judging abstracts, and there is a lack of published validity evidence to support score interpretation based on these tools. In order to equitably assess the quality of education research abstracts to scholarly venues, which may have downstream effects on researcher scholarship, advancement, and reputation, there is a need for a rigorously developed abstract scoring rubric that is based on a validity evidence framework.[9-10]

The aim of this paper is to describe the development and pilot testing of a dedicated rubric to assess the quality of both quantitative and qualitative medical education research studies. We describe the development process, which aimed to optimize content and response process validity, and initial internal structure and relation to other variables validity evidence to support score interpretation using these instruments. The rubrics may be of use to researchers developing studies, abstract and paper reviewers, and may be applied to medical education research assessment in other specialties.

Methods

Study Design

We utilized a modified Delphi technique to achieve consensus on items for a scoring rubric to assess quality of emergency medicine (EM) education research abstracts. The modified Delphi technique is a systematic group consensus strategy designed to increase content validity.[11] Through this method we developed individual rubrics to assess quantitative and qualitative EM medical education research abstracts. This study was approved by the Institutional Review Board of the David Geffen School of Medicine at UCLA.

178

179 **Study setting and participants**

180 The first author identified eight EM education researchers with successful publication
181 records from diverse regions across the United States and invited them to participate in the
182 Delphi panel. Previous work has suggested that 6-10 experts is an appropriate number for
183 obtaining stable results in the modified Delphi method.[12-14] All invited panelists agreed to
184 participate. The panel included one assistant professor, two associate professors, and 5
185 professors. All panelists serve as reviewers for medical education journals and four hold
186 editorial positions. We collected data in September and October, 2020.

187

188 **Study protocol**

189 We followed Messick's framework for validity that includes five types of validity
190 evidence; content, response process, internal structure, relation to other variables, and
191 consequential.[15] Our study team drafted initial items for the scoring rubrics after a review of the
192 literature and existing research abstract scoring rubrics to optimize content validity. We created
193 separate items for research abstracts reporting quantitative and qualitative data. We sent the
194 draft items to the Society of Academic Emergency Medicine (SAEM) education committee for
195 review and comment in order to gather stakeholder feedback and for further content and response
196 process validity evidence.[16] One author (JJ) who was not a member of the Delphi panel then
197 revised the initial lists of items based on committee feedback to create the initial Delphi
198 surveys. We used an electronic survey platform (SurveyMonkey) to administer and collect data
199 from the Delphi surveys.[17] Experts on the Delphi panel rated the importance of including each
200 item in a scoring rubric on a 1-9 Likert scale with 1 labeled as "not at all important" and 9
201 labeled as "extremely important". The experts were invited to provide additional written
202 comments, edits and suggestions for each item. They were also encouraged to suggest additional
203 items that they felt were important but not currently listed. We determined *a priori* that items
204 with a mean score of 7 or greater advanced to the next round and items with a mean score of 3 or
205 below were eliminated. The Delphi panel moderator (JJ) applied discretion for items scoring
206 between 4 and 6, with the aim of both adhering to the opinions of the experts and creating a
207 comprehensive scoring rubric. For example, if an item received a middle score but had

comments supporting inclusion in a revised form, the moderator would make the suggested revisions and include the item in the next round.

Each item consisted of a stem and anchored choices with associated point-value assignments. Panelists commented on the stems, content and assigned point-value of choices and provided narrative unstructured feedback. The moderator made modifications between rounds based on item scores and expert feedback. After each round, we provided panelists with aggregate mean item scores, written comments, and an edited version of the item list derived from the responses in the previous round. The panelists were then asked to rate the revised items and provide additional edits or suggestions.

We considered homogeneity of ratings in the Delphi process to be an indication of consensus. After consensus was achieved, we created final scoring rubrics for quantitative and qualitative medical education research abstracts. We then piloted the scoring rubrics to gather internal structure and further response process validity evidence. Five raters from the study group (JJ, LH, MG, CM, SB) participated in piloting. We piloted the final quantitative research rubric by scoring abstracts from publications identified in the most recent critical appraisal of EM education research by Academic Emergency Medicine/Academic Emergency Medicine Education and Training, "Critical Appraisal of Emergency Medicine Education Research: The Best Publications of 2016".[18] All 11 papers highlighted for excellent methodology in this issue were included in the pilot.[18] Additionally, we included an equal number of randomly selected citations that were included in the issue but not selected as top papers, for a total of 22 quantitative publications.[18] Given the limited number of qualitative studies cited in this issue of the critical appraisal series, we chose to pilot the qualitative rubric on publications from this series from the last 5 years available (2012-2016).[18-22] We randomly selected one qualitative publication that was highlighted for excellent methodology and one that was not from each year for a total of 10 qualitative publications.[18-22] The same five raters who performed the quantitative pilot also conducted the qualitative pilot.

