# The Revised METRIQ Score: A Quality Evaluation Tool for Online Educational Resources

Isabelle N. Colmers-Gray, MD, MSc[1], Keeth Krishnan, MD[2], Teresa M. Chan, MD, MHPE[3], N. Seth Trueger, MD, MPH[4], Michael Paddock, DO, MS[5], Andrew Grock, MD[6], Fareen Zaver, MD[7], and Brent Thoma, MD, MA, MSc[8]

## ABSTRACT

**Background:** With the rapid proliferation of online medical education resources, quality evaluation is increasingly critical. The Medical Education Translational Resources: Impact and Quality (METRIQ) study evaluated the METRIQ-8 quality assessment instrument for blogs and collected feedback to improve it.

**Methods:** As part of the larger METRIQ study, participants rated the quality of five blog posts on clinical emergency medicine topics using the eight-item METRIQ-8 score. Next, participants used a 7-point Likert scale and free-text comments to evaluate the METRIQ-8 score on ease of use, clarity of items, and likelihood of recommending it to others. Descriptive statistics were calculated and comments were thematically analyzed to guide the development of a revised METRIQ (rMETRIQ) score.

**Results:** A total of 309 emergency medicine attendings, residents, and medical students completed the survey. The majority of participants felt the METRIQ-8 score was easy to use (mean ± SD = 2.7 ± 1.1 out of 7, with 1 indicating strong agreement) and would recommend it to others (2.7 ± 1.3 out of 7, with 1 indicating strong agreement). The thematic analysis suggested clarifying ambiguous questions, shortening the 7-point scale, specifying scoring anchors for the questions, eliminating the "unsure" option, and grouping-related questions. This analysis guided changes that resulted in the rMETRIQ score.

**Conclusion:** Feedback on the METRIQ-8 score contributed to the development of the rMETRIQ score, which has improved clarity and usability. Further validity evidence on the rMETRIQ score is required.

With increasing expansion of emergency medicine (EM) blogs and podcasts, residents frequently use these open educational resources to supplement and potentially replace traditional tools.[1–4] Unlike textbooks and journals, these online resources are rarely peer-reviewed[5–7] and critics raise concerns that learners

From the [1]Department of Emergency Medicine, University of Alberta, Edmonton, Alberta, Canada; the [2]Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada; the [3]Division of Emergency Medicine, Department of Medicine, McMaster University, Hamilton, Ontario, Canada; the [4]Department of Emergency Medicine, Northwestern University, Chicago, IL; the [5]Department of Emergency Medicine, University of Minnesota, Minneapolis, MN; the [6]USC and LAC Departments of Emergency Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA; the [7]Department of Emergency Medicine, University of Calgary, Calgary, Alberta, Canada; and the [8]Department of Emergency Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.
Received January 26, 2019; revision received June 8, 2019; accepted June 17, 2019.
Presented at the Canadian Association of Emergency Physicians Annual Conference, Calgary, Alberta, Canada, May 2018.
Funding for this research was provided by the Canadian Association of Emergency Physicians (Junior Investigator Grant) and the Royal College of Physicians and Surgeons of Canada (Robert Maudsley Fellowship for Studies in Medical Education).
BT (ALiEM.com, Debrief2Learn.org, EMSimCases.com, and CanadiEM.org), FZ (ALiEM.com, and CanadiEM.org), NST (Mdaware.org), AG (ALiEM.com), and TMC (ALiEM.com, CanadiEM.org, FeminEM.org, ICENet.royalcollege.ca) edit or operate medical education blogs. NST receives salary support from the American Medical Association for his role as Digital Media Editor, *JAMA Network Open*, and a stipend for his role as Social Media Editor for *Emergency Physicians Monthly*; he previously received a stipend for his role as Social Media Editor for *Annals of Emergency Medicine* during portions of the METRIQ Study. The remaining authors declare no conflicts of interest.
Supervising Editor: Lalena M. Yarris, MD.
Address for correspondence and reprints: Brent Thoma, MD, MA, MSc; e-mail: brent.thoma@usask.ca.
AEM EDUCATION AND TRAINING 2019;3:387–392.


