ORIGINAL RESEARCH

Reliability of Peer Review of Abstracts Submitted to Academic Family Medicine Meetings

Joshua J. Fenton, MD, MPH, Hazel Tapp, PhD, Netra M. Thakur, MD, MPH, FAAFP, and Andrea L. Pfeifle, EdD, PT, FNAP

Purpose: **To assess the reliability of peer review of abstracts submitted to academic family medicine meetings in North America.**

Methods: **We analyzed reviewer ratings of abstracts submitted: 1) as oral presentations to the North American Primary Care Research Group (NAPCRG) meeting from 2016 to 2019, as well as 2019 poster session or workshop submissions; and 2) in 12 categories to the Society of Teachers of Family Medicine (STFM) Spring 2018 meeting. In each category and year, we used a multi-level mixed model to estimate the abstract-level intraclass correlation coefficient (ICC) and the reliability of initial review (using the abstract-level ICC and the number of reviewers per abstract).**

Results: **We analyzed review data for 1554 NAPCRG oral presentation abstracts, 418 NAPCRG poster or workshop abstracts, and 1145 STFM abstracts. Across all years, abstract-level ICCs for NAPCRG oral presentations were below 0.20 (range, 0.10 in 2019 to 0.18 in 2016) and were even lower for posters and workshops (range, 0.00-0.10). After accounting for the number of reviewers per abstract, reliabilities of initial review for NAPCRG oral presentations ranged from 0.24 in 2019 to 0.30 in 2016 and 0.00 to 0.18 for posters and workshops in 2019. Across 12 STFM submission categories, the median abstract-level ICC was 0.21 (range, 0.12-0.50) and the median reliability was 0.42 (range, 0.25-0.78).**

Conclusions: **For abstracts submitted to North American academic family medicine meetings, inter-reviewer agreement is often low, compromising initial review reliability. For many submission categories, program committees should supplement initial review with independent postreview assessments. (J Am Board Fam Med 2020;33:986–991.)**

Keywords: **Abstracting and Indexing, Biostatistics, Faculty, Observer Variation, Peer Review, Primary Health Care**

Introduction

Many scientific societies sponsor annual conferences where members congregate to present and discuss novel research findings and methods. For attendees, conferences are an opportunity to elicit feedback, encounter new ideas, and spawn collaborations. Oral conference presentations may be considered positively by academic promotion committees and may justify financial assistance to authors to defray meeting costs. Some societies issue awards for outstanding abstracts or articles.

To select abstracts for presentation or award, scientific societies rely on peer review of submitted abstracts. By definition, a valid and fair peer review process would have high reliability, where reliability refers to the reproducibility of the review outcome. Hence, if the same abstract were independently reassessed by a highly reliable peer review system, the result would be similar with each assessment. However, the reliability of peer review of confence abstracts has received limited study and findings have been variable when evaluated.[1–5] Reliability of peer review has been found to be low in

This article was externally peer reviewed.
Submitted 7 April 2020; revised 23 June 2020; accepted 10 July 2020.

From the Department of Family and Community Medicine and Center for Healthcare Policy and Research, University of California, Davis (JJF); Center for Primary Care Research, Department of Family Medicine, Atrium Health, Charlotte, NC (HT); University of North Carolina, School of Medicine, Advance Community Health Center, Raleigh (NMT); School of Medicine, Indiana University, Indianapolis (ALP).

Funding: None.
Conflicts of interest: JJF and HT are currently Vice-Chair and Chair, respectively, of the NAPCRG Program Committee. NMT and ALP are the current and former Chairs of the STFM Program Committee.

Corresponding author: Joshua J. Fenton, MD, MPH, 4860 Y St, Suite 2300, UC Davis Medical Center, Sacramento, CA 95817 (E-mail: jjfenton@ucdavis.edu).

analyses of biomedical journal submissions[6,7] and grant applications.[8–11]

Our principal objective was to assess the reliability of abstract review at recent annual conferences for 2 organizations representing North American academic family physicians: the North American Primary Care Research Group (NAPCRG) and the Society of Teachers of Family Medicine (STFM). While NAPCRG meetings focus on research, the STFM meeting prioritizes scholarship related to family medicine education, including research presentations. Both meetings include poster sessions and workshops. In 2018, NAPCRG Program Committee members conducted interim analyses of NAPCRG data from 2016 to 2018, which suggested suboptimal reliability of initial peer review of abstracts. In response, the Committee increased the number of reviewers per abstract in 2019 by assigning every abstract to at least 1 Committee member in addition to 2 other reviewers. We analyzed data from 2019 to assess for improvement.

