This is to certify that the
dissertation entitled

AN EXAMINATION OF THE PEER REVIEW PROCESS IN A LARGE RESEARCH ORGANIZATION

presented by

RICHARD P. BANGHART

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in CEPSE

Major Professor's Signature

11/23/06

Date

AN EXAMINATION OF THE PEER REVIEW PROCESS OF A LARGE RESEARCH ORGANIZATION

By

Richard P. Banghart

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

2006

ABSTRACT

AN EXAMINATION OF THE PEER REVIEW PROCESS OF A LARGE RESEARCH ORGANIZATION

By

Richard P. Banghart

Peer review is the "gold standard" of knowledge, and is the process through which all scientific and scholarly publications pass. Research into peer review is difficult and rare as the process is normally performed in secret and privacy.

This study was granted access to a large data set containing over 35,000 individual reviews collected over three years capturing the peer review practice of a large scholarly research organization. Through examination of this large data set, answers to questions about the underlying assumptions of peer review are sought. Do reviewers agree with one another? Do decision makers adhere to reviewers' findings? Are these findings robust across the three years and the divisions of the research organization?

DEDICATION

I dedicate this work to my wife, Zara. For more years than should have been necessary, she provided the environment that allowed me to complete this work.

ACKNOWLEDGMENTS

I wish to acknowledge the support of my committee: Yong Zhao, Richard Houang, Bob Floden, and Ann Austin. Each has provided loving encouragement through the process. Many others have contributed their support and ideas. Mark Urban-Lurain spent many hours with me as I struggled with some of the statistical ideas. Ed Wolfe was generous with his time and knowledge.

All who helped deserve credit for any good ideas that may appear, but all errors are mine alone.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER 1 BACKGROUND/PROBLEM 1
The Problem 1
Significance of the Problem 2

CHAPTER 2 PREVIOUS STUDIES ON PEER REVIEW 5
Unpacking Peer Review 9
Definition of Peer Review 9
Definition of Peer 10
Peer Review's "Built In" Problems 12

CHAPTER 3 THE CURRENT STUDY 20
Methodology/Analytical Framework 20
Data Description 21
Authors and Proposals 22
Reviewers and Reviews 23
Divisions 24
From Data to Measurement 28
Analytic Techniques Addressing Reviewer Agreement .. 30
Analysis of Variance (ANOVA) and G Study 30
Perfect Agreement 31
Harsh/Lenient Rater 32
Complete Disagreement 33
Apples with Multiple Criteria Resulting in a Single Latent Trait 36
G Study 37
Monte Carlo Simulation 43
Marble Draw Example 44
Analytic Techniques Addressing Decision Making 46
Editorial Choice 47
Cut Score 50
Editorial Reach 53
Questions/Hypotheses 54

CHAPTER 4 FINDINGS 59
The Broad Picture 59
Reviewer Summaries by Division and Year 61
Proposal Summaries by Division and Year 63
Author and Reviewer Characteristics 71
The Criteria 74

G Study 81
Results of Monte Carlo Simulation 82
Editorial Choice 89
Editorial Reach 91
Reviewer Characteristics 92
Author Characteristics 92

CHAPTER 5 DISCUSSION 94
Answers to Questions 94
Reviewer Agreement 94
Efficiency 97
Differences among Divisions and Years 98
Decision Based on Score 99
Implications/Recommendations 99
Central Tendency 100
Halo Effect 101
Editorial Influence 102
Other Conceptions of Peer Review 102
Limitations 103
Softer Science 104
Proposals, Not Completed Works 104
Future Research 105
Text Analysis 105
Editorial Influence 105
Ultimate Publication 106
Conclusion 106