doi: 10.1002/aet2.10376

are being misled.[8–10] Supporting these concerns, the Medical Education Translational Resources: Impact and Quality (METRIQ) study found that gestalt evaluations of these resources were unreliable.[11–14] This suggests that a systematized appraisal of these resources may be more appropriate.[11,15]

The METRIQ-8 score is a structured rating tool resulted from a rigorous derivation process, which included a systematic review and qualitative analysis designed to identify appropriate quality indicators for blogs,[16] a modified Delphi process with expert bloggers and podcasters,[17] a modified Delphi process with medical educators,[18] and a derivation study.[19] However, along with another structured assessment tool (the ALiEM AIR score[19,20]), the METRIQ study found that METRIQ-8 was no more reliable than staff physician gestalt in a general population of raters.[12] As part of a planned secondary analysis of data collected within the METRIQ study, we analyzed feedback on the METRIQ-8 score with the goal of improving its usability and reliability.

## METHODS

This was a planned secondary analysis of data from the METRIQ study (http://metriqstudy.org), which recruited students, EM trainees, and EM attendings to rate the quality of 20 clinically oriented EM blog posts via an online survey between March 1, 2016, and June 1, 2016.[11,13,14] After rating five blog posts with the METRIQ-8 score (outlined in Data Supplement S1, Table S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10376/full), participants rated the METRIQ-8 score on usability and whether they would recommend it using 7-point Likert scales (1 = strongly agree). Participants also reviewed the eight METRIQ-8 items to identify unclear elements. Each question was followed by an open-ended question asking participants to explain their rationale. Only data from participants who completed the full METRIQ study survey were used. The METRIQ study protocol was reviewed by the University of Saskatchewan's Research Ethics Board and met the requirements for exemption (BEH 16-09).

Quantitative analysis was conducted using parametric descriptive statistics and tests of significance. Analysis of variance with a two-tailed significance of $\alpha = 0.05$ was used to determine whether the ease of use or likelihood of recommendation differed significantly based upon level of training, frequency of blog reading, or region of origin.

Anonymized narrative data were analyzed using a content analysis to identify areas for improvement and common points of feedback.[21] The researchers included six staff emergency physicians, one emergency medicine resident, and a senior medical student. Two authors had formal training in qualitative methods (TMC, BT). All authors were familiar with open-access medical education resources. Two raters (INCG, KK) independently coded the data, compared their analyses, and resolved discrepancies through consensus. The final codebook was organized into themes and subthemes with quotes from participants demonstrating each subtheme.

The revised METRIQ (rMETRIQ) score was developed through an iterative process. A subgroup of the authors (INCG, KK, BT) modified each item of the METRIQ-8 score and developed specific scoring criteria for each revised item. The remaining authors provided feedback, and consensus on each item was reached via group discussion. This version was then piloted by the five authors not involved in revising the score (TMC, NST, MP, AG, FZ) on a new set of blog posts. Consistent with methods used in previous work,[12] one new clinically relevant blog post was selected for review from each of the 10 websites used in the METRIQ study. Average intraclass correlation coefficients (ICCs) were calculated for each item and the total score (the sum of each item's scores). Minor additional edits were made to the final version to clarify items with a lower ICC (indicating lower reliability).

## RESULTS

Participant demographics are described in Table 1. A total of 309 of the 330 (93.6%) individuals enrolled in the METRIQ study completed the survey. As outlined in Figure 1, the majority of participants agreed that "the METRIQ-8 score was easy to use" (mean ± standard deviation [SD] = 2.7 ± 1.1 on a 7-point scale, with 1 indicating "strongly agree") and "would recommend the METRIQ-8 score for the evaluation of blog posts" (mean ± SD = 2.7 ± 1.3 on a 7-point scale, with 1 indicating "strongly agree"). Neither ease of use nor recommendation of the METRIQ-8 score varied significantly by level of training, frequency of blog reading, or global region of origin.