Methods

Our study had a serial cross-sectional design based on analyses of deidentified peer review data from annual NAPCRG meetings from 2016 to 2019 and the Spring 2018 STFM meeting. For NAPCRG meetings, samples included abstracts submitted as oral presentations on completed research. For NAPCRG 2019, we also obtained abstract data on poster sessions (on "completed research" and "research in progress") and workshops. For the STFM meeting, samples included data for abstracts submitted in all 12 categories (see Table 1 for category names).

Depending on meeting and category, abstracts had variable numbers of reviews. Program Committees for each organization use the initial review ratings in postreview assessment, leading to final adjudication. Our analytic goal was to determine the reliability of the initial peer review process within each submission category for each of the 5 meetings. Reliability here refers to the reproducibility of the initial review, or the probability that the review result would be the same if it were repeated by a new set of randomly selected reviewers. Reliability increases as a function of 2 parameters: the correlation of individual reviewer assessments of the same abstract, and the total number of reviewers assigned to each abstract. If correlation is low across reviewers, a greater number of reviews will be needed to achieve adequate reliability.

For each abstract, we obtained reviewer ratings across all review subscales as well as global ratings. Reviewers were identified by anonymous study identification numbers, except for NAPCRG 2019 when reviewer identification was not possible due to a change in data systems. Because the study data were deidentified, the research was deemed exempt from human subject review by the University of California–Davis Institutional Review Board.

For NAPCRG 2016 and 2017 and most STFM 2018 categories, reviewers rated abstracts on 5 items, each item using a 5-point Likert scale. In 2016, the 5 NAPCRG items pertained to the importance of the topic, clarity of abstract's aims, the trustworthiness of the results, the importance of the findings, and the likely interest and value to NAPCRG attendees; in 2017, the NAPCRG items pertained to the abstract's relevance to primary care, the description of research methods, the validity of results, the clarity of writing and organization, and the newsworthiness of findings. For most STFM categories in 2018, the 5 items pertained to whether the abstract was clearly written, the clarity of the objectives, the relevance of the content to family medicine educators, whether the presentation was likely to be engaging and actionable for participants, and the overall quality. For 3 STFM categories (research projects and posters, and works-in-progress posters), reviewers rated abstracts' overall quality on a single 5-item Likert scale. For meetings and categories with multiple ratings per abstract, within-abstract item responses had high average interitem correlations (range: 0.54-0.69). Because of the high interitem correlations and to simplify analyses, we created standardized scales using all available items to reflect reviewers' overall judgments (range in Cronbach's α, 0.85-0.92). During NAPCRG 2018, reviewers rated abstracts on a single 1 to 10 scale, which we standardized for analyses. For NAPCRG 2019, reviewers rated abstracts on 5 items (clarity, relevance, methodological rigor, impact, and interest and value) before responding on a 5-point Likert scale to the item: "By definition, most NAPCRG abstracts should be rated 'average.' Based on your ratings above, rate the abstract overall compared with most NAPCRG abstracts submitted in this category." For 2019, we analyzed the standardized response to this summary question.

For statistical analyses, we used Stata MP (Version 15.1, College Station, TX). Using the standardized reviewer ratings, we quantified the correlation between reviewer assessments of the same

Table 1. Reliabilities of Abstract Reviews by Family Medicine Meeting, Year, and Submission Category