APPENDIX A Tables of Reviews per Proposal 108

APPENDIX B Figures Showing Accept and Reject by Division and Year 110

REFERENCES 128

LIST OF TABLES

Table 1 Summary of Proposals, Reviews and Reviewers by Year 27

Table 2 Perfect Agreement 32

Table 3 Harsh and Lenient Rating 33

Table 4 Complete Disagreement among Raters 34

Table 5 Variance Components for Rating with Harsh and Lenient Raters 43

Table 6 Example of Possible Proposal Scoring 49

Table 7 Summary of Proposals, Reviews and Reviewers by Year 60

Table 8 Proposals per Reviewer 62

Table 9 Proposals Submitted and Accepted 63

Table 10 Summary of Reviews per Proposal 65

Table 11 Reviewers per Division by Year 67

Table 12 Authors per Division by Year 68

Table 13 Cross Tabulation of Requested Format and Accepted Format 69

Table 14 Proposal Review Criteria Statements and Anchors 71

Table 15 Author Status Percentage across Divisions and Years 72

Table 16 Summary of Author Years 72

Table 17 Summary of Reviewer Status 73

Table 18 Summary of Reviewer Years 73

Table 19 Ratings Associated with Each Agreement Index ... 78

Table 20 Z Scores for Each Mean/Index 89

Table 21 Editorial Choice by Division and Year 90

Table 22 Editorial Reach across Divisions and through Years 91

Table A1 Reviews per Proposal by Division for 2001 108

Table A2 Reviews per Proposal by Division for 2002 109

Table A3 Reviews per Proposal by Division for 2003 109

LIST OF FIGURES

Figure 1. Accepted and rejected proposals. 76

Figure 2. Frequency of each possible variance of three scores. 79

Figure 3. Mean versus variance of accepted proposals. 80

Figure 4. Mean versus variance of rejected proposals 81

Figure 5. Median mean/variance outcome of unweighted Monte Carlo simulation. 84

Figure 6. Random, weighted mean/variance distribution. .. 85

Figure 7. Sample data mean and variance frequency. 86

Figure 8. Histogram of frequency of 1,1,1 in simulated data compared with actual. 87

Figure 9. Histogram of frequency of 3,3,4 in simulated data compared with actual. 88

Figure A1. Accept and reject, Division 1, Year 1 110

Figure A2. Accept and reject, Division 1, Year 2 110

Figure A3. Accept and reject, Division 1, Year 3 111

Figure A4. Accept and reject, Division 2, Year 1 111

Figure A5. Accept and reject, Division 2, Year 2 112

Figure A6. Accept and reject, Division 2, Year 3 112

Figure A7. Accept and reject, Division 3, Year 1 113

Figure A8. Accept and reject, Division 3, Year 2 113

Figure A9. Accept and reject, Division 3, Year 3 114

Figure A10 Accept and reject, Division 4, Year 1 114

Figure A11 Accept and reject, Division 4, Year 2 115

Figure A12 Accept and reject, Division 4, Year 3 115
Figure A13 Accept and reject, Division 5, Year 1 116
Figure A14 Accept and reject, Division 5, Year 2 116
Figure A15 Accept and reject, Division 5, Year 3 117
Figure A16 Accept and reject, Division 6, Year 1 117
Figure A17 Accept and reject, Division 6, Year 2 118
Figure A18 Accept and reject, Division 6, Year 3 118
Figure A19 Accept and reject, Division 7, Year 1 119
Figure A20 Accept and reject, Division 7, Year 2 119
Figure A21 Accept and reject, Division 7, Year 3 120
Figure A22 Accept and reject, Division 8, Year 1 120
Figure A23 Accept and reject, Division 8, Year 2 121
Figure A24 Accept and reject, Division 8, Year 3 121
Figure A25 Accept and reject, Division 9, Year 1 122
Figure A26 Accept and reject, Division 9, Year 2 122
Figure A27 Accept and reject, Division 9, Year 3 123
Figure A28 Accept and reject, Division 10, Year 1 123
Figure A29 Accept and reject, Division 10, Year 2 124
Figure A30 Accept and reject, Division 10, Year 3 124
Figure A31 Accept and reject, Division 11, Year 1 125
Figure A32 Accept and reject, Division 11, Year 2 125
Figure A33 Accept and reject, Division 11, Year 3 126
Figure A34 Accept and reject, Division 12, Year 1 126