Table 1
The METRIQ Study Participant Demographics

| Variable | Category | *n* (*N* = 309) | % |
|---|---|---|---|
| Age | Years | 31.1 (mean) | 7.3 (SD) |
| Sex | Female | 123 | 39.8 |
| | Male | 184 | 59.5 |
| | Other | 2 | 0.6 |
| Level of training | Medical student | 121 | 39.2 |
| | EM resident | 88 | 28.5 |
| | Emergency attending physician | 100 | 32.4 |
| Frequency of reading medical education blogs | Daily | 48 | 15.5 |
| | Several times weekly | 141 | 45.6 |
| | Once weekly | 43 | 13.9 |
| | Several times monthly | 38 | 12.3 |
| | Once monthly | 21 | 6.8 |
| | Less than once monthly | 15 | 4.9 |
| | Never | 2 | 0.6 |
| Manage, edit, own, or operate a medical education blog(s) | Yes | 45 | 14.5 |
| | No | 261 | 84.5 |
| | No response | 3 | 1.0 |

Qualitative analysis of the comments and feedback on the METRIQ-8 score revealed nine main themes and 51 subthemes (detailed in Data Supplement S1, Table S2). Main themes included usability, interpretation, length, application, structure, validity and reliability, scale, completeness, and comparison to the ALiEM AIR score.

The feedback summarized in the thematic analysis (Figure 2) informed the creation of the rMETRIQ score from the METRIQ-8 score. Significant changes from the METRIQ-8 score are summarized in Data Supplement S1, Table S1, and included clarifying ambiguous questions, shortening the 7-point scale to a 4-point scale, specifying scoring anchors for each question, eliminating the "unsure" option, and grouping-related questions. The choice of a 4-point scale was consistent with multiple participant recommendations to reduce the number of options. The scoring criteria and question refinement clarified terminology previously identified as ambiguous. Finally, we changed the order of the questions to group them into three broad categories: content, credibility, and review. The qualitative analysis of feedback provided on each item is outlined in Data Supplement S1, Table S3.

The results of pilot testing are shown in Table 2. Reliability of the aggregate score was high (ICC = 0.94, 95% confidence interval [CI] = 0.84–0.98). ICCs for individual items were also high (≥0.80) with the exception of rQ3 ("Is the resource well written and formatted?"; 0.72) and rQ5 ("Is it clear who created the resource and do they have any conflicts of interest?"; 0.59). Further changes to these items suggested by the pilot testers were made as outlined in Data Supplement S1, Table S2. The final rMETRIQ Score is presented in Figure 2.

## DISCUSSION

The rMETRIQ score was developed from the METRIQ-8 score by leveraging quantitative and qualitative feedback provided by a large population of users at various stages in training, geographic location, and levels of involvement with online medical education. This diverse group of participants mirrors the range of