Statistical Analysis

We calculated and reported descriptive statistics for item scoring during Delphi rounds. We used Cronbach's alpha to assess homogeneity of ratings in the Delphi process. Increasing homogeneity was considered to be an indication of consensus among the expert

panelists. We used intraclass correlation coefficient (ICC) estimates to assess reliability among raters during piloting based on a mean rating (k=5), absolute agreement, 2-way random-effects model. We performed all analyses in SPSS (IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp).

Results

Quantitative rubric:

Three Delphi rounds were completed, each with 100% response rate. Mean item scores for each round are depicted in Table 1. After the first round, three items were deleted, one item was added, and five items underwent wording changes. After the second round, one item was deleted and eight items underwent wording changes. After the third round items were re-ordered for flow and ease of use but no further changes were made to content or wording. Cronbach's alpha for the third round was 0.922 indicating high internal consistency. The final rubric contained nine items: quality of objectives, appropriateness of methods, outcomes, data analysis, generalizability, importance to medical education, innovation, quality of writing, and strength of conclusions (Appendix A). The ICC for the total scores during piloting was 0.893, indicating excellent agreement. ICCs for individual rubric items ranged from 0.406 to 0.878 (Table 3).

Qualitative rubric:

Three Delphi rounds were completed, each with 100% response rate. Mean item scores for each round are depicted in Table 2. After the first round two items were deleted, one item was added and nine items underwent wording changes. After the second round, three items were deleted and four underwent wording changes. After the third round no further changes were made. The resulting tool contained 7 items reflecting the domains of quality of study aims, general methods, data collection, sampling, data analysis, writing quality, and strength of conclusions (Appendix B). Cronbach's alpha for the third round was 0.913, indicating high internal consistency. ICC for the total scores during piloting was 0.788 indicating good agreement. The item on writing quality had an ICC of -0.301, likely due to the small scale of the item and sample size leading to limited variance. ICCs for the remainder of the items ranged from 0.176 to 0.897 (Table 3).

Discussion:

We developed novel and distinct abstract scoring rubrics for assessing quantitative and qualitative medical education abstract quality through a Delphi process. It is important to evaluate medical education research abstracts that utilize accepted education methods as a distinctly different class than basic, clinical, and translational research. Through our Delphi and piloting processes we have provided multiple types of validity evidence in support of these rubrics aligned with Messick's framework including content, response process and internal structure.[15] Similar to other tools assessing quality in medical education research, our rubrics assess aspects such as study design, sampling, data analysis, and outcomes that represent the underpinnings of rigorous research.[8, 23-26] Unlike many medical education research assessments published in the literature, our tool was designed specifically for the assessment of abstracts rather than full text manuscripts and therefore the specific item domains and characteristics reflect this unique purpose.

We deliberately created separate rubrics for abstracts reporting quantitative and qualitative data as each has unique methods. When designing a study, education researchers must decide the best method to address their questions. Often, in the exploratory phase of inquiry, a qualitative study is the most appropriate choice to identify key topics that merit further study. These often may be narrow in scope and may employ one or more qualitative methods (e.g., ethnography, focus groups, personal interviews). The careful and rigorous analysis may reveal points that can be studied later via quantitative methods to test a hypothesis gleaned during the qualitative phase.[27] Specific standards for reporting on qualitative research have been widely disseminated and are distinct from standards for reporting quantitative research.[28] Even an impeccably designed and executed qualitative study would fail to meet major criteria for excellent quantitative studies. For example, points may be subtracted for lack of generalizability or conduct of the qualitative study in multiple institutions, as well as for the absence of common quantitative statistical analytics. The qualitative abstract itself may necessarily lack the common structure of a quantitative report and lead to a lower score. The obvious problem is that a well-conducted study might not be shared with the relevant research community if it is judged according to quantitative standards. A similar outcome would occur if quantitative work were

judged by qualitative standards, therefore we advocate for using scoring rubrics specific to the type of research being assessed.

Our work has several possible applications. The rubrics we developed may be adopted as scoring tools for medical education research studies that are submitted for presentation to scientific conferences. The presence of specific scoring rubrics for medical education research may address disparities in acceptance rates and ensure presentation of rigorously conducted medical education research at scientific conferences. Further, publication of abstract scoring rubrics such as ours sets expectations for certain elements to be included and defines an acceptable level of submission quality. Dissemination and usage of the rubrics may therefore help improve research excellence. The rubrics themselves can serve as educational tools in resident and faculty training. For example, the rubrics could serve as illustrations or practice material in teaching how to prepare a strong abstract for submission. The inclusive wording of the items allows the rubrics to be adapted to medical education work in any medical specialty. Medical educators may also benefit from using the methods described here to create their own scoring rubrics or provide evidence-based best practice approaches for other venues. Finally, this study provides a tool that could lay the groundwork for future scholarship on assessing the quality of educational research.