Figure A35 Accept and reject, Division 12, Year 2 127

Figure A36 Accept and reject, Division 12, Year 3 127

CHAPTER 1
BACKGROUND/PROBLEM

The Problem

Peer review is the process through which "all important advances in science" (Rennie, Drummond et al. 1989) pass. It is the "primary institution responsible for processing and evaluating contributions to knowledge." (Lindsey, 1976). However, this "primary institution" has long been challenged for a number of faults. Problems of peer review include: prolonging the time to publication of important findings; expense in time and money and other resources; resistance to accepting innovative ideas; lack of civility in reviewer comments; bias toward accepting work that rejects the null hypothesis; bias toward accepting work from previously published authors (the Matthew effect); poor quality of reviews; theft of ideas by reviewers; and many others.

Weller analyzed more than 200 studies on peer reviewing from more than 300 journals. She affirmed, "Peer review's outstanding weakness is that error of judgment, either unintentional or intentional, are sometimes made. Asking someone to volunteer personal time evaluating the work of another, possibly a competitor, by its very nature

invites a host of potential problems, anywhere from holding a manuscript and not reviewing it to a careless review to fraudulent behavior." (Weller, 2002)

The fundamental issue about peer review is its worthwhileness. In other words, given the dependence that the scholarly and scientific communities have on the peer review process, and its expense in time, money and other resources, it is important to ask: **is the peer review process worthwhile?** The issue of the worthwhileness of peer review can be viewed from two angles: effectiveness and efficiency. Effectiveness is the degree to which the peer review process allows "good" knowledge claims to be differentiated from "bad" knowledge claims. That is, peer review's *effectiveness* is the degree to which we can truly trust it as an effective mechanism to advance knowledge. On the other hand, *efficiency* is the degree to which effectiveness is achieved at a minimum cost.

Significance of the Problem

Peer review is widely practiced. In all areas of science, (and, indeed, in other areas of scholarly pursuits) the very definition of acceptable knowledge includes its publication in a peer-reviewed journal. In virtually all areas of knowledge use and generation, peer

review is the tool invoked to guarantee the validity of the knowledge. In fact, as research methods change and paradigms rise and fall, one thing consistent since the introduction of the scientific method is the use of the peer-review process.

But peer review does much more than differentiate knowledge. Funding agencies use peer review to allocate research dollars. Especially since the rise of government funding of research after World War II, the use of peer review has been extended to determine not only what is to be published, but also what research is to be funded. The use of peer review to guide the allocation of resources for future funding adds a new function to peer review. Funding decisions are made in service to policy goals as well as scientific merit. While peer review was initially designed to ensure scientific quality, now the process is employed to ensure alignment with policy goals. This has had the result of causing the direction of research to be influenced by the peer review process in an even more direct way. (Chubin and Hackett, 1990)

Peer review affects the quality of knowledge and directions of new research. A report issued by the National Research Council recommends peer review as "the best available mechanism for identifying and supporting high-

quality research." The report goes on to say that beyond its role in promoting high quality research, peer review serves the "development of a culture of rigorous inquiry in the field." (National Research Council, 2004)

Therefore, it is important to better understand the effectiveness and efficiency of peer review.

This study sets out to do so by examining a large database of reviews. As such, this study is an opportunity to explore the effectiveness and efficiency of peer review as practiced in evaluating a large number of scholarly works. A large research organization has made available the electronic record of its peer review process in evaluating more than 10,000 scholarly submissions over three years. It is rare to have such data available for close analysis. Peer review is normally practiced in relatively small settings with secrecy and anonymity. The details of the findings of peers are generally not available for inspection. This large data set recording the peer review results of a large number of authors and reviewers can shed light on some of the assumptions of peer review.