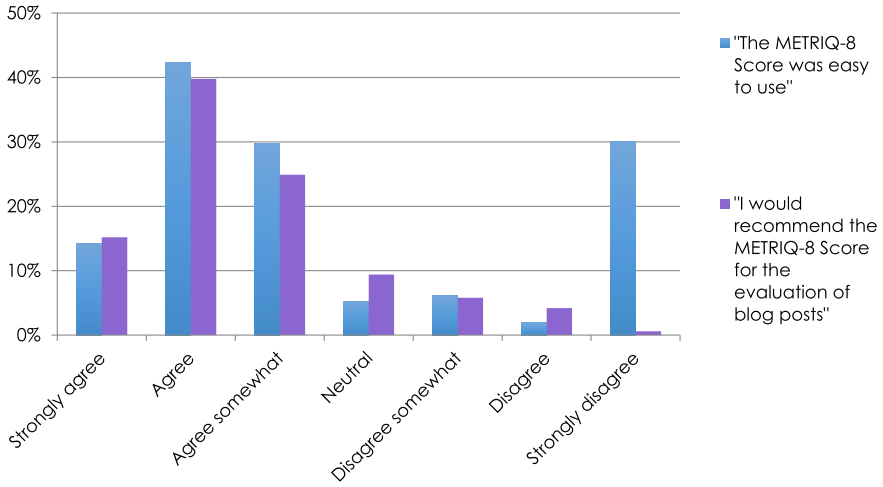


**Figure 1.** The METRIQ study participants' perspectives on the ease of use and recommendation of the METRIQ-8 score.

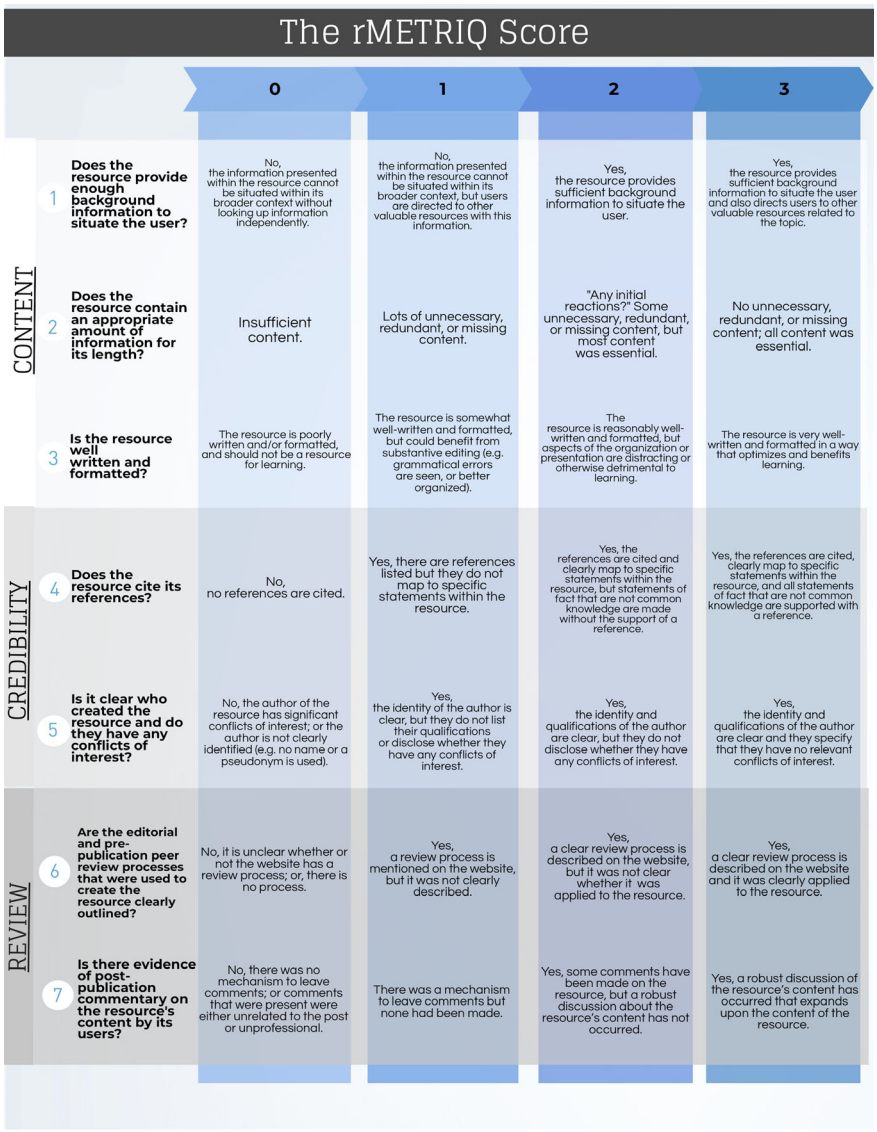


**Figure 2.** The revised METRIQ (rMETRIQ) score.

Table 2
ICCs for the Items of the Revised METRIQ (rMETRIQ) Score

| rMETRIQ Score Item | Average-measures ICC (95% CI) |
|---|---|
| Aggregate score (sum of rQ1–rQ7) | 0.94 (0.84–0.98) |
| Revised Question 1 (rQ1). Does the resource provide enough background information to situate the user? | 0.89 (0.70–0.97) |
| Revised Question 2 (rQ2). Does the resource contain an appropriate amount of information for its length? | 0.80 (0.51–0.94) |
| Revised Question 3 (rQ3). Is the resource well written and formatted? | 0.72 (0.30–0.92) |
| Revised Question 4 (rQ4). Does the resource cite its references? | 0.96 (0.90–0.99) |
| Revised Question 5 (rQ5). Is it clear who created the resource and do they have any conflicts of interest? | 0.59 (0–0.88) |
| Revised Question 6 (rQ6). Are the editorial and prepublication peer review processes that were used to create the resource clearly outlined? | 0.82 (0.56–0.95) |
| Revised Question 7 (rQ7). Is there evidence of postpublication commentary on the resource's content by its users? | 0.95 (0.87–0.99) |

ICC = intraclass correlation coefficient.

typical end-users that our instrument targets and identified correctable aspects for improvement.

The development of the rMETRIQ score is important given recent studies demonstrating that both gestalt[11,15] and the current structured evaluation tools (METRIQ-8 and ALiEM AIR)[12] lack reliability in general populations of raters. Reliability is an important component of modern validity theory and is generally felt to be a necessary (but not sufficient) aspect of validity.[22] Reliability improves with the number of raters, but the gestalt rating of blog posts requires a prohibitive number to achieve adequate reliability.[11] The reliability of instruments (i.e., higher value in an average-measures ICC) can be increased through rater training and instrument improvement.