Manuscript Quality before and after Peer Review and Editing at *Annals of Internal Medicine*

Steven N. Goodman, MD, PhD; Jesse Berlin, ScD; Suzanne W. Fletcher, MD, MSc;
and Robert H. Fletcher, MD, MSc

■ *Objective:* To evaluate the effects of peer review and editing on manuscript quality.

■ *Setting:* Editorial offices of *Annals of Internal Medicine*.

■ *Design:* Masked before-after study.

■ *Manuscripts:* 111 consecutive original research manuscripts accepted for publication at *Annals* between March 1992 and March 1993.

■ *Measurements:* We used a manuscript quality assessment tool of 34 items to evaluate the quality of the research report, not the quality of the research itself. Each item was scored on a 1 to 5 scale. Forty-four expert assessors unaware of the design or aims of the study evaluated the manuscripts, with different persons evaluating the two versions of each manuscript (before and after the editorial process).

■ *Results:* 33 of the 34 items changed in the direction of improvement, with the largest improvements seen in the discussion of study limitations, generalizations, use of confidence intervals, and the tone of conclusions. Overall, the percentage of items scored three or more increased by an absolute 7.3% (95% CI, 3.3% to 11.3%) from a baseline of 75%. The average item score improved by 0.23 points (CI, 0.07 to 0.39) from a baseline mean of 3.5. Manuscripts rated in the bottom 50% showed two- to threefold larger improvements than those in the top 50%, after correction for regression to the mean.

■ *Conclusions:* Peer review and editing improve the quality of medical research reporting, particularly in those areas that readers rely on most heavily to decide on the importance and generalizability of the findings.

Ann Intern Med. 1994;121:11-21.

From Johns Hopkins University School of Medicine, Baltimore, Maryland; the University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; and Harvard Medical School, Boston, Massachusetts. For current author addresses, see end of text.

Publication of medical research has high stakes: the communication and legitimization of medical research, the advancement of authors' careers, priorities in funding decisions, the direction of future research, and the visibility and prestige of journals themselves. Peer review and editing play central roles in the publication process, affecting the acceptance of a manuscript and the form in which it appears. The most commonly heard justification of peer review is that it is an indispensable aid to an editor in assessing the importance of a scientific question and in assessing how well that question has been answered (1, 2). However, it has also been criticized as being inherently conservative, censorial, and, perhaps worst of all, arbitrary (3). A frequently heard charge is that peer review delays the dissemination of crucial medical information without commensurate benefit (4-7).

During the last several years, some medical journal editors decided that the value of peer-review and editing practices should be examined with the same rigor demanded for testing medical hypotheses (8, 9). The First International Congress on Peer Review was organized in 1989 (10), bringing together medical journal editors and other interested scholars to present and discuss research on peer review; a second Congress was held in 1993.

The peer-review process has two components: the assessments by external reviewers and the decisions and actions taken by editors, which are partially affected by comments from the reviewers. To our knowledge, no study has evaluated the effects of peer review and editing on manuscript quality once the decision to accept has been made, and a computerized search of *Index Medicus* back to 1966 failed to locate any such studies. In this paper, we present the results of such a study, assessing the change in a manuscript between the times of provisional acceptance and final publication. We studied whether the quality of accepted manuscripts was improved by peer-review and editorial processes and, if it was, which aspects were most improved.

Methods

Setting

The study was conducted at the editorial offices of *Annals of Internal Medicine*. *Annals*, a specialty journal in internal medicine, is published twice monthly and has a circulation of approximately 100 000. *Annals* receives approximately 2400 manuscripts annually, of which half are reports of original research. During the period of this study, the investigators included the editors of *Annals* (RHF and SWF) and a statistical associate editor (SNG).

The Review Process

No change was noted in the usual review and editing procedures at *Annals* during the time of this study. All manuscripts received at *Annals* were initially reviewed by one of two full-time editors or one of two half-time deputy editors, as well as by one of seven associate editors, all of whom are faculty members of medical schools in Philadelphia and have subspecialty interests (for example, infectious disease, gastroenterology). Half of the submissions were returned to authors without further review and half were sent to at least 2 outside reviewers, selected by the associate editor from a database of about 7000 reviewers. After comments from the reviewers were received, the original editor and associate editor reassessed each manuscript and chose which ones would be discussed at a weekly editorial conference of editors, deputy editors, medical associate editors, and two statistical associate editors. Factors that affected acceptance decisions included the quality of the research, the importance of the question, the contribution of the finding to its field, the utility and interest for *Annals* readers, the quality of the presentation, the priority relative to other articles, and available space. Authors were notified either that the editors would not accept the paper, that the editors were willing to reconsider the paper after major revisions, or that the paper was provisionally accepted, pending satisfactory revision. Approximately one third of the articles evaluated by outside reviewers were accepted, 15% of submitted original research articles.

Papers to be considered further were sent to authors, along with the comments of the two outside reviewers, comments of one of the statistical editors, and a letter from one of the editors or deputy editors (which summarized the discussion at the weekly conference, the ideas of the associate editor, and suggestions from the editor). In addition, each manuscript was reviewed by a production editor, and directions for changes in manuscript wording or layout of figures and tables were included.

All revised manuscripts were reviewed by the editor or deputy editor in charge of the manuscript, the appropriate associate editor, the statistical editor, and the production editor. Some revised manuscripts were also reassessed by the original outside reviewers. Approximately half of the revised manuscripts were returned to authors for further revision. Most revised manuscripts (>95%) were ultimately published.

The time taken by this process was approximately 2 weeks for the initial decision to review or reject, 8 additional weeks to review and make an acceptance decision, 8 weeks until final acceptance, and about 4 months until publication. More than 95% of manuscripts submitted to *Annals* had a provisional acceptance or rejection decision sent to the authors within 3 months. The average time from submission to publication was about 7 months, with initial peer review accounting for approximately 6 weeks.