CHAPTER 2
PREVIOUS STUDIES ON PEER REVIEW

Three hundred fifty years ago, when peer review was first employed, it arose out of a commitment to a new kind of rational examination of knowledge claims. Rather than simply accepting the knowledge claims of an author, organizations with an interest in advancing knowledge sought a means of ensuring the validity of the content of their publications. With the development of reproducibility and methodological rigor as standards for supporting one's findings, it became sensible for those knowledgeable about the scholarly and scientific process to judge the work of their peers. (Chubin and Hackett, 1990)

Experimental studies of peer review are rare. (Speck, 1993) The necessity for deception and the diversion of the resources of busy people make the justification of such research tenuous. Peters and Ceci (Peters and Ceci, 1982) provided the most highly cited experimental study of peer review, and exposed what they described as a failure of the peer review system. The researchers re-submitted papers to journals that had previously accepted the same papers within 18 months of their original publication. The papers had been slightly modified by changing author name, institutional affiliation, title, and small changes to the

abstract. These changes were made to prevent titles, authors and keywords from appearing in a cursory search of articles. They found that 8 of the 12 articles were rejected on resubmission. Peters and Ceci point out that rejection would be appropriate if reviewers had noted that the manuscripts were not original work, but the review comments indicated that the articles were rejected for "serious methodological flaws."

Observational studies of peer review, while more common than experimental studies, are also sparse and generally performed on a small scale. The nature of the peer review process makes even observational studies difficult. It is common for reviews to be submitted anonymously. Then the editor makes a decision based in part on the reviews, but also on other factors (e.g., balance of articles in an issue). Further, the process of peer review is predicated on the variability of the quality of manuscripts and the task of determining it. Because peer review is the method employed to determine that quality, it is impossible to know whether the decision of the peer review process is appropriate. From observational studies, the judgment of the quality of peer review can only be made on the basis of patterns of acceptance and rejection, and

cannot examine the method behind the decision making process.

More recently, there are some new studies concerning peer-review. The Journal of the American Medical Association has held periodic International Congresses on Peer Review since 1986, most recently in 2005. In the prior Congress in 2001, no findings were presented to indicate that peer review had any effect on the quality of published work, (Rennie, 2002) while other studies indicate that a variety of long understood problems continue to appear in published articles.

David Kaplan, a highly cited author, has said, "Despite its importance as the ultimate gatekeeper of scientific publication and funding, peer review is known to engender bias, incompetence, excessive expense, ineffectiveness, and corruption. A surfeit of publications has documented the deficiencies of this system." (Kaplan, 1995, p. 10)

In all the studies of peer review, there is none that examines a large quantity of data over an extended period of time. Most of the studies are on a relatively small scale, involving at most a few hundred reviews. Where larger numbers of manuscripts are examined, the researchers look for trends in acceptance patterns, but not at the

inner workings of the review process. That is, they are able to discover resulting biases in the outcome, but they are unable to employ the fine-grained techniques that can reveal the strengths and weaknesses about what is going on "inside" the peer review process at the level of the individual manuscript and individual review.

The most recent studies continue to examine publication bias, blind reviewing, statistical errors, ethical issues, and continue to find that the peer review process permits articles to pass through the process to publication with problems intact. In an editorial introducing the JAMA issue devoted to the Fourth International Congress on Peer Review, Drummond Rennie quotes himself from the 1986 Congress:

> One trouble is that despite this system, anyone who reads journals widely and critically is forced to realize that there are scarcely any bars to eventual publication. There seems to be no study too fragmented, no hypothesis too trivial, no literature citation too biased or too egotistical, no design too warped, no methodology too bungled, no presentation of results too inaccurate, too obscure, and too contradictory, no analysis too self-serving, no argument too circular, no conclusions too trifling or too unjustified, and no

> grammar and syntax too offensive for a paper to end up in print. (Rennie, 2002, p. 2759)

Rennie went on to say that 16 years later one can continue to find "in abundance" all the problems he spoke of in 1986.

