# What Does It Take to Change an Editor's Mind? Identifying Minimally Important Difference Thresholds for Peer Reviewer Rating Scores of Scientific Articles

Michael Callaham, MD*; Leslie K. John, PhD

*Corresponding Author. E-mail: michael.callaham@ucsf.edu.

**Study objective:** We define a minimally important difference for the Likert-type scores frequently used in scientific peer review (similar to existing minimally important differences for scores in clinical medicine). The magnitude of score change required to change editorial decisions has not been studied, to our knowledge.

**Methods:** Experienced editors at a journal in the top 6% by impact factor were asked how large a change of rating in "overall desirability for publication" was required to trigger a change in their initial decision on an article. Minimally important differences were assessed twice for each editor: once assessing the rating change required to shift the editor away from an initial decision to accept, and the other assessing the magnitude required to shift away from an initial rejection decision.

**Results:** Forty-one editors completed the survey (89% response rate). In the acceptance frame, the median minimally important difference was 0.4 points on a scale of 1 to 5. Editors required a greater rating change to shift from an initial rejection decision; in the rejection frame, the median minimally important difference was 1.2 points. Within each frame, there was considerable heterogeneity: in the acceptance frame, 38% of editors did not change their decision within the maximum available range; in the rejection frame, 51% did not.

**Conclusion:** To our knowledge, this is the first study to determine the minimally important difference for Likert-type ratings of research article quality, or in fact any nonclinical scientific assessment variable. Our findings may be useful for future research assessing whether changes to the peer review process produce clinically meaningful differences in editorial decisionmaking. [Ann Emerg Med. 2017;■:1-5.]

Please see page XX for the Editor's Capsule Summary of this article.

0196-0644/$-see front matter

https://doi.org/10.1016/j.annemergmed.2017.12.004

## INTRODUCTION

Increasing the rigor of research on scientific publication, critical analysis, and other important disciplines requires performing well-designed and adequately powered randomized controlled trials, currently an uncommon event. Calculating power and the number of subjects needed for a reliable randomized controlled trial requires knowing how large a difference in outcome is important to detect, an issue that most of this research neither discusses nor reports.

Research on these topics typically relies on subjective data, structured as Likert-type questions with ratings of 1 to 5, indicating the strength of appraisal of some subjective variable (such as overall review quality). However, these Likert ratings are ordinal and have no intrinsic meaning; they do not convey what they mean to users or how users interpret changes in their value. In the rare circumstance in which peer review research even mentions effect sizes, they are typically only subjective guesses by 1 or 2 individuals. In the study that created the best-known quality rating score for peer reviews, itself composed of Likert ratings, this issue is not mentioned.[1]

The same challenge (the meaning of Likert scores) also occurs in clinical medicine, but has been well addressed by the federal mandate to be patient centered. Actual assessments by end users (typically patients) are systematically sought, comparing such ratings with other metrics (such as limitations in activities of daily living or desire for a larger medication dose).[2] In pain management, a reader or researcher therefore can easily determine the best standard for what size change matters clinically (ie, to the patient himself or herself).[3] Thus, statistically

**Editor's Capsule Summary**

*What is already known on this topic*

Article scoring is a standard part of the journal peer review process.

*What question this study addressed*

What is the minimum change in peer reviewer scoring that might alter an editor's decision in regard to publication?

*What this study adds to our knowledge*

In this survey of 41 editors at *Annals of Emergency Medicine* in regard to a hypothetical article, the median decrease in a reviewer rating (on a 1- to 5-point Likert scale) that might trigger their reversal of a tentative acceptance decision was 0.4, and conversely, the median increase in rating that might trigger their reversal of a tentative rejection was 1.2. There was substantial variability between editors and evidence that these ratings had little or no influence on decisionmaking for a large subset of editors.

*How this is relevant to clinical practice*

These findings provide insight into how journal editors interpret and apply the ratings their peer reviewers assign to articles.

"significant" but clinically irrelevant differences can be recognized and appropriately ignored.

### Goals of This Investigation

We sought to determine the minimally important difference for the Likert scores so widely used in assessing research and articles in many disciplines for publication. That is, we sought to identify the magnitude of change in a quality score required to warrant a change in an action threshold. For example, on a 1-to-5 scale of assessing the overall desirability of an article, what magnitude of change would be required for an editor to change his or her mind from rejecting to accepting that article?