Submission by Year, Meeting, Category	N, Submissions	Number of Reviews, Mean	Intracluster Correlation Coefficients (95% CI)		Reliability of Within-Abstract Mean Score (95% CI)
			Abstract-Level	Reviewer-Level	
North American Primary Care Research Group					
2016 oral presentation on completed research	392	2.0	0.18 (0.09-0.26)	0.22 (0.15-0.30)	0.30 (0.17-0.41)
2017 oral presentation on completed research	400	2.0	0.15 (0.08-0.24)	0.25 (0.17-0.33)	0.26 (0.15-0.39)
2018 oral presentation on completed research	388	2.0	0.17 (0.08-0.26)	0.15 (0.08-0.26)	0.29 (0.15-0.41)
2019 oral presentation on completed research	374	3.0	0.10 (0.03-0.16)	—*	0.24 (0.08-0.36)
2019 poster on completed research	139	2.0	0.03 (0.00-0.21)	—*	0.07 (0.00-0.34)
2019 poster on research in progress	238	2.0	0.11 (0.00-0.24)	—*	0.18 (0.00-0.38)
2019 workshop	41	2.9	0.00 (0.00-0.18)	—*	0.00 (0.00-0.39)
Society of Teachers of Family Medicine, 2018 Spring Meeting					
Completed research poster	30	3.0	0.33 (0.11-0.55)	0.11 (0.00-0.27)	0.60 (0.27-0.79)
Completed research project	75	3.0	0.50 (0.37-0.63)	0.08 (0.00-0.16)	0.75 (0.64-0.84)
Completed scholarly project	37	2.9	0.21 (0.02-0.41)	0.34 (0.00-0.72)	0.44 (0.06-0.67)
Completed scholarly project poster	49	2.5	0.18 (0.00-0.40)	0.15 (0.00-0.35)	0.35 (0.00-0.63)
Developing scholarly project poster	112	2.3	0.13 (0.02-0.24)	0.37 (0.16-0.58)	0.25 (0.04-0.42)
Trainee work-in-progress poster	244	2.0	0.27 (0.16-0.38)	0.22 (0.04-0.40)	0.42 (0.28-0.55)
Lecture discussion	238	2.9	0.17 (0.10-0.25)	0.22 (0.06-0.38)	0.37 (0.24-0.49)
Panel discussion	20	3.0	0.20 (0.00-0.44)	0.19 (0.00-0.48)	0.42 (0.00-0.70)
Pre-conference workshop	14	7.0	0.34 (0.12-0.56)	0.16 (0.00-0.35)	0.78 (0.49-0.90)
Scholarly topic roundtable discussion	89	2.9	0.12 (0.01-0.24)	0.16 (0.01-0.31)	0.28 (0.03-0.48)
Seminar	180	3.0	0.16 (0.07-0.24)	0.15 (0.00-0.29)	0.36 (0.18-0.49)
Workshop	57	3.0	0.33 (0.17-0.49)	0.10 (0.00-0.26)	0.60 (0.38-0.74)

*Reviewer identifiers were not available for the 2019 North American Primary Care Research Group meeting, preventing estimation of reviewer-level ICCs.
CI, confidence interval; ICCs, intracluster correlation coefficients.

abstract by fitting multi-level mixed linear regression models with the standardized summary scores as dependent variables and abstract- and reviewer-level random effects to estimate abstract- and reviewer-level variance components. The mixed-effects models accounted for cross-nesting of abstracts within reviewers and vice versa. For each model, we computed the abstract- and reviewer-level intracluster correlation coefficients (ICCs) by dividing the between-abstract and between-reviewer variance components, respectively, by the sum of all variance components (between-abstract, between-reviewer, and residual error).[12–14] The abstract-level ICC quantifies inter-reviewer agreement on ratings of the same abstract, while the reviewer-level ICC quantifies the agreement in ratings by the same reviewer across abstracts. For NAPCRG 2016 to 2018, models adjusted for reviewer characteristics (self-rated research experience and principal investigator status), although these adjustments had no substantive impact on abstract- or reviewer-level ICCs during these years. For both NAPCRG 2019 and STFM 2018, we lacked data on reviewer characteristics, so we estimated ICCs using empty models; for NAPCRG 2019, the model only

included an abstract-level random effect due to lack of data identifying reviewers.

Based on the abstract-level ICCs and the average number of reviewers for each abstract within submission categories, we used the Spearman-Brown prophecy formula to estimate reliability (ranging from 0 to 1) of the initial review process.[15] For high-stakes assessments (eg, comparing performance of individual physicians based on quality metrics), experts have argued that measures should have reliabilities of 0.80 or higher.[16,17] For abstract reviews, lower reliabilities may be acceptable, though reliabilities ≤0.20 imply largely idiosyncratic processes.

Results

As shown in Table 1, we analyzed review data for 1544 abstracts submitted to the NAPCRG category of oral presentation on completed research from 2016 to 2019 (range per year, 374 to 400). We also assessed 2019 NAPCRG poster and workshop categories. NAPCRG abstracts were reviewed by a mean of 2.0 reviewers (range, 1 to 3), except in 2019 when oral presentation and workshop abstracts were reviewed by approximately 3 reviewers (range, 1 to 3). For the 2018 STFM Spring Meeting, we analyzed data for 1145 abstracts in 12 submission categories (range, 14 for preconference workshops to 244 trainee work-in-progress posters). The modal number of reviewers per STFM abstract was 3.0 (range, 1 to 7) although preconference workshop abstracts were reviewed by a mean of 7.0 reviewers.