Unpacking Peer Review

Definition of Peer Review

The term *peer review* (as used in this study) refers to the process by which scientific and scholarly work is deemed acceptable for funding or publication. While the process of *funding* peer review is somewhat different from *publication* peer review, the two share some essential features. In both cases, the work is submitted to a number of people who are thought to be knowledgeable in the field of research being reviewed. The work is given to the reviewers with the author and institutional affiliation removed. That is, the review is said to be "blind" where the reviewers are intended not to know whose work they are reviewing. The reviewers submit their review of the work to the publisher or funding agency, and from those reviews a decision about whether to accept or reject is made. The author is often given some indication of the reviews. Sometimes the actual reviews are provided, while other

times the publisher or funding agency may summarize the content of the reviews. Most often, the names of the reviewers are withheld from the author.

Publication and grant peer review are similar also in their allocation of scarce resources. While each kind of peer review seeks to assure that published and funded works are of sufficiently high scientific merit, each also may reject good work because of a lack of funds or publication space.

Funding peer review and publication peer review differ in some ways. In funding peer review the outcome is usually acceptance or rejection, while in publication peer review the outcome often includes encouragement to resubmit the work with changes that are suggested by reviewers (or editor). The result is that in publication peer review the process often serves to help an author develop a paper. In this way peer review actually serves to guide authors in shaping their research and publication.

Definition of Peer

The word "peer" typically means one who is of similar capability and rank, as used in the phrase "a jury of one's peers." In the case of publication peer review, the word takes on an additional sense of "one having particular

expertise" enabling an informed decision about the knowledge claims being made in a manuscript. In a sense, "peer" can be defined (almost tautologically) as one who has the ability to reliably discern the qualities required of good scientific work.

Scientists are expected to participate as reviewers. As part of the culture of science, the fact that all published works are peer reviewed leads all to understand the necessity of competent reviewers. Scientists contribute their time and energy in providing this function--often with no formal training, minimal guidelines, no compensation and very little oversight.

In this study the concept of "necessary and sufficient" becomes important in several circumstances. To accomplish peer review effectively and efficiently there are a number of things that may be necessary, but yet might not be sufficient. The implication of this idea is that necessary things, where lacking, are certain to prevent the process from being effective. But even where necessary things are present, those may not be sufficient to produce the desired result. Flour is a necessary, but not sufficient, ingredient for making bread. Flour, salt, sugar, water, shortening, and yeast are both necessary and sufficient ingredients for making bread.

Peer Review's "Built In" Problems

Even if peer review were perfectly *effective* (allowing the publication of all good works, and preventing the publication of all bad works), and perfectly *efficient* (performing at the lowest possible expense) the process is subject to criticism for some of its inevitable consequences, and some of its traditional implementations.

Even at its most efficient, peer review is expensive. Historically, manuscripts were copied and delivered to reviewers who prepared reviews that were copied and delivered back to the journal or funding agency. Even in today's world of electronic communication, considerable resources are expended in support of the peer review process. Many hours of time are given by some of the most capable scientists and scholars in preparing reviews.

In addition to its expense, peer review violates one of the precepts of scientific and scholarly work--transparency. While research and the presentation of research findings are assumed to be open to inspection and replication, peer review is most often done "blindly." That is, reviewers read the work of authors without knowing their name or institutional affiliation, and authors never discover who provided the reviews for their work. It is

ironic that the process that is intended to guarantee openness and transparency is closed and opaque.

Another negative side effect of the peer review process is the delay it causes in bringing new information to the public. Peer review takes weeks or months (and sometimes years) to accomplish. This results in considerable delay before the research can reach publication and be made available to the public. For the knowledge consumer, peer reviewed publication is required before practice can be informed by research findings. "Prudent patient care demands that ultimate judgment await submission of a formal paper and the obligatory process of editorial peer review." (Soffer, 1980) It's difficult to know how lives may be affected by the delay of knowledge reaching those who need it.

While the intent is that reviewers do not know the authors of the work they are reviewing, the reality is that (especially in highly specialized fields) reviewers are frequently able to infer the author (or the institutional affiliation) owing to the fact that so few people are engaged in such research. In those cases, not only is the "blindness" of the review compromised, but also the highly specialized work of a competitor is revealed to those most capable of taking advantage of the knowledge. Because the

author of the work does not know who the reviewer is, it can be difficult for the author to protect "proprietary" information that may be revealed to the reviewer.