The approach of asking participants whether their choice would change as a function of score changes has been used in other research and for varied goals, from determining consumers' willingness to pay[4] to assessing whether patients have successfully articulated their values.[5] In the present research, we sought to assess minimally important difference values for the editors of a peer review process so that, in turn, these values could be used in research to assess whether changes to the peer review process produce clinically meaningful differences in outcomes.

As such, the goal of the present research was not to determine the validity of this common article scoring system, nor was it to assess other inputs to the article evaluation process (eg, reviewers' written comments, editors' direct assessment of the article itself).

## MATERIALS AND METHODS

### Study Design and Setting

The study was an online survey administered to editors making decisions about article acceptance at *Annals of Emergency Medicine*, a medical journal in the top 6% of all scientific journals by impact factor.[6] All participants gave consent and the study was designated exempt by the institutional review board of the university.

### Selection of Participants

All 46 decision editors at *Annals* who had more than 1 year of editorial decisionmaking experience were eligible for enrollment. These editors make the primary decision about acceptance or rejection of articles for publication. Their number reflects the many subspecialty areas of expertise in this specialty. All have academic appointments, which include the top research institutions in the United States. Collectively, these editors had made decisions on 6,429 articles at *Annals* in the previous 5 years, overwhelmingly of original research and excluding case reports, editorials, and other nonoriginal material. Some of them were not in this role at *Annals* for the entire 5-year period, and many had previous editorial experience at other journals. The mean volume per editor across this 5-year period was 149.5 articles (SD 134.2; 95%; confidence interval 108.2 to 190.8) and the median was 122.

### Interventions and Methods of Measurement

An online survey consisting of a thought experiment was constructed to determine minimally important differences (Appendix E1, available online at http://www.annemergmed.com). Each editor was asked to imagine that he or she was making a decision about a specific but typical article with reviews in hand, as he or she had done often in the past. Decision choices included acceptance, substantive revision, or rejection. The editors were asked to suppose that they had made a preliminary decision to accept the article, based on a reviewer's overall rating of 4 for that article (on the journal's "overall desirability" scale, which ranged from 1 to 5). Critically, they further supposed that before sending the decision letter, the reviewer indicated that he or she had made a mistake and was changing the rating. Editors were asked how, if at all, their decision

would change as a function of the reviewer's new rating. Specifically, they were asked whether their decision would change for each of 5 possible reviewer rating changes: from 4 to 3.8, 3.6, 3.4, 3.2, and 3.0.

Because question framing affects responses,[7] editors also went through a mirror "rejection frame" scenario in which they supposed the reviewer's initial rating to be a 3 and that their preliminary decision was to reject the article. Then, they indicated whether a reviewer rating changes from 3 to 3.2, 3.4, 3.6, 3.8, and 4.0 would cause them to change their decision. The order of presentation of the 2 scenarios was randomly counterbalanced between participants.

We also administered 2 additional questions at the end of the survey. The first assessed, given 2 articles, one with a rating of 4 and another of 5, how likely editors would be to favor the 5 based primarily on that rating. The second assessed how much attention editors typically pay to the "overall desirability for publication" rating. Finally, a text box was provided for editors to provide comments as they wished.

### Data Collection and Processing and Primary Data Analysis

Elsevier, the publisher of *Annals*, provides article tracking and peer review evaluation software, Editorial Manager, which is the industry leader used by thousands of scientific journals and the majority of major scientific publishers.[8] This software uses Likert-style ratings as the default format for evaluation scores.

For each editor, we calculated the score change needed for him or her to make any change to the decision. For example, suppose an editor indicated that with a reviewer score of 4 out of 5, he or she would accept the article, and that when successively asked to imagine that the reviewer score changed to 3.8 and 3.6, he or she indicated that the decision would be unchanged, but that at 3.4, he or she would change the decision. In this case, 3.4 was the tipping point for a decision by that editor for that question and represented a score change of 0.6 Likert points as the minimally important difference. We use descriptive statistics to summarize the results. For the purpose of comparison, the same survey was administered to peer reviewers with no editorial experience, as discussed in more detail in Appendix E2, available online at http://www.annemergmed.com.