Manuscript Selection and Study Design

All original research manuscripts (articles) accepted for publication by *Annals* from March 1992 to March 1993 were entered into the study after obtaining the author's consent. Commentaries, reviews, expository pieces, editorials, and brief reports were not included. This study had a before-after design, in which two versions of each manuscript were evaluated: the version originally submitted and the version sent to the printer for publication after all modifications based on peer review, editors' comments, and copyediting. All "before" and "after" manuscripts were in electronic form and were reformatted to make the appearance of the two versions identical. Authors' names and affiliations were removed. The design of the study was approved by the Institutional Review Board of the University of Pennsylvania School of Medicine.

Definition of "Quality"

Manuscript "quality" can be separated conceptually into two components: the quality of the research itself, and the quality of the research *report*. The quality of the research *report* was evaluated in this study. It was defined as follows on the cover sheet of the quality assessment instrument: "*Whether the authors have described their research in enough detail and with sufficient clarity so a reader could make an independent judgment about the strengths and weaknesses of their data and conclusions.*"

Manuscript Quality Assessment Instrument

A 34-question instrument was developed to structure the assessment of the quality of a manuscript (Appendix). Items were derived from published checklists (11-14), articles about the contents of journal articles (15-18), the authors' editorial experience, and the comments of journal editors and methodologists who reviewed drafts of the instrument. Each question could be answered on a 5-point ordinal scale, where 1 was worst and 5 was best. The instrument was organized along the same dimensions as a standard journal article: Title and Abstract (2 items), Introduction (2 items), Methods (7 items), Results (15 items), Discussion and Conclusions (4 items), and General Evaluation (4 items). An additional question asked for a subjective assessment of the manuscript's overall quality on a 10-point scale.

The instrument differed from previously published quality scoring schemes in several ways. It was not a checklist but rather was a set of structured judgments, graded ordinally, allowing users the discretion not to penalize a manuscript if a detail was omitted that was not critical to the study's interpretation. Also, in keeping with the definition of quality given above, each question was about the adequacy of the reporting rather than the quality of the research itself.

Assessment

A panel of 44 physicians and epidemiologists with training in research methods and in critically assessing the medical literature was recruited to serve as an independent panel of expert assessors ("experts"). They did not receive any formal training in the use of the assessment instrument, although general guidelines were given on the cover sheet (Appendix). The panel was masked to the design and aims of the study; they were told only that they were participating in a "study of manuscript quality" for *Annals*.

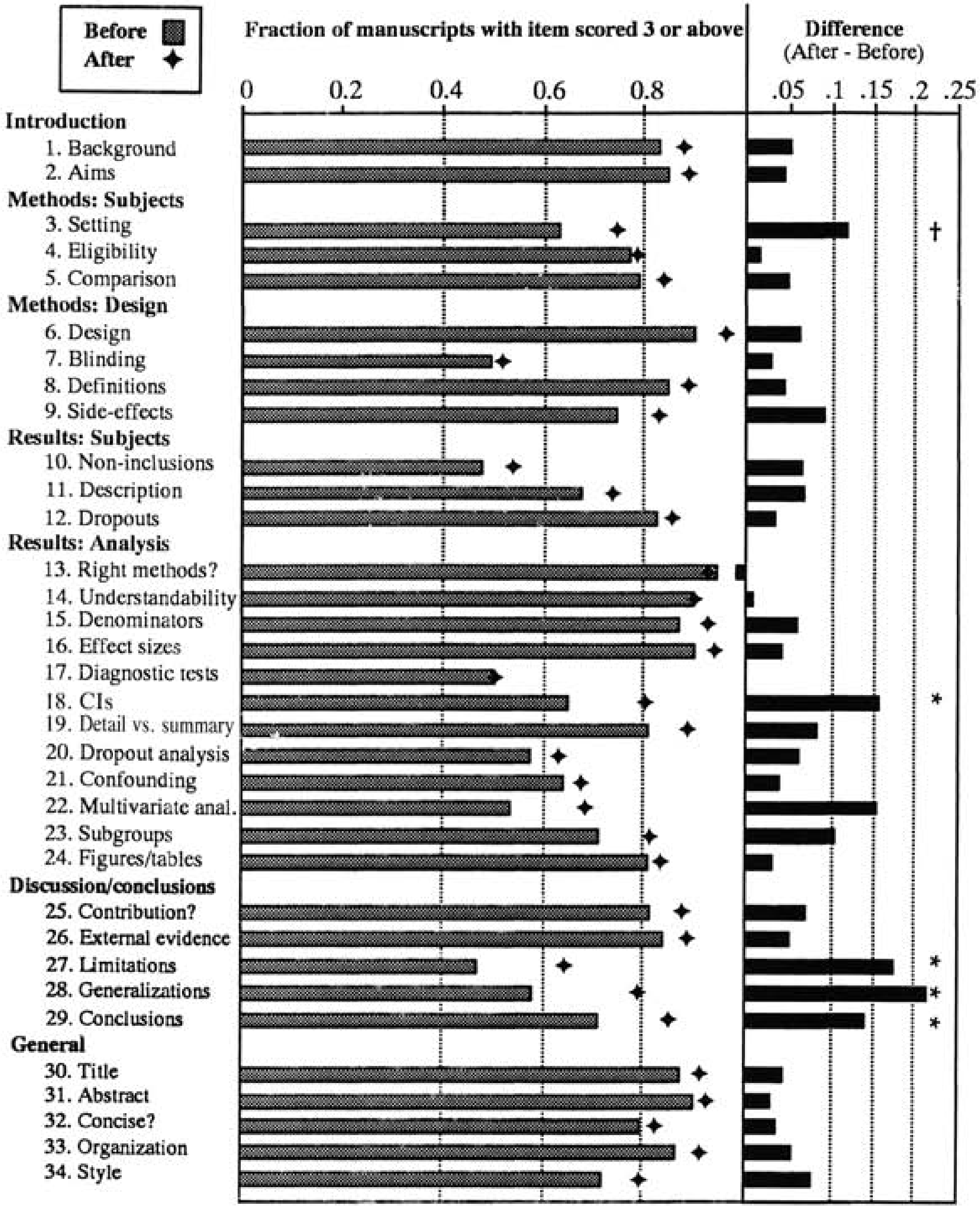
"Before" and "after" versions of each manuscript were randomly assigned to different experts to prevent the bias that might have been introduced if they could infer the design of the study and thereby which manuscript had been through the editorial process. Thirty-two manuscript versions were given to two or three experts to assess the reliability of the instrument; all others were assessed by only one expert.

