Effect of a Scoring Rubric on the Review of Scientific Meeting Abstracts

KEY WORDS: meeting abstract; peer review; reproducibility of results; health services research.

J Gen Intern Med 36(8):2483–5
DOI: 10.1007/s11606-020-05960-6

INTRODUCTION

Scientific meeting abstract review is susceptible to poor inter-rater agreement, which can lead to decreased differentiation among abstracts. A rubric is "a scoring guide… with three essential features: evaluative criteria, quality definitions, and a scoring strategy."[1] Abstract review guided by a detailed rubric could improve inter-rater reliability and lead to presentation of higher quality abstracts.

The 1991 Society of General Internal Medicine (SGIM) scientific abstract committee analyzed inter-rater agreement.[2] At that time, there were three criteria: interest to SGIM audience, quality of methods, and quality of presentation. Score options were as follows: 1= poor, 2 = fair, 3 = good, 4 = very good, and 5 = outstanding. Given significant reviewer disagreement, the authors suggested a 7-point scoring scale with explicit descriptions of the scores.

By 2016, there were four criteria, with sparse instructions ("1, lowest; 7, highest"). In 2017, a large-scale rubric modification was initiated, retaining four review criteria (Importance, Methods, Conclusions, and Writing), but adding detailed descriptions for each score on the 7-point scale within each criterion (see Text Box 1). We examined whether the 2017 rubric addressed scoring issues including leniency bias (abstract mean scores), inter-rater reliability (within-abstract standard deviations), and discriminability of abstracts (across-abstract standard deviations).

METHODS

We analyzed all abstracts submitted from 2014 to 2018, with 2014–2016 designated as "old" and 2017–2018 as "new" rubric periods. We calculated the *composite* score for each abstract-reviewer combination as the mean of the four individual *criteria* scores (Importance, Methods, Conclusions, and Writing) provided by a reviewer for a given abstract. We calculated the *final* score for each abstract as the unweighted mean of the composite scores from all submitted reviews for that abstract.

All analyses compared "old" to "new" rubric abstracts. First, we calculated the mean composite score per abstract (i.e., final score) and the standard deviations (SDs) of the composite scores for a given abstract. These are within-abstract statistics, reflecting the distribution of composite scores across reviews within each abstract. For each within-abstract statistic, we took a weighted mean of the statistic in the old and new rubric periods, using the number of reviews as the weighting factor. Then, we calculated the old to new ratio of the weighted mean of the statistic. To test the hypotheses that the new rubric would (1) decrease scores (i.e., reduce leniency), (2) increase inter-rater reliability, and (3) cause reviewers to use more of the scoring range across abstracts, we calculated the old to new ratio of (1) weighted mean final scores, (2) weighted mean of within-abstract SDs for composite scores, and (3) across-abstract SDs for final scores, respectively.

We used approximate permutation to estimate the sampling distribution of old to new ratios under the null hypothesis that the rubric had no effect.[3] We used sampling with replacement by drawing 1000 samples of 3523 abstracts from the original sample of 3523 abstracts, randomly allocating 2078 as "old" and 1445 as "new" rubric, based on the original ratio of abstracts. We calculated the old to new ratio for each statistic of interest. If the observed old to new ratio falls outside the range of ratios calculated from the 1000 random samples, the null hypothesis can be rejected.

Prior Presentations *Selected findings from this paper have been featured in an oral presentation at the Society of General Internal Medicine annual meeting (Washington DC, May 2019).*

Received April 22, 2020
Accepted June 4, 2020
Published online August 3, 2020

Text Box 1 Scientific abstract review instructions for 2017–2018 ("new" rubric period)

Importance of the Research Question [Importance]: To what extent does the abstract address a topic that is important? To what degree will the results advance concepts in General Internal Medicine?

1	2	3	4	5	6	7
Does not address a topic important to general internists.	Addresses a topic important to only a *few* general internists.	Addresses a topic important to *some* general internists.	Addresses a topic important to *about half* of general internists.	Addresses a topic that is important to *many* general internists; or *somewhat* expands current concepts.	Addresses a topic that is important to *most* general internists; or *greatly* expands current concepts.	Addresses a topic that is important to *nearly all* general internists; or *introduces* a *new* concept.

Strength and Appropriateness of Methods [Methods]: Is the study design clearly described? Are sampling procedures adequately described, including inclusion and exclusion criteria; is there potential selection bias? Are the measures reliable and valid? Are possible confounding factors addressed? Are the statistical analyses appropriate for the study design, and are they the best that could have been used? Is there discussion of the statistical power?
[*Please note that not all issues described apply to all abstract types. For example, qualitative studies may not have statistical analyses; however, they should still be evaluated on the quality of study design description and appropriateness of the methods.*]

1	2	3	4	5	6	7
Study design and sampling procedures not described. Possible confounders not discussed. Statistical analyses are not discussed.	Study design and sampling procedures *poorly* described. Possible confounders not discussed.	Study design and sampling procedures *adequately* described. Possible confounders *not* discussed. Statistical analyses are *adequate.*	Study design and sampling procedures *fully* described. Measures are *probably* reliable and valid. Possible confounders *partially* discussed, but may not be controlled. Statistical analyses are *appropriate.*	Study design and sampling procedures *fully* described. No selection bias exists. Measures *probably* reliable and valid. Possible confounders *fully* discussed and controlled for as needed. Statistical analyses are *appropriate*.	Study design and sampling procedures *well* described. No selection bias exists. Measures *are* reliable and valid. Possible confounders *fully* discussed and controlled for as needed. Statistical analyses are *strong.*	Study design and sampling procedures *very clearly* described. No selection bias exists. Measures *are* reliable and valid. Possible confounders *fully* discussed and controlled for as needed. Statistical analyses are *the best that could have been used.*

Validity of Conclusions and Implications [Conclusions]: Are conclusions clearly stated and justified by the data? Are implications strong enough to influence how clinicians/teachers/researchers "act" in clinical practice, teaching, or future research?