Limitations

Our study has several limitations. First, the modified Delphi technique is a consensus technique which can force agreement of respondents and the existence of consensus does not denote a correct response.[11] Since the method is implemented electronically, there is limited discussion and elaboration. Second, the team of experts were all researchers in EM, therefore the rubrics may not generalize to other specialties. The rubrics were intended for quantitative and qualitative education research abstract submission, so it may not perform well for abstracts that include *both* quantitative and qualitative data or those focused on early work, innovations, instrument development, validity evidence, or program evaluation. Finally, there are two limitations to the pilot testing. An *a priori* power calculation to determine sample size was not possible since the rubrics were novel. The ICCs of individual items on the scoring rubrics were variable and we chose not to eliminate items with low ICCs given the small sample size during piloting and a desire to create a tool comprehensive of key domains. Future studies of use of

these tools incorporating larger samples may provide data for additional refinement. Faculty who piloted the rubrics were familiar with the constructs and rubrics, and it is not known how the rubrics would have performed with general abstract reviewers nor what training might be required. The success of separate rubrics may rely on the expertise of the reviewers in the methodology being assessed.

We offer two medical education abstract scoring rubrics with supporting preliminary reliability and validity evidence. Future studies could add additional validity evidence including use with trained and untrained reviewers and relationship to other variables, e.g. a comparison between rubric scores and expert judgement. Additional studies could be done to provide consequential validity evidence by comparing the number and quality of accepted medical education abstracts before and after the rubric's implementation or whether the number of abstracts that eventually lead to publication increases.

Conclusions

Using the modified Delphi technique for consensus building, we developed two scoring rubrics to assess quality in quantitative and qualitative medical education research abstracts with supporting validity evidence. Application of these rubrics demonstrated high reliability.

References

1. Padayachy A, Rodrigues G, Tahar A. Comment rédiger un abstract scientifique ? [How to write a scientific abstract]. Rev Med Suisse 2019;15(664):1703-1706.

2. Timmer A, Sutherland LR, Hilsden RJ. Development and evaluation of a quality score for abstracts. BMC Med Res Methodol 2003;3:2.

3. Mitchell NS, Stolzmann K, Benning LV, Wormwood JB, Linsky AM. Effect of a Scoring Rubric on the Review of Scientific Meeting Abstracts. J Gen Intern Med 2020. Epub ahead of print.

4. Poolman RW, Keijser LC, de Waal Malefijt MC, et al. Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings. Acta orthop 2007;78(2):278–284.

5. van der Steen LP, Hage JJ, Kon M, et al. Reliability of a structured method of selecting abstracts for a plastic surgical scientific meeting. Plast Reconstr Surg 2003;111(7):2215–2222.

6. Kuczmarski TM, Raja AS, Pallin DJ. How do Medical Societies Select Science for Conference Presentation? How Should They? West J Emerg Med 2015;16(4):543-50.
7. Stephenson CR, Vaa BE, Wang AT, et al. Conference presentation to publication: a retrospective study evaluating quality of abstracts and journal articles in medical education research. BMC Med Educ 2017;17(1):193.
8. Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med 2015;90(8):1067-76.
9. Sullivan GM. A primer on the validity of assessment instruments. J Grad Med Educ 2011;3: 119-120.
10. Cook DA, Beckman TJ. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. Am J of Medicine 2006:119:166e10-199.e16.
11. Hasson F, Keeney S, Mckenna HP. Research guidelines for the Delphi Survey Technique Research guidelines for the Delphi survey. J Adv Nurs 2000;32(4):1008-15.
12. Thangaratinam SK, Redman CWE. The Delphi Technique. The Obstetrician and Gynaecologist 2005;7:120-125.13. Pines JM, Alfaraj S, Batra S, et al. Factors important to top clinical performance in emergency medicine residency: results of an ideation survey and Delphi panel. AEM Educ Train 2018;2(4):269-276.
14. Eubank BH, Mohtadi NG, Lafave MR, et al. Using the modified Delphi method to establish clinical consensus for the diagnosis and treatment of patients with rotator cuff pathology. BMC Medical Research Methodology 2016;16:56.
15. Messick S. Validity. In Linn R, ed. Educational Measurement, 3rd Edition. New York, NY: American Council on Education and Macmillan, 1989.
16. Society of Academic Emergency Medicine. Join a committee. *(Accessed on July 5th at https://www.saem.org/about-saem/saem-membership/committees)*
17. SurveyMonkey. *(Accessed on April 7th at https://www.surveymonkey.com/)*
Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev. 2007 Apr;18(2):MR000005.
18. Dubosh NM, Jordan J, Yarris LM, et al. Critical Appraisal of Emergency Medicine Educational Research: The Best Publications of 2016. AEM Educ Train 2018;3(1):58-73.