Statistical Analyses

The study was designed to have 90% power to detect a 0.5 unit change in average score, assuming a within-manuscript standard deviation of 1 scale unit, using $\alpha = 0.05$. The main outcome measure was the percentage of items that were scored 3 or higher on the 5-point scales (percentage score). The average of all score components (average score) was also analyzed. Linear regression was used to assess the effect of revision on each of these outcome measures, with terms controlling for manuscript and reviewer.

Item-specific analyses were done on dichotomized item scores (0 for ratings ≤ 2 and 1 for ratings ≥ 3). The change from before revision to after revision in individual items was statistically assessed with conditional logistic regression, which allowed for variable group sizes (because of the reliability study) and for the correlation of scores within the manuscript. For displaying the item results, manuscripts that had more than one evaluation had their original numerical score (1 to 5) averaged and then converted into a dichotomous variable. The reported percentages represent the fraction of manuscripts that rated a specific item of 3 or more. Reliability of the instrument was assessed with the intraclass correlation coefficient.

Linear regression was also used to assess the effect of initial quality on the before-after change. Control for regression to the mean was achieved by a median split of the manuscripts according to the average of their before and after scores. Thus, manuscripts rated in the bottom 50% were not those with the lowest initial scores but were those whose before-after average was in the bottom 50%. All confidence intervals (CI) are 95%. JMP (SAS Institute, Carey, North Carolina) and Egret (Seattle, Washington) statistical software packages were used for all of the analyses.



Item Analysis

Figure 1. Item analysis of the study. The length of the gray bars in the left-hand bar graph represents the fraction of unrevised manuscripts in which the given item was rated three or above (n = 111). The diamonds (◆) are the corresponding scores of the revised versions. The black bars on the right-hand graph represent the change because of revision, that is, the difference between the after scores and the before scores (or distance in the left-hand graph between the end of the bar and the diamond). *$P < 0.05$, †$P = 0.07$.

Results

From 1 March 1992 through 1 March 1993, 113 reports of original research were accepted for publication in *Annals*, and all were entered into the study. There were a total of 258 evaluations: 111 manuscripts were evaluated before and after revision, 30 in duplicate and 2 in triplicate, and 2 had only 1 version evaluated. Most of the replications were done on "before" manuscripts; 143 evaluations of manuscripts were done before revision and 115 after revision. The 2 manuscripts with only 1 version evaluated were excluded from the analysis.

Baseline Quality: Item Analysis

Figure 1 presents, for each item in the questionnaire,

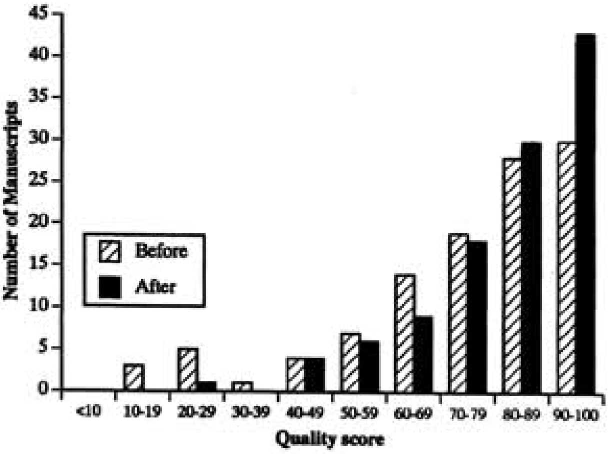


Figure 2. The distribution of manuscript quality scores before and after revision. The quality score displayed here is the percentage of items for each manuscript that was rated above 3 (out of 5).

the percentage of manuscripts before and after revision for which that item was rated 3 or more. The five items that were rated most deficient at the time of submission were discussion of limitations (47% rated ≥3), the description of participants not included in the study (48%), description of blinding (49%), reporting of summary statistics for diagnostic tests (50%), and quality of multivariate reporting (53%). (Only 17 manuscripts assessed diagnostic tests and 78 used multivariate methods.) The five items scoring highest at the time of submission were use of correct analytical methods (94%), clarity of description of research design (90%), quality of the abstract (90%), understandability of the quantitative presentation (90%), and reporting of effect sizes (90%).

Changes after Revision: Item Analysis

Thirty-three of the 34 items improved after peer review and editing, with 1 decreasing by a negligible amount (1%). Four items showed statistically significant improvements, with a fifth having borderline significance. (By chance, less than 1 out of the 34 univariate comparisons would be expected to improve in a statistically significant extent.) The 5 items were discussion of limitations (before-after change from 47% to 65%, $P < 0.001$), acknowledgment and justification of generalizations (58% to 79%. $P < 0.001$), appropriateness of the strength or tone of the conclusions (71% to 85%, $P = 0.01$), use of confidence intervals (65% to 81%, $P < 0.001$), and description of the setting 67% to 74%, $P = 0.07$). Analysis of the ordinal scores showed similar results. The clarity of multivariate analyses seemed to have a sizable change (53% to 69% $P = 0.23$), but the estimate of the change was imprecise because only 78 of the manuscripts used these methods and the rating changed in only 17 manuscripts.