Across all years, abstract-level ICCs for NAPCRG oral presentations were below 0.20 (range, 0.10 in 2019 to 0.18 in 2016). Abstract-level ICCs in 2019 were very low for posters on completed research (0.03), posters on research in progress (0.10), and zero for workshops. After accounting for the number of reviewers per abstract, reliabilities of initial review were limited for oral presentations (range, 0.24 in 2019 to 0.30 in 2016) and poor for posters and workshops in 2019 (range, 0.00 to 0.18). For oral presentations from 2016 to 2018, the abstract-level ICC was comparable to the reviewer-level ICC, implying that reviewer identity had as much influence on ratings as abstract content.

Across the 12 STFM submission categories, the median abstract-level ICC was 0.21 (range, 0.12 for scholarly topic roundtable discussions to 0.50 for completed research projects), while the median reliability was 0.42 (range, 0.25 for developing scholarly project poster to 0.78 for preconference workshops). For several STFM categories, reviewer-level ICCs were similar to or greater than the abstract-level ICC, although abstract-level ICCs were clearly higher than reviewer-level ICC for completed research and workshop abstracts, resulting in more favorable review reliabilities for these categories (range, 0.60 for both completed research posters and workshops to 0.78 for preconference workshops).

Discussion

Our analyses of abstract review data from recent academic family medicine meetings reveal typically low correlation between ratings by separate reviewers of the same abstract, frequently leading to low reliability of the initial review process. Reliability was low in all NAPCRG years and submission categories and in selected STFM categories in 2018. Reliability was boosted to acceptable levels in other STFM categories due to greater inter-reviewer agreement and the use of 3 or more reviewers.

The low reliabilities in this study are consistent with other studies of medical society meetings,[1,2,5] as well as analyses of inter-reviewer agreement of submitted journal articles and National Institutes of Health grant applications, wherein substantial influence individual reviewer preferences was also observed.[7,9] On the other hand, reliability of abstract review has been more favorable in 2 academic medical meetings.[3,4] Each of these meetings emphasized research, used longer, more detailed ratings scales, and had a smaller pool of reviewers than NAPCRG or STFM. Family medicine organizations may seek to lengthen or improve rating scales or consider brief reviewer orientation or training to improve the accuracy and reproducibility of reviewer ratings.[18]

Because of the breadth of primary care practice, NAPCRG and STFM attendees and reviewers are likely to have diverse interests and priorities and hence to differ in rating abstracts based on dimensions such as importance, interest, or impact. Indeed, as reviewer preferences are likely to remain influential, higher reliability of initial review may require a large number of reviewers.[19] Using the Spearman-Brown prophecy formula, we estimate that 8 reviewers would be required to increase NAPCRG oral presentation review reliability above 0.6. While it may be infeasible to obtain such a large number of reviews per abstract, committees should attempt to augment the number

of independent reviews per abstract in submission categories with currently low reliabilities.

Although NAPCRG increased the number of reviews per abstract by assigning 2019 oral presentation abstracts to at least 1 Program Committee member, our analyses do not suggest resultant improvement. It is possible that Program Committee members rate abstracts differently than other reviewers, which could have compromised between-reviewer correlations. Aware of low reliabilities in 2018 and 2019, the Committee undertook extensive reassessment of abstracts ranked highly or lowly after initial review to guide selection of abstracts as distinguished papers or rejection of a small percentage of abstracts. In 2019, this reassessment consisted of all Program Committee members reading and rerating all abstracts in the top and bottom quartiles after initial review. The STFM Program Committee also conducts extensive reassessment of abstracts after initial peer review during which reviewers calibrate ratings after discussion with coreviewers. The STFM Committee also considers the breadth and balance in its representation of topics within the meeting. Our analyses focus on the initial review of abstracts and do not evaluate the impact or reliability Program Committees' postreview assessments.

Reliabilities were statistically significantly higher for reviews of STFM completed research projects and preconference workshop as compared with many other STFM categories and NAPCRG reviews. While the use of 7 reviewers for the preconference workshops likely boosted reliability for this category, the mean of 3 reviewers for the research submissions had relatively favorable agreement on abstract ratings (0.50) and a low reviewer-level ICC (0.08). Reviewers assigned to these submissions could have had greater baseline agreement on review criteria or priorities, or the submitted abstract pool could have been more easily parsed if a substantial fraction of submissions happened to be of very high or very low quality.