[23] In light of the disappointing results of reliability testing, we felt that revising our instrument would be the next reasonable approach to improving evaluation of these resources. Pilot testing of the rMETRIQ score suggests that its reliability has improved. However, it will require further evaluation in a larger validation study with general readers of emergency medicine blogs. We anticipate that the rMETRIQ score will impact three separate areas within EM: first, by guiding quality assessment of online resource among readers; second, by improving quality of online content by providing a framework of quality metrics for content producers to incorporate into future online content; and finally, by supporting the development of more robust methods of reviewing and assessing the online emergency medicine.

The rMETRIQ score was recently used to appraise the quality of blog posts in the new SAEM Systematic Online Academic Resource (SOAR) review series of online educational content on EM renal and genitourinary conditions.[24] Although our work and the METRIQ study are centered around EM content, the quality principles of the rMETRIQ score can easily be applied to other domains within medicine. Additionally, we anticipate that with minor modification of the wording of the instrument, it will be possible to apply it to other types of online resources such as podcasts, videos, and other open educational resources that are not vetted through traditional peer review processes. Further studies will also be required to classify the numeric scores (i.e., what score constitutes high vs. medium vs. low quality).

## LIMITATIONS

First, the data used in this study was collected in 2016 and it is possible that the feedback received on the METRIQ-8 score may have differed with a sample of blog posts published more recently. Second, given the significant modifications made to develop the rMETRIQ score, new validity evidence will need to be collected before its use can be recommended broadly. Finally, the rMETRIQ score was developed specifically using blogs and will need to be modified for application to other popular online educational resources.

## CONCLUSIONS

Direct feedback on the METRIQ-8 score spurred the development of the revised METRIQ score with improved usability and reliability. We anticipate that it will be used widely to assess the quality of blog posts and, potentially, other online resources. Further validity evidence for use of the revised METRIQ score will be required before it can be broadly recommended.

The authors thank the medical students, residents, and staff physicians who participated in the METRIQ study.