Changes after Revision: Summary Scores

Figure 2 shows the distribution of percentage scores before and after revision. A noticeable change occurred in the two extremes of the distribution of quality scores almost all of the lowest scores were eliminated after revision, and an increase of 43% (from 30% to 43%) occurred in the number of manuscripts that scored more than 90%. The mean initial percentage score was 75% (25th to 75th percentile: 66% to 90%), the mean average score was 3.5 (SD 0.8), and the mean subjective score (using a 10-point scale) was 5.4 (SD 2.5).

The results of linear regression, which included terms for manuscript and revision status, showed that the percentage score increased by an absolute 7.3% (CI, 3.3% to 11.3%; $P < 0.001$), and the average score increased by 0.23 scale units (CI, 0.07 to 0.39; $P = 0.005$). An analysis eliminating the five individual items with the smallest P values showed that the percentage score still changed by 5.7% (CI, 1.7% to 9.7%; $P = 0.006$). The 10-point subjective score showed no statistically discernible change, increasing by 0.29 units (CI, −0.25 to 0.83; $P = 0.3$).

Peer review and editing improved poor manuscripts more than those that were already good at the time of submission. Among manuscripts rated in the bottom 50%, the percentage score increased by 10.3% (CI, 3.9% to 18.7%) from a mean initial score of 63%. Manuscripts rated in the top 50% increased by 4.5% (CI, 1% to 8%) from a mean baseline of 88%. This analysis stratified the manuscripts by the average of the before and after scores to control for regression to the mean.

Reliability of the Instrument

The reliability of the instrument, measured in 32 manuscripts, was low. The intraclass correlation coefficient calculated on this subsample was 0.12 (CI, −0.22 to 0.44) for the average score and was 0.02 (CI, −0.30 to 0.36) for the percentage score.

A different estimate of reliability emerged from the complete sample. Ignoring the before-after status and using all 111 manuscripts, the intraclass correlation coefficient was 0.25 (CI, 0.14 to 0.36) for the average and percentage scores. The intraclass correlation of the best subscales was approximately the same. Because before-after differences were observed in the total sample, ignoring them should underestimate the intraclass correlation. The subsample of manuscripts used for the reliability study may have by chance underestimated the true reliability. The overlapping confidence intervals on the estimates from the two samples show that they are statistically consistent, although unlikely to be as low as indicated in the subsample.

Reviewer Effect

Because each expert rated four to seven manuscripts, it was theoretically possible to estimate how severely each graded on average (thus controlling for the average "toughness" of the reviewers). Adding the 44 reviewers to the linear regression model (which had terms for "manuscript" and "before-after status") increased the model R-squared (percentage of variance accounted for by the model) from 61% to 80% for the percentage score and from 60% to 85% for the average score. Similar changes were seen in the adjusted R-squared, which takes into account the number of terms in the model. Including reviewers in the model decreased the estimate of the revision effect by about 20% for percentage and average scores. This occurred because a substantial proportion (75%) of reviewers had an uneven before-after split of manuscripts, which produced a correlation between a reviewer's average score and the before-after status of the manuscript. Therefore, it was judged that omitting the reviewer term from the model produced the more accurate measure of change.

Discussion

We found that peer review and editing increased the quality of articles reporting original research in *Annals of Internal Medicine*. Almost all components of the articles improved (*see* Figure 1). Of the five instrument items that showed a statistically discernible change, two related to the generalizability of the results (description of achieved sample and justification of generalizations), two related to the weight authors put on the findings (discussion of limitations and tone of conclusions), and one related to certainty of the point estimates of effect (confidence intervals). Although the changes were small for most other individual components, they were consistent enough so that even when the five most statistically significant items were eliminated, the change in the percentage score was still significant ($P = 0.006$). These changes were apparent to two independent readers without specialized content expertise, each of whom saw only one manuscript version. By not comparing the original and modified manuscripts side by side, the design mirrored how articles are actually read and also probably underestimated the true change.

The manuscript components with the largest changes (with the exception of the confidence intervals) were those that involved the most complex and subjective judgments by the raters. The types of changes observed were consistent with the view that peer review is a "negotiation between author and journal about the scope of the knowledge claims that will ultimately appear in print" (17, 18). Descriptions of generalizations, limitations, and conclusions are powerful determinants of how research findings are received by the scientific community and the media.

Among individual items that changed to a statistically significant extent, the presentation of limitations was rated lowest in the submitted manuscripts. This may reflect a perception among authors that presenting a study's shortcomings weakens a report. It is not possible to say whether the peer-review and editing process made authors aware of limitations they had not previously considered or whether it forced them to acknowledge weaknesses they already knew about. Of the other items, the use and proper interpretation of confidence intervals have been a policy at *Annals* for several years. Their use is important because they are an excellent way to convey the variability of results and to shift a reader's perspective from "testing" to "estimating" an effect (19–22). One of the largest improvements was seen in the quality of reporting of multivariate methods, although this was not statistically significant, probably because only a subset ($n = 78$) of manuscripts used them. We believe the observed change is probably real because of our experience that these presentations are frequently changed in the editing process, and our observation that many investigators do not understand these methods well. Problems with the reporting of multivariate methods in the medical literature have been documented (18, 23).