It is ultimately important for committees using peer review to appreciate its limitations. For higher stakes decisions where reliability is low, such as selection of abstracts for award or distinction, we recommend postreview reassessment by multiple independent reviewers to account for potentially spurious ratings by individual reviewers. For lower-stakes decisions where reliability is low (eg, trainee poster sessions), committees might consider lotteries as more transparent means of selecting abstracts for acceptance versus rejection, hybrid methods (eg, random rejection from the bottom third of abstracts ranked after initial review), or abstract prioritization based on themes selected by Program Committees.

Our analysis has several limitations. First, program committees make final decisions on abstracts based only in part on initial review results; our analysis could not account for the influence of postreview assessments by program committees. Second, we lacked data on reviewer identities for 2019 NAPCRG reviewers so could not account for within-reviewer correlations, which may have affected abstract-level ICCs and estimated reliability for 2019. Third, our analyses only incorporated quantitative data submitted by reviewers and does not assess reviewer text comments. Fourth, our analysis only considered a single year of STFM data; generalizability of results to other STFM years is uncertain.

We conclude that the reliability of peer review of abstracts submitted to academic family medicine meetings is low for many submission categories. While reliability was acceptable in some STFM submission categories, reliability of the initial review process remained low for NAPCRG in 2019 despite an increase the number of peer reviewers compared with previous years. For higher-stakes program committee decisions, such as the selection of distinguished abstracts, our results support supplementing initial peer review with extensive, postreview reassessment by program committee members.

The authors thank Dean Seehusen, MD, MPH for constructive comments on the manuscript.

To see this article online, please go to: http://jabfm.org/content/33/6/986.full.

References

1. Kemper KJ, McCarthy PL, Cicchetti DV. Improving participation and interrater agreement in scoring Ambulatory Pediatric Association abstracts. How well have we succeeded? Arch Pediatr Adolesc Med 1996;150:380–3.
2. Montgomery AA, Graham A, Evans PH, Fahey T. Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. BMC Health Serv Res 2002;2:8.
3. Poolman RW, Keijser LC, de Waal Malefijt MC, Blankevoort L, Farrokhyar F, Bhandari M. Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings. Acta Orthop 2007;78:278–84.
4. Rowe BH, Strome TL, Spooner C, Blitz S, Grafstein E, Worster A. Reviewer agreement trends from four years of electronic submissions of conference abstract. BMC Med Res Methodol 2006;6:14.

5. Rubin HR, Redelmeier DA, Wu AW, Steinberg EP. 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;8:255–8.
6. Ingelfinger FJ. Peer review in biomedical publication. Am J Med 1974;56:686–92.
7. Kravitz RL, Franks P, Feldman MD, Gerrity M, Byrne C, Tierney WM. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? PloS one 2010;5:e10072.
8. Fogelholm M, Leppinen S, Auvinen A, Raitanen J, Nuutinen A, Vaananen K. Panel discussion does not improve reliability of peer review for medical research grant proposals. J Clin Epidemiol 2012;65:47–52.
9. Pier EL, Brauer M, Filut A, et al. Low agreement among reviewers evaluating the same NIH grant applications. Proc Natl Acad Sci USA 2018;115:2952–7.
10. Reinhart M. Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. Scientometrics 2009;81:789–809.
11. Mayo NE, Brophy J, Goldberg MS, et al. Peering at peer review revealed high degree of chance associated with funding of grant applications. J Clin Epidemiol 2006;59:842–8.
12. Carrasco JL, Jover L. Estimating the generalized concordance correlation coefficient through variance components. Biometrics 2003;59:849–58.
13. Kraemer HC. Correlation coefficients in medical research: from product moment correlation to the odds ratio. Stat Methods Med Res 2006;15:525–45.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
15. Carmines EG, Zeller RA, eds. Reliability and validity assessment. Newbury Park, CA: Sage Publications; 1979.
16. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. New York, NY: Oxford University Press; 2006.
17. Nelson EC, Gentry MA, Mook KH, Spritzer KL, Higgins JH, Hays RD. How many patients are needed to provide reliable evaluations of individual clinicians? Med Care 2004;42:259–66.
18. Sattler DN, McKnight PE, Naney L, Mathis R. Grant peer review: improving inter-rater reliability with training. PloS One 2015;10:e0130450.
19. Kaplan D, Lacetera N, Kaplan C. Sample size and precision in NIH peer review. PloS One 2008;3:e2761.