While there are many problems associated with peer review, and many questions that deserve answers, this study will focus on a narrow set of questions whose answers might possibly be gleaned from the analysis of a large data set. The questions have to do with the mechanics of translating numeric responses of reviewers into decisions of acceptance or rejection. They are about the effectiveness and efficiency of the process as a measurement task.

To determine the effectiveness and efficiency of the peer review process, its underlying assumptions must be examined. Peer review makes two related assumptions: (a) Manuscripts vary in their quality (i.e., some manuscripts are worth accepting and some are worth rejecting); and (b) "Peers" share knowledge of the criteria used in judging a manuscript, and can discern the extent to which a manuscript adheres to those criteria.

The first assumption can reasonably be accepted on its face (no one suggests that all manuscripts deserve to be published), while the second assumption is much more complicated and requires closer examination. The second

assumption has within it three words that need to be more fully fleshed out: "peer", "criteria", and "discern."

The criteria that peers are presumed to share knowledge of are the defining characteristics of good scholarly or scientific work. Specifically they include: originality, importance, methodology, analysis, and writing. (Wolff, 1970; Wilson, 1979; Armstrong, 1982) These criteria are broad and overlapping, and subject to a variety of interpretations. Because of their breadth, they are often expressed as sets of narrower characteristics each of which is a part of a larger criterion. For example, methodology is sometimes examined as a single construct, and sometimes viewed as the combination of data selection, study design and choice of statistical procedures. At other times, statistical procedures are thought to be part of the analysis criterion.

Like the review process, the criteria are rarely questioned. The guidelines that journals offer to reviewers embody these criteria. Surveys of most important criteria for the acceptance of manuscripts universally refer to five general criteria. The number one criterion for a publishable manuscript is its originality. (Hackett and Chubin, 2003) The other criteria are generally understood to include writing, methodology, theory, data, analysis,

conclusions, topic choice, and contribution to the field. Each of these qualities is a construct of considerable complexity, increasing the likelihood that reviewers will have difficulty in reliably discerning the degree to which a manuscript manifests these qualities.

Identifying multiple criteria as necessary in scientific work implies that these criteria exist in some way independently of one another. Those involved in peer review recognize multiple criteria as contributing to the quality of a paper. While originality is considered the main criterion for acceptance, papers must also be well written, the data involved must be collected properly, and the statistics employed must be properly chosen and interpreted. There is no reason to think that these criteria are inherently related to one another. That is, an author might have collected data well, but that is no reason to assume the writing is good.

Some of these criteria may be more important than others. For example, good data collection and analysis might trump punctuation. The knowledge community might be served by the publication of a work that has excellent theoretical and research technique, but includes some poorly phrased sentences. On the other hand, beautifully crafted prose may not redeem a work with invalid

statistical procedures producing questionable answers to hypotheses. An effective and efficient peer review system will account for such differences among the criteria.

To "discern" in this context takes on the meaning conventionally assigned to a measurement task. That is, the job of the reviewer is to measure the degree to which the qualities that are thought to exist in some quantity within each manuscript are present. This also implies that the reviewers are making a shared decision about these inherent qualities. That is, that they have the same understanding of what the qualities are and how they can be expressed in the manuscript. If peer review is to be effective and efficient, it is reasonable to expect both that there will be some differences among reviewers (otherwise there is no need for multiple reviewers), and that they have general agreement among themselves (otherwise there is no confidence they are responding to the qualities that are found in the manuscript being reviewed).

A problem with the concept of a peer as used in scientific peer review is that in the case of a truly innovative idea there may be no peers. This can happen in the case of breakthrough kinds of research where the findings challenge the current paradigm. The first researcher who challenges conventional wisdom faces the

flywheel effect of the peer review process. Effective and efficient peer review needs a way of recognizing novel developments in a field, or it runs the risk of halting progress in that area of research.