## RESULTS

The online survey was sent to 46 editors, 41 of whom completed it, for a response rate of 89%. In the acceptance frame, 4 respondents gave internally inconsistent responses; for example, indicating that a score decrease from 4 to 3.8 would shift the editor from acceptance to rejection, but that a decrease from 4 to 3.6 would not. This rate was similar in the rejection frame, in which 5 respondents gave incoherent responses. Given that the editors were asked to engage in a somewhat cognitively complex thought experiment, this rate of incoherent responses struck us as low and increased our confidence that respondents took the task seriously and provided meaningful responses. Results are restricted to coherent responses.

For the acceptance frame, in which the initial reviewer score was 4 and respondents imagined reviewer-revised scores decreasing by 0.2 points down to 3, 14 editors (38%) specified a minimally important difference of 0.2 points; another 14 (38%) never changed their initial decision within the offered maximum 1-point range of this study. The median change in rating needed to trigger a decision change (the minimally important difference) was 0.4 points (mode 1.2; interquartile range 1.0) (Figure 1).

For the rejection frame, in which the initial reviewer score was 3 and respondents imagined reviewer-revised scores increasing by 0.2 points up to 4, only 4 (11%) had a minimally important difference of 0.2 points; 19 (51%) never changed the decision within the 1-point range. The minimally important difference median was 1.2 (mode 1.2; interquartile range 0.6), statistically significantly higher than that of the acceptance frame.

Sixty-five percent of editors elaborated free-text comments on their responses to the above scenarios. Half of these comments conveyed that their editorial decisions did not depend solely on a rating score in isolation for their decisions and that the article and details in the reviews were also important.

When asked how likely they would be to favor an article with a rating of 5 over one with a rating of 4, primarily according to the rating, editors' median response was 2 ("somewhat likely"); no ratings were more positive than "neutral." Thirty-six percent answered 1 ("not likely at all"), 36% 2 ("somewhat likely"), and 28% 3 ("neutral").

When asked how much attention they typically paid in routine practice to the "overall desirability for publication" rating, the median response was 1.9; none selected the higher value ratings (4 and 5), 36% chose 1 ("not likely at all"), 36% chose 2 ("somewhat likely"), and 28% chose 3 ("neutral").

## LIMITATIONS

Our study examined editors from a single journal; a more definitive study would involve hundreds of editors from a large number of journals. The decision editors at our journal were chosen because of the uniformity of their practice and the likelihood of high response rates.

Although respondents were experienced editors performing a judgment that is a routine part of their responsibilities, they knew this was a