We studied the quality of the research *report*, as distinguished from the quality of the research, for two reasons. First, with the exception of the statistical analyses, the quality of completed research generally cannot be improved, whereas the report of that research can be. Second, the quality of the underlying research can only be evaluated if the report is clear and complete. If the excellence of research cannot be appreciated from its written description, the results will not be accorded their appropriate weight by readers. Conversely, an imperfect study, with flaws hidden, can be accorded more importance than it deserves. This focus on the quality of reporting implicitly emphasizes the role of the journal as a vehicle for scientific exchange rather than as a static repository of scientific facts. As the editor of *Chest* has suggested (24), published research can be seen as "the *onset* of a dialogue in the establishment of scientific truths." This is a different perspective than that expressed in some of the heated debates about peer review (25–27).

Although we found that peer review and editing at *Annals* produce measurable and substantive improvements, the degree of overall improvement was modest, and there was still substantial room for improvement in quality scores after manuscript revision. Several possible explanations exist for this observation. Limitations on editors' time means that accepted manuscripts can only be improved, not made perfect. Also, the design of the study may have accounted for finding only moderate improvements. Many experts said they had difficulty separating judgments about the quality of the research from the quality of the report. Although the covering letter emphasized the difference (Appendix) and the questions reinforced it, written and verbal comments of the experts indicated that they did not always make that distinction. Improvement in some manuscripts may have made research flaws more apparent, resulting in paradoxically lower ratings. This might have been corrected if the experts had formal training in the use of the questionnaire. If the same expert saw both versions of the manuscript, larger improvements probably would have been found, but then their judgments would not have been masked. Although our study suggests that further efforts to improve manuscript quality are necessary, it may be equally important to minimize unrealistic expectations of peer review, to increase readers' awareness of problems in the literature, and to enhance critical reading abilities.

The study's design did not allow us to distinguish the effects of external peer review from those of internal editing. We do not know how many of the observed changes could have been produced solely by in-house editing, nor could we measure the effect of peer review without a journal or editor, such as with "Clinical Alerts" from the National Cancer Institute (28, 29). In many instances, the effects are not separable because comments by outside reviewers often identify areas for editorial scrutiny. Some of the improvements we found, particularly those related to the use of confidence intervals and multivariate methods, are more often requested by editors at *Annals* than by peer reviewers.

Cicchetti (25) has published a comprehensive review of the peer-review research conducted in a broad range of disciplines up to 1991. Most previous studies of peer review have investigated the reliability and agreement of reviewers' and editors' decisions, and the degree to which unrelated factors (for example, author's sex or institutional affiliation) affect these assessments (25, 27). Lock (30) did one such study at the *British Medical Journal*; he informally measured the change (from submission to publication) of accepted articles and rejected ones that were subsequently published in other journals. He reported that 50% of articles published in the *British Medical Journal* were changed in some manner compared with 20% of rejected articles published elsewhere, although the nature of that change was not determined. Gardner (31) conducted an unmasked before-after study (using one assessor) of the statistical components of submitted and published papers, focusing on the process of in-house statistical review. Of 45 papers, he found that only 5 (11%) were statistically acceptable at submission and 38 (84%) were adequate by the time of publication. Researchers who are not editors at journals lack accessibility to the peer-review process, which has made investigation in this area difficult, as has resistance to such studies by some journal editors (24–26).

Our study is different from previous research in two main respects: It describes how peer review and editing improve a manuscript, rather than guide the selection of what is published, and it tried to directly and comprehensively assess manuscript quality, instead of using surrogate measures like reviewers' global assessments or subsequent citation rates. The effect of peer review and editing after manuscript acceptance is important because many biomedical research reports are ultimately published (25, 27, 30, 32). The extent to which this effect differs across journals may be a partial explanation for the variable degree of weight accorded to research results published in different journals.

Our results suggest why specific comments from reviewers are more useful than global assessments that do not agree as shown in many studies (25, 33). We found that the structured quality assessments, inquiring about the quality of individual elements of a manuscript, detected improvements where a global assessment of quality did not. Moreover, the reliability of the subjective score, even measured on the total sample, was lower (intraclass correlation coefficient = 0.12) than the instrument-based score or the score on many of the subscales (intraclass correlation coefficient = 0.25). Reviewers' opinions about specific aspects of a manuscript that need improvement may be more stable and more likely to result in consensus than global assessments. This is consistent with our experience at *Annals* and that of other editors (8, 30, 32) who report that the content of reviewer comments is more useful than summary recommendations for or against publication.

There are several limitations to this study. First, the manuscripts examined in the study were already highly selected, first by the author's decision to submit them to *Annals* and then by the journal's selection process, which accepts only 15% of submissions. Room for improvement in accepted manuscripts may have been small compared with rejected manuscripts. This possibility is supported by our data showing that the higher the initial quality of a manuscript, the smaller the subsequent improvement. Less selective journals may have the potential to improve research reporting even more than observed at *Annals*. Second, the relatively large editorial staff at *Annals* is not typical of any but the largest medical journals, and the generalization to others with different selection, review, and editing processes cannot be easily made. Third, the instrument we used reflected the perspectives of the same editors who engaged in the revision process. It may be that the instrument was particularly sensitive to the kind of changes our editing style would produce. However, based on the outside review of the instrument during its development, we believe a consensus exists among methodologists that the dimensions used in the questionnaire are important at most medical journals publishing original research.

Another problem was that the reliability of the instrument was low, although the total study sample suggested higher reliability (intraclass correlation coefficient = 0.25) than did the replication subsample (intraclass correlation coefficient = 0.12). Correlation on individual items and sections was somewhat higher but not dramatically so. In general, an intraclass correlation coefficient less than 0.40 is regarded as poor reliability (34). Because the pattern of item results (33 of 34 improving) and the effect of the categorical variable for the manuscript in the regression could not have plausibly occurred at random (P = 0.0005), we believe that the subsample results may have been because of an unfortunate play of chance. However, even with an intraclass correlation coefficient of 0.25 or somewhat higher, this instrument may be better suited for evaluating groups of manuscripts rather than individual studies, unless there are multiple raters, or unless user training can substantially increase reliability.