19. Lin M, Fisher J, Coates WC, et al. Critical Appraisal of Emergency Medicine Education Research: The Best Publications of 2012. Acad Emerg Med 2014;21:322-333.
20. Farrell SE, Kuhn GJ, Coates WC, et al. Critical Appraisal of Emergency Medicine Education Research: The Best Publications of 2013. Acad Emerg Med 2014;21(11):1274-83.
21. Yarris LM, Juve AM, Coates WC, et al. Critical Appraisal of Emergency Medicine Education Research: The Best Publications of 2014. Acad Emerg Med 2015; 22(11):1327-36.
22. Heitz CR, Coates WC, Farrell SE, Fisher J, Juve AM, Yarris LM. Critical Appraisal of Emergency Medicine Education Research: The Best Publications of 2015. AEM Educ Train 2017;1(4):255-268.
23. Moralejo D, Ogunremi T, Dunn K. Critical appraisal toolkit (CAT) for assessing multiple types of evidence. Can Commun Dis Rep 2017;43(9):176-181.
24. Reed DA, Beckam TJ, Wright SM, Levine RB, Kern DE, Cook DA. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's medical education special issue. J Gen Intern Med 2008;23(7):903-907.
25. Johnson JL, Adkins D, Chauvin S. A review of the quality indicators of rigor in qualitative research. American Journal of Pharmaceutical Education 2020;84(1):7120.
26. Yang LJ, Chang KW, Chung KC. Methodologically rigorous clinical research. Plast Reconstr Surg 2012;129(6):979e-988e.
27. Schneider NC, Coates WC, Yarris LM. Taking Your Qualitative Research to the Next Level: A Guide for the Medical Educator. AEM Educ Train 2017;1(4):368-378.
28. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med. 2014;89(9):1245-51.

Table 1. Items and mean scores of expert review during Delphi process for quantitative scoring rubric

Item	Mean score

	(SD) N= 8
Round 1	
Clarity of objectives 0 = No clear objective or hypothesis 1 = Objective(s) are stated but unclear 2 = Clearly stated objective(s)	8.88 (0.35)
Quality of objectives 0 = No stated objective or hypothesis 1 = Poorly chosen objective(s) or stated hypothesis is difficult to test 2 = Well thought out study objective(s) or testable hypothesis	7.71 (1.70)
Study design 0 = Inappropriate study design for objective(s) 0.5 = Single group cross sectional or single group post-test only 1 = Single group pre-test and post-test 1.5 = Two or more non-randomized groups (quasi-experimental study) 2 = Two or more randomized groups (experimental study)	6.5 (2.07)
Sampling: institutions 0 = Single institution 2 = Multi-institutional	5.38 (1.92)
Sampling: response rate 0 = Less than 50% or not reported 1 = 50-74% 2 = Greater than or equal to 75%	5.29 (2.43)
Type of data 0 = Not described 1 = Assessment by study participant 1.5 = Subjective assessment by someone other than the study participant (i.e. an observer) 2 = Objective assessment	6.50 (2.67)
Power/sample size 0 = No power/sample size calculation was performed 2 = A power/sample size calculation was calculated and satisfied	5.63 (2.83)

Data Analysis 0 = No analysis described or inappropriate data analysis for study design 1 = Descriptive analysis only (i.e. frequency, mean, median) 2 = Beyond descriptive analysis (i.e. any comparative statistics or test of statistical inference)	7.88 (0.99)
Generalizability 0 = not at all generalizable, results are only applicable to very specific population/setting 0.5 = Minimally generalizable 1 = Moderately generalizable 1.5 = Very generalizable, results apply to most EM educational populations/settings 2 = Extremely generalizable, results apply to educational populations/settings beyond EM	8.13 (0.64)
Relevance and importance of topic to medical education 0 = This topic is only of interest to a very small group of people and is unlikely to result in important knowledge 0.5 = This is an important topic to EM medical education that will lead to information of interest to many EM educators and learners 1 = This topic is essential to EM medical education and is likely to be important and relevant for every EM educator and learner to know 2 = This topic is essential to medical education other specialties beyond EM and is likely to be important for every medical educator and learner to know	7.5 (2.73)
Publication readiness/quality of writing 0 = Poorly written, unclear, difficult to understand 1 = Generally well-written, but leaves room for confusion on some concepts or has one or two errors 2= Exceptionally well-written, clear, logical organization and presentation of ideas.	7.38 (1.85)
Outcome(s) 0.5 = Kirkpatrick level 1 – satisfaction, attitudes, perceptions, opinions, general facts (i.e. demographics) 1 = Kirkpatrick level 2 – knowledge, skills (includes behaviors in a test setting such as simulation) 1.5 = Kirkpatrick level 3 – behaviors in real context or clinical setting	7.63 (2.13)