## References

1. von Muhlen M, Ohno-Machado L. Reviewing social media use by clinicians. J Am Med Inform Assoc 2012;19:777–81.
2. Matava CT, Rosen D, Siu E, et al. eLearning among Canadian anesthesia residents: a survey of podcast use and content needs. BMC Med Educ 2013;13:59.
3. Purdy E, Thoma B, Bednarczyk J, et al. The use of free online educational resources by Canadian emergency medicine residents and program directors. Can J Med Educ 2015;17:101–6.
4. Mallin M, Schlein S, Doctor S, et al. A survey of the current utilization of asynchronous education among emergency medicine residents in the United States. Acad Med 2014;89:598–601.
5. Azim A, Beck-Esmay J, Chan TM. Editorial processes in free open access medical educational (FOAM) resources. AEM Educ Train 2018;2:204–12.
6. Thoma B, Chan T, Desouza N, et al. Implementing peer review at an emergency medicine blog: bridging the gap between educators and clinical experts. Can J Emerg Med 2015;17:188–91.
7. Sidalak D, Purdy E, Luckett-Gatopoulos S, et al. Coached peer review: developing the next generation of authors. Acad Med 2017;92:201–4.
8. Genes N. Pro/con: why #FOAMed Is Not Essential to EM Education[Internet]. Emergency Physicians Monthly. 2014. Available at: http://epmonthly.com/article/pro-con-is-foam-essential-to-em-education-no/. Accessed June 6, 2019.

9. Brabazon T. The Google effect: googling, blogging, wikis and the flattening of expertise. Libri 2007;56:157–67.
10. Schriger DL. Does everything need to be "scientific?". Ann Emerg Med 2016;68:738–9.
11. Thoma B, Sebok-Syer SS, Krishnan K, et al. Individual gestalt is unreliable for the evaluation of quality in medical education blogs: a METRIQ study. Ann Emerg Med 2017;70:394–401.
12. Thoma B, Sebok-Syer SS, Colmers-Gray I, et al. Quality Evaluation Scores are no more reliable than gestalt in evaluating the quality of emergency medicine blogs: a METRIQ study. Teach Learn Med 2018;30:294–302.
13. Thoma B, Paddock M, Purdy E, et al. Leveraging a virtual community of practice to participate in a survey-based study: a description of the METRIQ study methodology. AEM Educ Train 2017;1:110–3.
14. Thoma B, Chan TM, Kapur P, et al. The social media index as an indicator of quality for emergency medicine blogs: a METRIQ study. Ann Emerg Med 2018;72:696–702.
15. Krishnan K, Thoma B, Trueger NS, et al. Gestalt assessment of online educational resources may not be sufficiently reliable and consistent. Perspect Med Educ 2017;6:91–8.
16. Paterson QS, Thoma B, Milne WK, et al. A systematic review and qualitative analysis to determine quality indicators for health professions education blogs and podcasts. J Grad Med Educ 2015;7:549–54.
17. Thoma B, Chan TM, Paterson QS, et al. Emergency medicine and critical care blogs and podcasts: establishing an international consensus on quality. Ann Emerg Med 2015;66(396–402):e4.
18. Lin M, Thoma B, Trueger NS, et al. Quality indicators for blogs and podcasts used in medical education: modified Delphi consensus recommendations by an international cohort of health professions educators. Postgrad Med J 2015;91:546–50.
19. Chan TM, Thoma B, Krishnan K, et al. Derivation of two critical appraisal scores for trainees to evaluate online educational resources: a METRIQ study. West J Emerg Med 2016;17:574–84.
20. Chan TM, Grock A, Paddock M, et al. Examining reliability and validity of an online score (ALiEM AIR) for rating free open access medical education resources. Ann Emerg Med 2016;68:729–35.
21. Cooper S, Endacott R. Generic qualitative research: a design for qualitative research in emergency care? Emerg Med J 2007;24:816–9.
22. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006;166.e7–16.
23. Norman GR, Streiner DL. Measurement Scales: A Practice Guide to Their Development and Use, 3rd ed. Oxford: Oxford University Press, 2008.
24. Grock A, Bhalerao A, Chan TM, et al. Systematic Online Academic Resource (SOAR) review: renal and genitourinary. AEM Educ Train 2019;19:000–00.

## Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10376/full

**Data Supplement S1.** Supplemental material