One area for which these results have implications is meta-analysis, or, more generally, the synthesis of evidence from many sources. A controversial issue in this area has been how to factor the quality of component research into the summary numbers (35, 36). We gave our assessors the freedom not to penalize a research report if the omission of an item would have had a negligible impact on the interpretation of the finding. The tone and strength of conclusions, the justifiability of generalizations, and the nature of limitations are qualitative measures of the strength of the scientific evidence that are not a simple function of design features. Our results raise the question of whether important dimensions of quality exist that are not captured by checklists that record the presence or absence of various features of an experiment.

These results indicate that most medical research reporting has substantial room for improvement and that peer review and editing can improve it in ways that are particularly important to readers, the media, and the lay public. A report, once published, is a permanent part of the medical literature; this study indicates that peer review and editing can make it more temperate and well balanced than if published as originally submitted.

Appendix: Manuscript Quality Assessment Instrument

The purpose of the questionnaire is to provide a structured assessment of the quality of a medical research report—*not the quality of the research itself.* Your main focus should be on making judgments as to whether the authors have described their research in enough detail and with sufficient clarity so a reader could make an independent judgment about the strengths and weaknesses of their data and conclusions. An imperfect study, which is well described and with its limitations well presented and discussed, can result in an excellent report. Conversely, a model experiment can be misunderstood because of a poor presentation. Please try to keep this distinction in mind as you review the enclosed paper.

As you will see, this quality assessment instrument is not a "checklist." It will require you to make many judgments. When in doubt, if an omitted detail would have a negligible impact on your interpretation (for example, the reason why 1 out of 150 persons dropped out), the manuscript should not be penalized. If, however, you find that the presentation makes it difficult to understand what was done, to follow the reasoning or to trust the conclusions, that should be reflected in your scoring. In other words, apply high but realistic standards—the same standards that you would want in the best medical journals and ones that would actually help you make clinical judgments. If a question is not applicable to the study you are evaluating, circle N/A.

Introduction

1. How clear are the background and rationale for this study?

   The frequency and severity of the clinical problem, what remains unknown about the research question, and how patients could benefit from the answer.

1	2	3	4	5
Not Clear		Somewhat clear		Clear

2. How clear are the specific aims of this study?

   The research questions (distinguishing main from secondary) and, if appropriate, hypotheses about what will be found.

1	2	3	4	5
Not Clear		Somewhat clear		Clear

Methods: *Subjects*

3. How adequate is the description of the setting of the study and source of subjects?

   To help readers understand whether the patients in the study are like theirs, the manuscript should provide information on when and where the research took place, a description of the level of care (community, primary care, referred), and if patients were referred, the pattern (source, distance, route) of referral.

1	2	3	4	5
Inadequate		Fair		Adequate

4. How clear are the eligibility (inclusion and exclusion) criteria?

1	2	3	4	5
Not Clear		Somewhat clear		Clear

5. For studies in which two groups are compared, is there enough information to judge the suitability of the comparison groups?

   How well was it reported how patients were chosen (for observational studies) or allocated (for experiments) so that readers can judge whether the researchers have compared like with like?

1	2	3	4	5
No information		Some information		All necessary information

Methods: *Design*

6. How clear is the study design?

   Do you understand what the authors set out to do and how they did it (the study design)?

1	2	3	4	5	N/A
Not Clear		Somewhat clear		Clear	

7. How adequate is the description of the masking (i.e. blinding) procedure?

   Is it clear who was blinded, what blinding procedure was used, and the degree to which blinding was achieved?

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

Methods: *Variable measurement*

8. Is the operational definition of major variables clear enough so their strengths and limitations can be assessed?

For example, in surveys, case definitions; in cohort studies, definitions for exposure and disease status; in diagnostic studies, the test procedure; for case control studies, is it clear how cases and controls were defined? Other major variables might include important confounders, compliance, etc.

1	2	3	4	5	N/A
Not Clear		Somewhat clear		Clear	

9. How adequate is the reporting of important side-effects?

For example, what are the types and numbers?

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

Results: *Subjects*

10. How complete is the information (reasons and numbers) on eligible subjects who were not included?

For example, subjects might not be included because they refused to participate, their records were lost, or they were not compliant during a run-in period. Is there enough information to judge, even in a general way, the comparability of the participants and non-participants in the study?

1	2	3	4	5	N/A
No information		Incomplete		Complete	

11. How adequate is the description of the enrolled sample, including potential cofounders, effect modifiers, co-interventions, comorbidities and spectrum of disease? (In comparative studies, this would mean description by group).

Is there a description (a table when necessary) of the characteristics of the enrolled sample, including potentially important demographic and prognostic factors or other descriptors that would help you to evaluate the comparability of the groups and/or the generalizability of the study results?

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

12. How clear are the outcomes for everyone enrolled in the study?

In addition to the main outcomes of the study, how well do the authors document the number of protocol violations, dropouts, crossovers, subjects with incomplete data, subjects who died for reasons other than the main reason under study, etc?

1	2	3	4	5	N/A
Not Clear		Somewhat clear		Clear	

Results: *Quantitative Reporting*

13. Are the quantitative methods the right ones for the research questions and data?

Are the methods appropriate for the unit of analysis (e.g., person, events, or clusters), sample size and type of outcome (e.g., dichotomous or continuous, time to event)?

1	2	3	4	5	N/A
Not right		Partly right		Right	

14. Are quantitative results reported in a manner that most of the intended audience could understand?

Consider whether units are clear (particularly of regression coefficients), whether the results are in the most accessible scale (e.g., non-logarithmic), and whether there should be additional effort to interpret technical results for the reader.