2 = Kirkpatrick level 4 = patient or health care outcome (actual effects on real patients, programs, or society)	
Innovation of study 0 = Not innovative or novel 1 = Moderately innovative (i.e. new method of instructing in a standard environment or standard instructional method in a novel area/environment) 2 = Completely novel idea	7.25 (1.39)
Global Rating 0 = No clear conclusions can be drawn 0.5 = Results ambiguous but appears to show a trend 1 = Conclusions can probably be based on results 1.5 = Results are clear and likely to be true 2 = Results are unequivocal	8.00 (1.31)
Round 2	
Quality of objectives 0 = No stated objective 1 = Poorly chosen or ambiguous objective(s) 2 = Clear, well thought out objective(s)	9.00 (0)
Appropriateness of methods 0 = Inappropriate methods for objective(s) 1 = Chosen methods were sub-optimal, but did address the objective(s) (i.e. acceptable methodology) 2 = Chosen methods were the best feasible for the objective(s) (i.e. rigorous methodology)	8.38 (1.06)
Study design 0 = Study design not described 0.5 = Single group cross sectional or single group post-assessment only 1 = Single group pre- and post-assessment 1.5 = Two or more non-randomized groups (quasi-experimental study) 2 = Two or more randomized groups (experimental study)	5.25 (2.66)
Data Analysis 0 = No analysis described or inappropriate data analysis for study design 1 = Descriptive analysis only (i.e. frequency, mean, median)	7.50 (1.31)

2 = Beyond descriptive analysis (i.e. any comparative statistics or test of statistical inference)	
Generalizability 0 = Results are only applicable to a very specific population/setting 1 = Results are applicable to most EM educational populations/settings 2 = Results are applicable to educational populations/settings beyond EM.	7.00 (1.51)
Relevance and importance of topic to medical education 0 = This topic is only of interest to a very small group of people and is unlikely to result in important knowledge 0.5 = This is an important topic to EM medical education that will lead to information of interest to many EM educators and learners 1 = This topic is essential to EM medical education and is likely to be important and relevant for every EM educator and learner to know 2 = This topic is essential to medical education in other specialties beyond EM and is likely to be important for every medical educator and learner to know	7.00 (1.31)
Publication readiness/quality of writing 0 = Poorly written, unclear, difficult to understand 1 = Generally well-written, but leaves room for confusion on some concepts or has one or two errors 2= Exceptionally well-written, clear, logical organization and presentation of ideas.	7.25 (2.05)
Outcome(s) 0 = Chosen outcomes are inappropriate for study objective 0.5 = Kirkpatrick level 1 – satisfaction, attitudes, perceptions, opinions, general facts (i.e. demographics) 1 = Kirkpatrick level 2 – knowledge, skills (includes behaviors in a test setting such as simulation) 2 = Kirkpatrick level 3 – behaviors in real context or clinical setting 3 = Kirkpatrick level 4 = patient or health care outcome (actual effects on real patients, programs, or society)	6.25 (2.25)
Innovation of study 0 = Not innovative or novel 1 = Moderately innovative (i.e. new method of instructing in a standard environment or standard instructional method in a novel area/environment)	7.75 (1.04)

2 = Completely novel idea	
Strength of conclusion(s) 0 = No clear conclusions can be drawn 0.5 = Results ambiguous but appears to show a trend 1 = Conclusions can probably be based on results 1.5 = Conclusions are clear and likely to be true 2 = Conclusions are unequivocal	7.00 (1.51)
Round 3	
Quality of objectives 0 = No stated objective 1 = Poorly chosen or ambiguous objective(s) 2 = Clear, well thought out objective(s) that logically follow from the background information	8.63 (0.52)
Appropriateness of methods 0 = Inappropriate methods for objective(s) 1 = Chosen methods were sub-optimal, but did address the objective(s) 2 = Chosen methods were the best feasible for the objective(s) (i.e. rigorous methods)	8.75 (0.46)
Data analysis 0 = No analysis described or inappropriate data analysis for study design 1 = Descriptive analysis only (e.g frequency, mean, median) 2 = Beyond descriptive analysis (e.g. any comparative statistics or test of statistical inference)	8.38 (0.74)
Generalizability 0 = Results are only applicable to a very specific population/setting 1 = Results are applicable to most EM educational populations/settings 2 = Results are applicable to educational populations/settings beyond EM.	7.25 (1.58)
Relevance and importance of topic to medical education 0 = This topic is only of interest to a very small group of people and is unlikely to result in important knowledge 1 = This topic is essential to EM medical education and is likely to be important and relevant for every EM educator and learner to know 2 = This topic is essential to medical education in other specialties beyond EM and is likely to be important for every medical educator and learner to know	6.88 (1.46)