A related issue is the difference that is seen between the hard sciences (e.g., physics, astronomy) and the "soft" sciences (e.g., psychology, economics). The hard sciences have clearly defined paradigms. As such, authors and reviewers have clear ideas of what counts as good research and knowledge claims, there is high agreement among reviewers, and the journals have high acceptance rates (authors know what it takes to be accepted). In the soft sciences, the paradigms are not as firmly established: There is much more variability and less agreement among reviewers and authors. As a result, the acceptance rate is much lower in the journals of the soft sciences. (Newman, 1966; Zuckerman and Merton, 1971; Adair, 1982)

The most obvious assumption behind peer review is that the quality of submitted work varies, i.e., some work deserves to be published while other work deserves to be rejected. The use of multiple reviewers implies a couple of assumptions. One is that individual reviewers might make errors, but that having multiple reviewers will reveal the idiosyncratic judging of an individual. Another assumption

is that reviewers both know what is required of an acceptable work, and can detect those qualities within a work.

While some suggest that reviewers are selected with a variety of skills and abilities, and are to review that aspect of the work that relates to their area of expertise, the truth is that reviewers assigned to evaluate a work are all given the same instructions, and asked to rate a work on all of the criteria. Because of this, we should expect general agreement among reviewers.

CHAPTER 3
THE CURRENT STUDY

Methodology/Analytical Framework

Many problems with peer review have been identified, and its critical role in the scientific enterprise has been established. Many of the problems of peer review are beyond the scope of this study, however it is possible to focus on an area of importance that is very difficult to address simply because of the paucity of data normally available. By using the raw data of the many thousands of reviewers' ratings and the decisions made, along with information about the reviewers and the authors, the efficiency and effectiveness of the process as a measurement task can be evaluated.

These issues have been difficult to address because of the lack of data about the responses of individual reviewers across a large number of items being reviewed. This study takes advantage of a large body of data that was collected over three years by a large research organization's electronic peer review system. The web-based system mediated the assignment of reviewers to proposals that were submitted to the research organization's annual

meeting. The system also collected and stored the reviewers' responses.

The peer review process as practiced by this large research organization has some of the characteristics of both funding and publication peer review. Like funding peer review, this process resulted in acceptance or rejection with no opportunity of resubmitting the proposal with modifications suggested by reviewers. But like publication peer review, the work submitted was justified and funded elsewhere, so the process was unable to direct research energies to the funding targets (the goals of the funding agency). This peer review process allocated the limited resource of a presentation venue.

Data Description

This section describes the broad outlines of the nature, extent and limitations of the data collected by the organization during their peer review process. First, information that describes characteristics of the **proposals** submitted, the **authors** of the proposals, the **reviewers** and the **reviews** will be discussed. This is followed with an explanation of how all of these elements are contained within separate **divisions** of the organization, and repeated for each of three years.

Authors and Proposals

Authors submitted proposals to the research organization in hopes of being accepted to present their work at the organization's annual meeting. Proposals were submitted electronically to the organization through the web-based application. Authors intending to submit a proposal visited the website, and registered on the system. The registration process involved filling in forms on a web page that collected contact information and information about the authors' institutional affiliation, professional status, and years of membership in the organization. After the author's information was entered into the system, the author then submitted the written proposal in another form on the website, or uploaded a document.

Additionally, authors submitted their proposals to be presented in one of several preferred "requested formats." The formats are ways in which the work is presented to the membership of the organization. The most prestigious format is the "paper presentation," and accounted for more than 70% of the format requests by authors. Other formats (in order of decreasing prestige) included the "round table" (or "paper discussion"), "poster," and "new member poster." There are a variety of other formats available, but they are rarely requested.

To summarize, each proposal has the following attributes associated with it: Author Status (Professor, Assistant Professor, Associate Professor, Graduate Student, or Other), Author Years of Membership (<1, 1-5, 6-15, 16-25, >25), Division, Requested Format, Decision (accepted or rejected), and Accepted Format (if accepted).

Reviewers and Reviews

Like the authors, reviewers also registered with the web-application, and completed the on-line forms. Each reviewer selected a division, and gave information about his or her professional status and years of organization membership. Each reviewer also entered a brief biographical description about his or her interests and expertise.