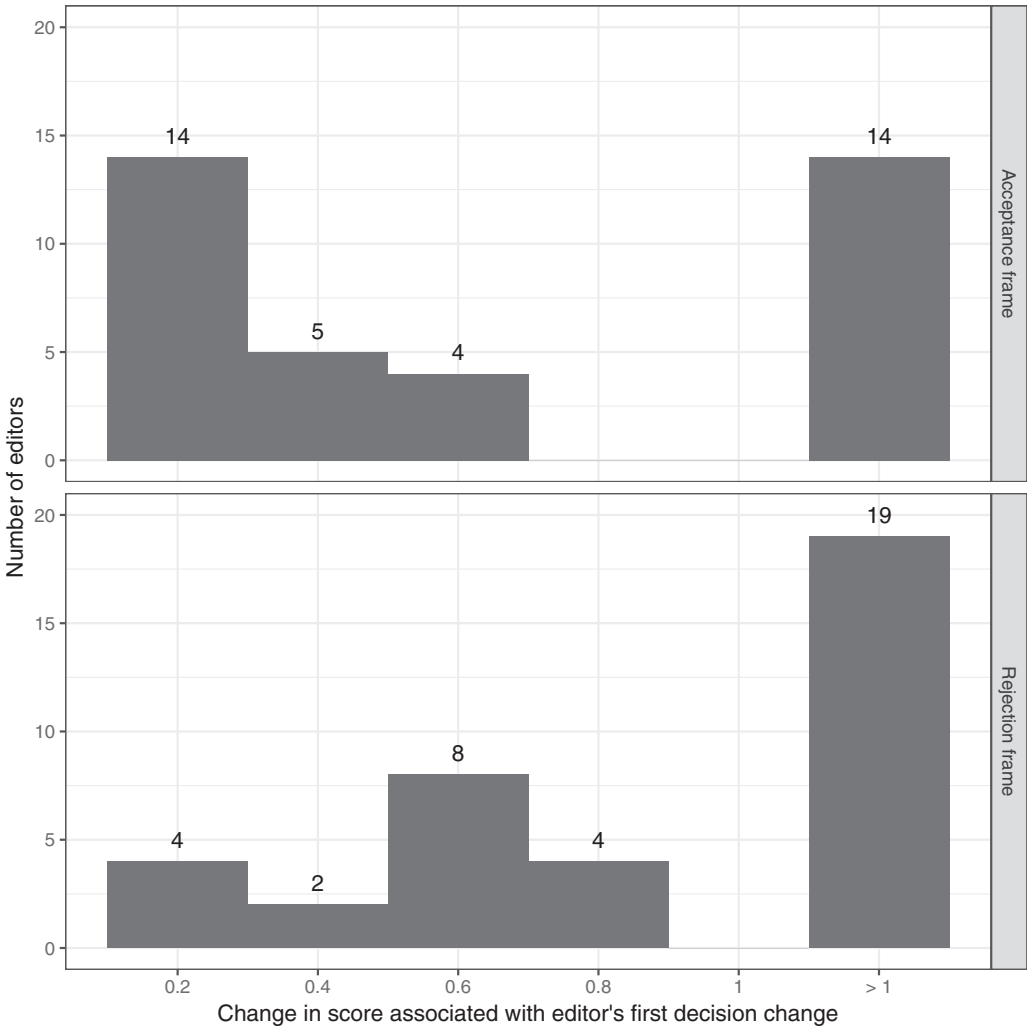


**Figure 1.** Distribution of tipping points invoked by various editors, ie, the magnitude of change needed to lead to a change in decision about article acceptance. Editors who had not changed their decision by the time the quality score had changed 1 full point were assigned to ">1."

hypothetical situation, so the results demonstrate what they thought they would do and not necessarily their actual decisions.

We studied a maximum point change of 1 (on a scale of 1 to 5) rather than the entire spectrum to keep the task from becoming too complex and decreasing compliance, and also because editors believed that the minimally important change was likely to be smaller than 1 point.[9] This point change was located between 3 and 4 on our scale, which ought to best reflect the importance of the score (because at this journal, virtually all submissions with a score less than 3 are rejected, and ratings of 5 are rare).

We studied the minimally important difference in isolation, absent corresponding additional inputs typically factored into editorial decisions. Therefore, we necessarily focused on quantitative scores.

## DISCUSSION

To our knowledge, this is the first study to have examined the minimally important difference for Likert-type ratings of research article quality, or in fact any nonclinical scientific assessment variable. In patient care, it is well recognized that such a rating (as for pain) is inaccurate and of poor utility unless it is known what changes in ratings mean to the user (in that case, the patient; in ours, editors and researchers). The practice of determining such minimally important differences for scales used to assess clinical conditions is now commonplace in the medical literature.[2]

However, in the scientific publication literature, this concern about minimally important differences is almost completely absent. In the few instances in which an effect size or minimally important difference is mentioned and

used to calculate power, it has been an intuitive and unexplained approximation. We are not aware of any previous systematic attempt to determine this threshold for purposes of powering controlled trials.

Consistent with research in psychology and allied fields documenting a "negativity bias,"[10] editors had a bias in favor of rejection; they were more reluctant to change their decision when starting with an unfavorable decision than a favorable one. In the acceptance frame, 38% of editors changed their mind with the smallest possible change (of 0.2). In contrast, in the rejection frame only 11% of editors did so. In both frames, a considerable percentage of respondents did not change their decision even with the largest possible change of 1.0 (acceptance frame 38%; rejection frame 51%).

Broadly, the framing effect raises the question, What is the "correct" frame? And what is the true minimally important difference? Is there even such a true value? At minimum, our results indicate that it is contained within the range bounded by the minimally important difference produced by the 2 frames. More broadly, the framing effect is consistent with the broader conclusion of behavioral decision research: people's preferences and choices can be highly malleable and shaped by seemingly arbitrary contextual factors. Here, responses differed as a function of 2 formally equivalent scenarios. Although each scenario asked the same essential question (how much would the reviewer's score have to change for you to have made a different editorial decision?), significantly different answers were given as a function of the frame. From a practical standpoint, our study additionally speaks to the importance of careful consideration of question wording and frames.

The observed heterogeneity in minimally important differences, both within and across editors, may raise the question of whether Likert ratings alone are a useful article assessment method. It may be best to think of Likert (quantitative) and written (qualitative) article evaluation measures as complementary, the strengths and weaknesses of which offset each other. However, the latter pose significant logistical difficulties for everyday use.