Quality of writing 0 = Poorly written, unclear, difficult to understand 0.5 = Generally well-written 1= Exceptionally well-written, clear, logical organization and presentation of ideas.	7.50 (1.93)
Outcome(s) 0 = Chosen outcomes are inappropriate for study objective 1 = Chosen outcomes are reasonable for study objective, but not the best measure 2 = Chosen outcomes are ideal for study objective	8.50 (0.93)
Innovation of study 0 = Not innovative or novel 1 = Moderately innovative (e.g. new method of instructing in a standard environment or standard instructional method in a novel area/environment) 2 = Completely novel idea (e.g. new method of instructing in a novel area/environment)	7.63 (1.19)
Strength of conclusion(s) 0 = No clear conclusions can be drawn or conclusions do not follow directly from results 1 = Conclusions can probably be based on results 2 = Conclusions are unequivocal	8.25 (0.89)

Table 2. Items and mean scores of expert review during Delphi process for qualitative scoring rubric

Item	Mean score (SD)

	N = 8
Round 1	
Quality of objectives 0 = No stated objective 1 = Poorly chosen or ambiguous objective(s) 2 = Clear, well thought out objective(s) that logically follow from the background information	8.13 (1.36)
Study design 0 = Qualitative design is not appropriate for study objective(s) 1 = Qualitative approach is appropriate for study objective, but specific design not identified (i.e. phenomenology, ethnography, grounded theory, etc.) 2 = Specific qualitative design identified and appropriate for study objective	8.25 (0.89)
Data collection methods 0 = Data collection methods (participant observation, interviews, document review, etc.) not identified 1 = Data collection methods identified but inappropriate for study objective 2 = Data collection methods identified and appropriate for study objective	7.88 (1.64)
Sampling: method (Sampling is defined as the process of selecting participants) 0 = sampling method not described 1 = sampling method described, but not clear or not theoretically justified 2 = Clear description of sampling method that is theoretically justified	7.25 (1.49)
Sampling: saturation (Saturation is defined as the point at which no new information is being learned from continued data collection) 0 = Saturation of data not achieved or not described 2 = Saturation of data achieved	4.75 (2.92)
Trustworthiness (Trustworthiness is a marker of quality and can be supported with evidence of credibility, transferability, dependability, and confirmability) 0 = No clear description of researcher role, study context, or triangulation 1 = Provides some evidence of trustworthiness, but not comprehensive 2 = Provides significant evidence of trustworthiness such as clear description of researcher role, study context and triangulation	6.75 (1.49)
Data Analysis 0 = No analysis described or inappropriate data analysis for study objectives/design	7.50 (2.00)

1 = Some description of data analyses, but not entirely clear 2= In depth description of systematic data analyses appropriate to study objective with clear description of how themes and concepts were derived	
Relevance and importance of topic to medical education 0 = This topic is only of interest to a very small group of people and is unlikely to result in important knowledge 1 = This topic is essential to EM medical education and is likely to be important and relevant for every EM educator and learner to know 2 = This topic is essential to medical education in other specialties beyond EM and is likely to be important for every medical educator and learner to know	7.50 (2.07)
Quality of writing 0 = Poorly written, unclear, difficult to understand 0.5 = Generally well-written 1= Exceptionally well-written, clear, logical organization and presentation of ideas.	7.50 (2.00)
Innovation of study 0 = Not innovative or novel 1 = Moderately innovative 2 = Innovative or novel	6.00 (2.51)
Strength of conclusion(s) 0 = No clear conclusions can be drawn or conclusions do not follow directly from results 1 = Conclusions can probably be based on results 2 = Conclusions are unequivocal	7.63 (1.69)
Round 2	
Quality of study aims/objectives 0 = No stated aim or objective 1 = Poorly chosen or ambiguous aim/objective(s) 2 = Clear, well thought out aim/objective(s) that logically follow from the background information	8.75 (0.46)
General methods 0 = Qualitative methods are not appropriate for study aim/objective(s) 1 = Qualitative methods are appropriate for study aim/objective(s), but specific approach (e.g. phenomenology, ethnography, grounded theory, etc.) or paradigm (e.g. postpositivist, constructivist/interpretivist) not stated or not ideal	8.13 (0.83)