1	2	3	4	5	6	7
Conclusions and implications *not* included. Does *not* influence action.	Conclusions *present* but not justified. Does *not* influence action.	Conclusions *present* and *weakly* supported. Provides knowledge but *likely will not* change action.	Conclusions *clearly* stated and *supported*. *Absent or weak* implications. Provides knowledge but *likely will not* change action.	Conclusions *clearly* stated and supported. Implications *weak*. Provides knowledge that *may* change action.	Conclusions *clearly* stated and supported. Implications *moderately* appropriate. Provides knowledge that *may* change action.	Conclusions *clearly* stated and supported. Implications *fully* appropriate. Provides knowledge that *likely will* change action.

Quality of Writing [Writing]: Is the writing clear and organized to effectively communicate the findings?

1	2	3	4	5	6	7
Writing is *poor* and *disorganized.*	Writing is *adequate* and *somewhat* disorganized.	Writing is *adequate* and *minimally* disorganized.	Writing is *clear* and *organized*.	Writing is *above average* and *organized*.	Writing is *high quality* and *well organized*.	Writing is *masterful* and *well organized*.

Table 1 Effect of Rubric on Composite Scores

Year	Abstracts (*n*)	Reviews (*n*)	Weighted mean final score	Weighted mean composite score SD (within-abstract)	Final score SD (across abstracts)
Old rubric	2078	13,895	4.961	0.943	0.646
New rubric	1445	8347	4.764	0.904	0.647
Ratio old/new	n/a	n/a	*1.041*	*1.042*	0.998
Permutation range	n/a	n/a	0.989–1.015	0.955–1.027	0.915–1.069

Italic values indicate statistical significance

RESULTS

During the study period, 3523 abstracts were submitted, 2078 in the old period and 1445 in the new period. The effect of the 2017 rubric on composite scores is shown in Table 1. The weighted mean final scores in new rubric years were significantly lower than those in old rubric years. Weighted mean within-abstract SDs of composite scores similarly show statistically significant decreases in new rubric years. Final score SDs *across* abstracts indicated no statistically significant change.

DISCUSSION

Our new rubric successfully lowered final scores on scientific abstracts, reflecting a shift away from leniency bias (i.e., tendency toward the upper portion of a scoring range). The rubric also decreased the composite score SDs *within* abstracts, indicating improvement in inter-rater agreement. The rubric did not lead to more variable scores overall across all abstracts; however, scores did shift toward the lower end of the scoring range, such that fewer abstracts received high scores and more received low scores.

Objective evaluation of abstract submissions ensures the rigor of scientific meeting presentations. Efforts should continue to refine and implement tools to improve abstract scoring and maintain a high-integrity environment for disseminating scientific discovery.

Nia S. Mitchell, MD, MPH[1,2]
Kelly Stolzmann, MS[3]
Lauren V. Benning, DO[4]
Jolie B. Wormwood, PhD[5,6]
Amy M. Linsky, MD, MSc[3,7,8]

[1]Division of General Internal Medicine, Duke University School of Medicine,
200 Morris St., 3rd Floor, Durham, NC, USA
[2]Center for Community and Population Health Improvement, Duke University School of Medicine,
Durham, NC, USA
[3]Center for Healthcare Organization and Implementation Research, VA Boston Healthcare System,
Boston, MA, USA
[4]Family Medicine Residency, McLeod Regional Medical Center,
Florence, SC, USA
[5]Department of Psychology, University of New Hampshire,
Durham, NH, USA
[6]Center for Healthcare Organization and Implementation Research, Edith Nourse Rogers Memorial VA Hospital,
Bedford, MA, USA
[7]Section of General Internal Medicine, VA Boston Healthcare System,
Boston, MA, USA
[8]Section of General Internal Medicine, Boston Medical Center,
Boston, MA, USA

Corresponding Author: *Nia S. Mitchell, MD, MPH; Division of General Internal Medicine, Duke University School of Medicine, 200 Morris St., 3rd Floor, Durham, NC, USA (e-mail: nia.s.mitchell@duke.edu).*

Funding Information *Dr. Mitchell was supported by an NIH/NHLBI career development award (K01HL115599). Dr. Linsky was supported by a Department of Veterans Affairs (VA), Veterans Health Administration, Health Services Research and Development Career Development Award (CDA12-166).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Disclaimer: *The views expressed in this article are those of the authors and do not necessarily represent the views of the NIH nor the Department of Veterans Affairs. Neither the NIH nor the Department of Veterans Affairs had a role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; nor the decision to submit the manuscript for publication.*

REFERENCES

1. **Popham WJ**. What's wrong - and what's right - with rubrics. *Educ Leadership.* 1997;55(2):72-75.
2. **Rubin H**, **Redelmeier D**, **Wu A**, **Steinberg E**. How Reliable Is Peer Review of Scientific Abstracts? Looking Back at the 1991 Annual Meeting of the Society of General Internal Medicine. *J Gen Intern Med.* 1993;8:255-258.
3. **Ludbrook J.** Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin Exp Pharmacol Physiol.* 1994;21(9):673-686.