Our results represent an initial effort to improve the quality of study design in scientific assessment. Future research is needed to assess whether the minimally important differences documented herein generalize to other journals.

*The authors gratefully acknowledge Andrew Marder, MA, statistician and programmer, Harvard Business School, for his help with statistical design and analysis.*

*Supervising editor:* Steven M. Green, MD

*Author affiliations:* From the UCSF School of Medicine, San Francisco, CA (Callaham); and Harvard Business School, Boston, MA (John).

*Author contributions:* MC conceived the study, designed the trial and survey instrument, supervised data collection and analysis, and drafted the manuscript. LJ designed the trial and survey instrument, interpreted the results, and substantially revised the manuscript. AM performed the statistical analysis. MC takes responsibility for the paper as a whole.

All authors attest to meeting the four ICMJE.org authorship criteria: (1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND (2) Drafting the work or revising it critically for important intellectual content; AND (3) Final approval of the version to be published; AND (4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

*Funding and support:* By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). Dr. Callaham is paid a stipend by ACEP to serve as editor of the journal.

*Publication dates:* Received for publication September 19, 2017. Revision received November 21, 2017. Accepted for publication November 30, 2017.

Dr. Green was the supervising editor on this article. Dr. Callaham did not participate in the editorial review or decision to publish this article.

## REFERENCES

1. Van Rooyen S, Black N, Godlee F. Development of the Review Quality Instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol*. 1999;52:625-629.
2. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312:1342-1343.
3. Serlin RC, Mendoza TR, Nakamura Y, et al. When is cancer pain mild, moderate, or severe? grading pain severity by its interference with function. *Pain*. 1995;61:277-284.
4. Becker GM, Degroot MH, Marschak J. Measuring utility by a single-response sequential method. *Behav Sci*. 1964;9:226-232.
5. John LK, Fischhoff B. Changes of heart: the switch-value method for assessing value uncertainty. *Med Decis Making*. 2010;30:388-397.
6. Thomson Reuters. InCites Journal Citation Reports. Available at: https://jcr.incites.thomsonreuters.com/JCRJournalHomeAction.action?SID=A1-GyPbP6YZeDYieYywwAexxNopcEYQRyzvp-18x2dS9HIDaFHCFcPrQKQ8eH2vwx3Dx3DXEET5rNUAh0CLdCBUwhCswx3Dx3D-iyiHxxh55B2RtQWBj2LEuawx3Dx3D-1iOubBm4x2FSwJjjKtx2F7IAaQx3Dx3D&SrcApp=IC2LS&Init=Yes Accessed August 8, 2017.
7. Schwarz N. Self-reports: how the questions shape the answers. *Am Psychol*. 1999;54:93-105.
8. Aries Systems. Editorial Manager. Available at: https://www.ariessys.com/software/editorial-manager/. Accessed August 15, 2017.
9. Justice AC, Cho MK, Winder MA, et al; PEER Investigators. Does masking author identity improve peer review quality? a randomized controlled trial. *JAMA*. 1998;280:240-242.
10. Baumeister RF, Bratslavsky E, Finkenauer C, et al. Bad is stronger than good. *Rev Gen Psychol*. 2001;5:323-370.

## APPENDIX E2
## COMPARISON SURVEY OF PEER REVIEWERS

### Reviewer survey results

For comparison to the editors discussed in our article, we also surveyed 97 of our "superreviewers." Each year, our 50 top-performing reviewers are identified according to review volume, acceptance and completion rates, timeliness, and quality of review. Superreviewers are defined as those who qualified for that list at least twice in a 4-year period during the preceding decade; 36% of this group did so more than twice, to a maximum of 10 times. However, these reviewers do not make decisions about acceptance, nor do they know the views of the other reviewers of an article, before they complete their own assessment.

An identical survey was administered to the 97 current superreviewers to determine how their responses compared with those of the editors. Sixty-three responded with a complete survey (65% response rate). Eight answers to the acceptance frame (question 1) and 3 to question 2 (rejection frame) were internally inconsistent and were excluded from analysis.