2 = Specific qualitative approach and paradigm stated and aligned with study aim/ objective(s)	
Data collection 0 = Data collection methods (observation, interviews, document review, etc.) not identified or inappropriate for study aim/objective(s) 1 = Data collection methods appropriate for study aim/objective(s), but not ideal 2 = Data collection methods are ideal for study aim/objective(s)	7.63 (1.06)
Sampling (Sampling is defined as the process of selecting participants) 0 = Sampling not described 1 = Sampling described, but flawed (e.g. unclear, inappropriate, not theoretically justified) 2 = Sampling clearly described and theoretically justified	7.50 (0.76)
Trustworthiness (Trustworthiness is a marker of quality and can be supported with evidence of credibility, transferability, dependability, confirmability, and reflexivity. Examples of specific techniques used to enhance trustworthiness include member checking, audit trail, triangulation, etc.) 0 = No clear description of methods to enhance trustworthiness. 1 = Provides some evidence of trustworthiness, but not comprehensive 2 = Provides significant evidence of trustworthiness such as clear description of researcher role, member checking, audit trail, study context or triangulation, with supported rationale	6.88 (2.59)
Data Analysis 0 = No analysis described or inappropriate data analysis for study objectives/design 1 = Some description of data analyses, but unclear or not justified 2= In depth description of systematic data analyses appropriate to study objective with clear description of how themes and concepts were derived	7.75 (1.39)
Importance of topic to medical education 0 = This topic is unlikely to result in important knowledge 1 = This topic is essential to EM medical education and is likely to be important for EM educators and learners to know 2 = This topic is essential to medical education in other specialties beyond EM and is likely to be important for medical educators and learners to know 2 = This topic is essential to medical education in other specialties beyond EM	6.38 (2.50)

and is likely to be important for medical educators and learners to know	
Quality of writing 0 = Poorly written, unclear, difficult to understand 0.5 = Generally well-written 1= Consistently well-written, clear, logical organization and presentation of ideas.	7.13 (2.36)
Strength of conclusion(s) 0 = No clear conclusions can be drawn or conclusions do not follow directly from results 1 = Conclusions can probably be based on results, but inference is necessary to draw conclusions 2 = Conclusions are well supported by results	8.25 (0.89)
Study implications 0 = Does not provide valuable information for future research 1 = Provides information that contributes to the field, but has limited implications for future research 2 = Provides a foundation for future hypothesis testing research	6.00 (1.93)
Round 3	
Quality of study aims/objectives 0 = No stated aim or objective 1 = Poorly chosen or ambiguous aim/objective(s) 2 = Clear, well thought out aim/objective(s) that logically follow from the background information	8.88 (0.35)
General methods 0 = Qualitative methods not appropriate for study aim/objective(s) 1 = Qualitative methods appropriate for study aim/objective(s), but specific approach (e.g. phenomenology, ethnography, grounded theory, etc.) or paradigm (e.g. postpositivist, constructivist/interpretivist) not stated or not ideal 2 = Specific qualitative approach and paradigm stated and aligned with study aim/ objective(s)	8.38 (0.52)
Data collection 0 = Data collection methods (observation, interviews, document review, etc.) not identified or inappropriate for study aim/objective(s) 1 = Data collection methods appropriate for study aim/objective(s), but not ideal 2 = Data collection methods ideal for study aim/objective(s)	8.00 (1.07)

Sampling (Sampling is defined as the process of selecting participants) 0 = Sampling not described 1 = Sampling described, but flawed (e.g. unclear, inappropriate, not theoretically justified) 2 = Sampling clearly described and theoretically justified	7.63 (0.74)
Data Analysis 0 = No analysis described or inappropriate data analysis for study objectives/design 1 = Some description of data analyses, but unclear or not justified 2= In depth description of systematic data analyses appropriate to study objective with clear description of how themes and concepts were derived	8.50 (0.76)
Quality of writing 0 = Poorly written, unclear, difficult to understand 1= Consistently well-written, clear, logical organization and presentation of ideas.	8.00 (1.20)
Strength of conclusion(s) 0 = No clear conclusions can be drawn or conclusions do not follow directly from results 1 = Conclusions require reader inference to draw conclusions 2 = Conclusions are well supported by results	8.38 (0.74)