1	2	3	4	5	N/A
No		Possibly		Yes	

15. How adequate is the reporting of denominators?

For averages, percentages, rates, ratios, etc.

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

16. Are the magnitudes of effects reported?

"Effects" include odds ratios, risk differences, differences between means, regression coefficients, etc. (but not P values), and should be either stated directly or readily apparent from the data presented.

1	2	3	4	5	N/A
No	They are omitted in some important places.			Yes, whenever appropriate.	

17. In studies of diagnostic tests, how adequate is the reporting of summary statistics for test performance?

Summary statistics include sensitivity, specificity, predictive value, ROC curve, or likelihood ratio.

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

18. Are confidence intervals or standard errors reported for main outcomes?

If the main outcome is a difference between groups, or within patients, the statistical precision of that difference should be reported.

1	2	3	4	5	N/A
Never		They are omitted in some important places.		Whenever appropriate.	

19. How appropriate is the balance between detail and summary results?

1	2	3	4	5	N/A
Inappropriate		Fair		Appropriate	

20. How appropriately are dropouts, crossovers, or subjects with incomplete data dealt with in the analysis?

Techniques to deal with these problems include intention-to-treat analyses, comparison of these groups at baseline, analyses stratified by these factors, and sensitivity analyses.

1	2	3	4	5	N/A
Inappropriate		Fair		Appropriate	

21. How adequate is the method used to control or assess the effects of multiple measured variables?

If multiple variables are considered only singly, should joint effects be evaluated? Is a reasonable multivariate method chosen (e.g. stratification, adjustment, regression, ANOVA)? Does the variable coding permit adequate control?

1	2	3	4	5	N/A
Inappropriate		Fair		Appropriate	

22. How adequate is the reporting of analyses of multiple variables?

Are we told how the initial pool of possible predictors was chosen, how the final ones were selected, the coefficients or effects (in interpretable units) of all terms in the final model, the coding of each variable, and the number of subjects with each predictor or the spread of predictor variables?

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

23. Are clinically relevant subgroup effects explored in appropriate detail (neither too much nor too little)?

1	2	3	4	5	N/A
Inappropriate		Fair		Appropriate	

24. Do the figures and tables effectively summarize important data?

Include in your judgment whether tables and figures are accurate and clear, whether tabular data would be better presented graphically or vice-versa, and whether the balance between text and figures/tables is appropriate.

1	2	3	4	5	N/A
No		Somewhat		Yes	

Discussion and Conclusions

25. Is it clear what this study adds to the body of knowledge in its field?

1	2	3	4	5	N/A
Unclear		Somewhat clear		Clear	

26. How appropriate is the presentation of other supporting evidence that may be relevant to these conclusions (including theoretical reasoning, basic science results)?

An appropriate presentation would be neither too detailed nor deficient.

1	2	3	4	5	N/A
Inappropriate		Fair		Appropriate	

27. How appropriate is the discussion of study limitations?

An appropriate discussion would be neither too detailed nor deficient.

1	2	3	4	5	N/A
Inappropriate		Fair		Appropriate	

28. Is it clear if the authors are generalizing? If so are these generalizations justified?

For example, for different patients, interventions, follow-up times, outcomes, etc.?

1	2	3	4	5	N/A
There is no acknowledgment of the generalizations.		The generalizations are acknowledged, but not well justified.		Any generalizations are acknowledged and reasonably justified.	

29. Is the strength and/or tone of the conclusions appropriate to the design and results?

1	2	3	4	5	N/A
Inappropriate		Somewhat appropriate		Appropriate	

Title

30. How good is the title?

For example, clear, concise, and accurate?

1	2	3	4	5	N/A
Poor		Fair		Excellent	

Abstract

31. Does the abstract adequately summarize the data and conclusions?

1	2	3	4	5	N/A
Inadequate		Fair		Adequate	

General Evaluation

32. Is the manuscript concise?

1	2	3	4	5	N/A
No. (The text could be tightened by >25%.)		Somewhat. (About 10% to 15% could be cut.)		Yes	

33. How good is the organization of this report?

For example, are all methods in the methods section, all results in the results section?

1	2	3	4	5	N/A
Poor		Fair		Excellent	

34. How would you describe the style of the presentation?

1	2	3	4	5	N/A
Opaque		Workmanlike		Elegant	

Summary Scale

35. How would you describe the overall quality of this report?

1	2	3	4	5	6	7	8	9	10
Poor		Fair		Acceptable			Good		Superb

Presented in part at the Second International Congress on Peer Review in Biomedical Publication sponsored by the American Medical Association in Chicago, Illinois, 9 to 11 September 1993.

Acknowledgments: The following "expert assessors" are thanked for their contributions to this project: Wendy Adams, MD; Eric Bass, MD; Greta Bunin, PhD; Jeffrey Carson, MD; Neal Dawson, MD; John Dickenson, MD; Kay Dickersin, PhD; Jose Escarce, MD; Harold Feldman, MD; Daniel Ford, MD, MPH; Martha Gerrity, MD; Matthew Gillman, MD, PhD; Deborah Grady, MD; Alan Hillman, MD; Thomas Imperiale, MD; Sharon Inouye, MD; Jeane Grisso, MD; Walter Kernan, MD; Thomas Keyserling, MD; Michael Klag, MD; Kurt Kroenke, MD; Seth Landefeld, MD; Suzanne Landis, MD; Valerie Lawrence, MD; Richard Lofgren, MD; Peter Margolis, MD, PhD; Wayne McCormick, MD; Alfredo Morabia, MD, PhD; Ann Naftinger, MD; Roberta Ness, MD; Warren Newton, MD; Sandra Norman, PhD; Patrick O'Connor, MD; Michelle Petri, MD, MPH; John Philbrick, MD; Jennifer Pinto-Martin, PhD; Peter Pompei, MD; Bruce Psaty, MD; Jeffrey Silber, MD, PhD; Rebecca Silliman, MD; John Steiner, MD; Ramon Velez, MD; Paul Whelton, MD, MSc; Larry Wissow, MD, MPH.