Their responses differed from the editors', presumably because of their lack of actual decisionmaking experience. Unlike that of the editors, reviewer responses to both acceptance and rejection frame questions were very similar, as were their histograms (Figure E1), with half the respondents making a decision with the smallest possible minimally important difference (0.2 points, more sensitive to rating change than the editors) and only 7% still having made no decision even at 1 full point. On the question of how likely they would be to accept an article with a score of 5 rather than a 4, based primarily on the rating, 38% of respondents chose "very likely" and 35% "quite likely"; only 5 chose scores less than 4 ("neutral" to "not likely at all"). On the question of

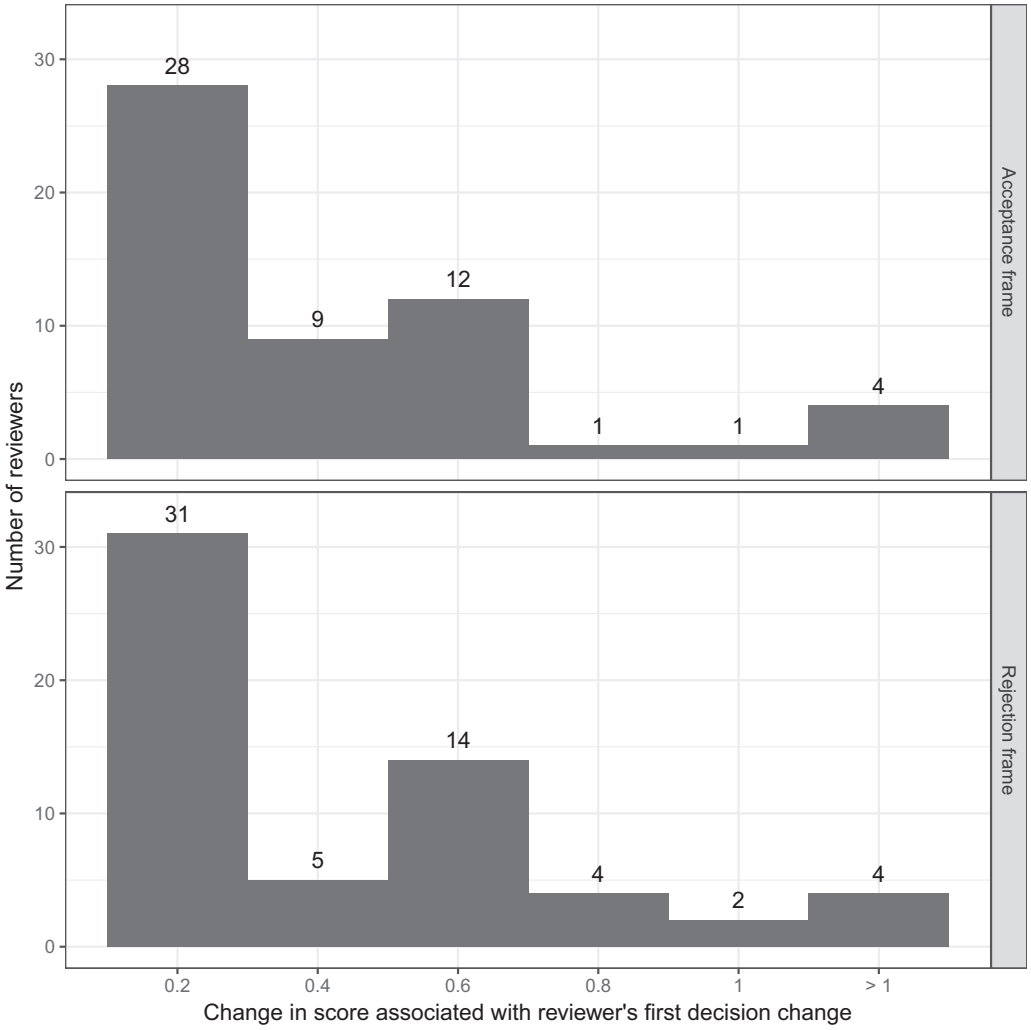


**Figure E1.** Tipping points invoked by superreviewers (no editorial decisionmaking experience).

how likely they were to review this rating in real life, 42% chose "almost always" and 35% "often"; only 14 (23%) chose a score of 3 ("about half the time") or less.

The heterogeneity of our superreviewer vs editor results suggests that future research will need to define the minimally important difference for a much larger number of editors, journals, and assessors. Uniformity should not be assumed, and there may be important variation between editors in specific specialties and journals.