Table 3. Interrater Reliability Results During Piloting

Item	Intraclass correlation coefficient [95% CI]
Quantitative rubric	

1. Quality of objectives	0.406 [-0.006, 0.705]
2. Appropriateness of methods	0.821 [0.671, 0.916]
3. Outcome(s)	0.661 [0.365, 0.843]
4. Data analysis	0.753 [0.548, 0.883]
5. Generalizability	0.878 [0.767, 0.944]
6. Relevance and importance of topic to medical education	0.747 [0.530, 0.882]
7. Innovation of study	0.786 [0.607, 0.900]
8. Quality of writing	0.726 [0.500, 0.870]
9. Strength of conclusions	0.739 [0.512, 0.878]
Total score	**0.893 [0.802, 0.950]**
Qualitative rubric	
1. Quality of objectives	0.176 [-0.466, 0.711]
2. General methods	0.897 [0.749, 0.971]
3. Data collection	0.635 [0.158, 0.892]
4. Sampling	0.531 [-0.106, 0.863]
5. Data analysis	0.874 [0.574, 0.950]
6. Quality of writing	-0.301 [-1.083, 0.489]
7. Strength of conclusions	0.753 [0.415, 0.927]
Total score	**0.788 [0.469, 0.939]**

Appendix A. Quantitative Education Research Abstract Scoring Rubric

1. Quality of objectives

0 = No stated objective

1 = Poorly chosen or ambiguous objective(s)

2 = Clear, well thought out objective(s) that logically follow from the background information

2. Appropriateness of methods

0 = Inappropriate methods for objective(s)

1 = Chosen methods were sub-optimal, but did address the objective(s)

2 = Chosen methods were the best feasible for the objective(s) (i.e. rigorous methods)

3. Outcome(s)

0 = Chosen outcomes are inappropriate for study objective

1 = Chosen outcomes are reasonable for study objective, but not the best measure

2 = Chosen outcomes are ideal for study objective

4. Data analysis

0 = No analysis described or inappropriate data analysis for study design

1 = Descriptive analysis only (e.g frequency, mean, median)

2 = Beyond descriptive analysis (e.g. any comparative statistics or test of statistical inference)

5. Generalizability

0 = Results are only applicable to a very specific population/setting

1 = Results are applicable to most EM educational populations/settings

2 = Results are applicable to educational populations/settings beyond EM.

6. Relevance and importance of topic to medical education

0 = This topic is only of interest to a very small group of people and is unlikely to result in important knowledge

1 = This topic is essential to EM medical education and is likely to be important and relevant for every EM educator and learner to know

2 = This topic is essential to medical education in other specialties beyond EM and is likely to be important for every medical educator and learner to know

7. Innovation of study

0 = Not innovative or novel

1 = Moderately innovative (e.g. new method of instructing in a standard environment or standard instructional method in a novel area/environment)

2 = Completely novel idea (e.g. new method of instructing in a novel area/environment)

8. Quality of writing

0 = Poorly written, unclear, difficult to understand

0.5 = Generally well-written

1= Exceptionally well-written, clear, logical organization and presentation of ideas.

9. Strength of conclusion(s)

0 = No clear conclusions can be drawn or conclusions do not follow directly from results

1 = Conclusions can probably be based on results

2 = Conclusions are unequivocal

555
556
557
558
559
560
561
562

Appendix B. Qualitative Education Research Abstract Scoring Rubric

564

Are you familiar with qualitative research study design?

Yes: Proceed with scoring

No: Decline

568

1. Quality of study aims/objectives

0 = No stated aim or objective

1 = Poorly chosen or ambiguous aim/objective(s)

2 = Clear, well thought out aim/objective(s) that logically follow from the background information

574

2. General methods

0 = Qualitative methods not appropriate for study aim/objective(s)

1 = Qualitative methods appropriate for study aim/objective(s), but specific approach (e.g. phenomenology, ethnography, grounded theory, etc.) or paradigm (e.g. postpositivist, constructivist/interpretivist) not stated or not ideal

2 = Specific qualitative approach and paradigm stated and aligned with study aim/ objective(s)

581

3. Data collection

0 = Data collection methods (observation, interviews, document review, etc.) not identified or inappropriate for study aim/objective(s)

1 = Data collection methods appropriate for study aim/objective(s), but not ideal

2 = Data collection methods ideal for study aim/objective(s)

4. Sampling (Sampling is defined as the process of selecting participants)

0 = Sampling not described

1 = Sampling described, but flawed (e.g. unclear, inappropriate, not theoretically justified)

2 = Sampling clearly described and theoretically justified

5. Data Analysis

0 = No analysis described or inappropriate data analysis for study objectives/design

1 = Some description of data analyses, but unclear or not justified

2= In depth description of systematic data analyses appropriate to study objective with clear description of how themes and concepts were derived

6. Quality of writing

0 = Poorly written, unclear, difficult to understand

1= Consistently well-written, clear, logical organization and presentation of ideas.

7. Strength of conclusion(s)

0 = No clear conclusions can be drawn or conclusions do not follow directly from results

1 = Conclusions require reader inference to draw conclusions

2 = Conclusions are well supported by results