The authors also thank Ms. Torii Ransome and Ms. Grace Lobb for their efforts in coordinating this project and Mr. Andrew Langman for assisting with the design of the questionnaire.

Requests for Reprints: Steven Goodman, MD, PhD, Johns Hopkins University, Oncology Center, Division of Biostatistics, 550 North Broadway, Suite 1103, Baltimore, MD 21205.

Current Author Addresses: Dr. Goodman: Johns Hopkins University, Oncology Center, Division of Biostatistics,
550 North Broadway, Suite 1103, Baltimore, MD 21205.
Dr. Berlin: University of Pennsylvania School of Medicine, Center for Clinical Epidemiology and Biostatistics, Philadelphia, PA 19104.
Drs. Fletcher and Fletcher: Harvard Medical School, Department of Ambulatory Care and Prevention, 126 Brookline Avenue, Suite 200, Boston, MA 02215.

References

1. Relman AS, Angell M. How good is peer review? [Editorial]. N Engl J Med. 1989;321:827-9.
2. Relman AS. Peer review in scientific journals— what good is it? West J Med. 1990;153:520-2.
3. Peters D, Ceci S. Peer-review practices of psychological journals: The fate of submitted articles, submitted again. Behavioral and Brain Sciences. 1982;5:187-255.
4. Bower B. Peer review under fire. Science News. 1991;(Jun 22):394-5.
5. Altman L. The myth of passing peer review. In: Bailar J, Angell M, Boots S, eds. Ethics and Policy in Scientific Publication. Bethesda, MD: Council of Biology Editors, Inc.; 1990.
6. Medical Journals: slowing flow of news on life saving discoveries? Washington Post. 1991;(Jan 25):A25.
7. Horrobin DF. The philosophical basis of peer review and the suppression of innovation. JAMA. 1990;263:1438-41.
8. Bailar JC 3d, Patterson K. The need for a research agenda. N Engl J Med. 1985;312:654-7.
9. Rennie D. Guarding the guardians: a conference on editorial peer review [Editorial]. JAMA. 1986;256:2391-2.
10. Rennie D. Editorial peer review in biomedical publication. The first international congress. [Editorial]. JAMA. 1990;263:1317.
11. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. Br Med J. 1986;292:810-2.
12. Chalmers TC, Smith HJ Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Controlled Clin Trials. 1981;2:31-49.
13. Meinert CL, Tonascia S, Higgins K. Content of reports on clinical trials: A critical review. Controlled Clin Trials. 1984;5:328-47.
14. Meinert CL, Tonascia S. Clinical Trials: Design, Conduct and Analysis. New York: Oxford University Press; 1986:1-469.
15. Mosteller F, Gilbert J, McPeek B. Reporting standards and research strategies for controlled clinical trials; Agenda for the Editor. Controlled Clin Trials. 1980;1:37-58.
16. Simon R, Wittes R. Methodologic guidelines for reports of clinical trials [Editorial]. Cancer Trt Reps. 1985;69:1-3.
17. Bailar JC 3d, Mosteller F. Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations. Ann Intern Med. 1988;108:266-73.
18. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med. 1987;317:426-32.
19. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. Br Med J. 1986;292:746-50.
20. Berry G. Statistical significance and confidence intervals [Editorial]. Med J Aust. 1986;144:618-9.
21. Braitman LE. Confidence intervals extract clinically useful information from data [Editorial]. Ann Intern Med. 1988;108:296-8.
22. Braitman LE. Confidence intervals assess both clinical significance and statistical significance [Editorial]. Ann Intern Med. 1991;114:515-7.
23. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med. 1993;118:201-10.
24. Soffer A. Can you believe what you read in medical journals? Chest. 1992;101:1417-9.
25. Cicchetti D. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences. 1991;14(1):119-35.
26. Bailar J, Angell M, Boots S, et al. Ethics and Policy in Scientific Publication. Bethesda, Maryland: Council of Biology Editors, Inc.; 1990.
27. Chubin DE, Hackett EJ. Peerless Science: Peer Review and US Science Policy. Albany: State University of New York; 1990.
28. Feinberg BA. Peer review and the NCI's clinical alert on node-negative breast cancer [Letter]. JAMA. 1989;261:695-6.
29. DeVita VT. Is a mechanism such as the NCI's Clinical Alert ever an appropriate alternative to peer review? In: DeVita VT, Hellman S, Rosenberg SA, eds. Important Advances in Oncology. Philadelphia: Lippincott; 1991:241-54.
30. Lock S. A difficult balance: editorial peer review in medicine. Philadelphia: ISI Press; 1986:1-172.
31. Gardner MJ, Bond J. An exploratory study of statistical assessment of papers published in the *British Medical Journal*. JAMA. 1990;263:1355-8.
32. Wilson J. Peer review and publication: Presidential address before the 70th annual meeting of the American Society for Clinical Investigation. San Francisco, California, 30 April 1978. J Clin Invest. 1978;61:1697-701.
33. Ernst E, Saradeth T, Resch KL. Drawbacks of peer review [Letter]. Nature. 1993;363:296.
34. Fleiss J. Statistical Methods for Rates and Proportions. Second edition. New York: Wiley; 1981.
35. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol. 1992;45(3):255-65.
36. Greenland S. A critical look at some popular meta-analytic methods. Am J Epidemiol. 1994; [In press].