After registering with the web-based system, reviewers were contacted by e-mail and informed when proposals had been assigned to them to be rated. The reviewers then visited the web-based application, logged in to the system, and there were able to read the proposals and respond to ten rating criteria. Their responses consisted of clicking on one of five "radio buttons," thereby indicating a choice of 1 to 5 for each of the ten criteria. A comments field was also offered for the reviewers to type extended text comments.

Each review was stored electronically and identified by a reviewer identification number (revid), and a proposal identification number (propid). The review consisted of answers to 10 "questions" (q1 - q10) each of which could hold a value of one to five, along with text comments. The system did not compel a response for each criterion, so criteria that received no response were coded in the system as a 0, but not included in any summary statistics.

Divisions

Each proposal was submitted to one of 12 divisions of the organization. The divisions represent areas of interest within the larger organization. Each reviewer also selected a division for which to review. The divisions of the organization were individually responsible for recruiting reviewers, managing the assignment of proposals to reviewers and making the decisions about acceptance or rejection. Divisions were also able to customize the criteria to allow them to conform to conventions of their areas of interest. Those divisions that chose to exercise that option generally chose only to re-order the criteria. One division for the second and third years of the data substantially modified the criteria.

The "chair" of the division was charged with making decisions about acceptance or rejection of proposals, and selecting a presentation format if the proposal was accepted. The web-based system provided several methods of decision support. The chair was able to view summary statistics about reviews given across the division as well as within a given proposal. Specific criteria could be selected to be included in these summary displays.

The unique circumstances of this form of peer review process enable a rare look at not only the outcomes, but also the individual findings of each of the reviewers. This, combined with information about the author and the reviewers, allows a fine-grained analysis that may shed light on the peer review process.

For analysis in this study, data were extracted from the on-line database that served the research organization's web-based system, and transferred into a local database. To preserve anonymity, all identifying information about individuals was removed, and user id numbers were assigned having no relation to the original identification numbers or names.

The data collected through the system includes information about the author, the reviewer, and about the proposal itself. This permits an analysis that takes into

account more than just the acceptance or rejection of a proposal. Both authors and reviewers are identified by their institutional affiliation, their professional status, and their years of membership in the organization. The proposals are submitted in two categories, and in a number of different modes of submission that vary in status. The combination of these factors permits the testing of some earlier suggestions about peer review.

In addition to the individual differences, the divisions serve as a broader classification of the people and works involved. Each of the 12 divisions within the research organization is involved in researching a different area of the larger field. Some of the divisions are involved in researching issues of policy, others focus on legal issues, some are involved in research methodology and measurement and statistics, others explore social and psychological issues, while some are involved in research around practice in the field. Across the divisions, there is a wide range of what counts as appropriate research questions, research methodology, and standards for making truth claims. Yet each division uses the same peer review method to determine which of the submitted proposals will be accepted for presentation.

The decision of acceptance or rejection was based on the answers each reviewer gave to each of 10 questions (the 10 criteria). For each criterion the reviewers offered a response of 1, 2, 3, 4 or 5 (or no response). This permits possible mean scores between 1 and 5. The data were maintained without any reduction. That is, each of the responses is available for analysis.

The resulting data consist of a total of nearly 34,000 reviews. Table 1 summarizes the data by year, number of proposals, reviewers and reviews.

Table 1

Summary of Proposals, Reviews and Reviewers by Year

Year	Reviews	Reviewers	Proposals
2001	10248	1860	3206
2002	12310	2641	3900
2003	11382	2179	3969

Three sets of data were available for three consecutive years (2001, 2002 and 2003). While there are some differences among the data sets owing to the continued development and improvement of the web-based system, all three data sets contain information about the proposal (ratings received, decision made, author information), the

reviewers (ratings given, years in research organization, professional status), and the criteria (scores given, anchors, prompts).

As mentioned, this large collection of data permits a level of analysis that is extremely rare (and perhaps unprecedented) in peer review. Through the analysis of these data it is possible to conceive of the process of peer review as an attempt at measuring the quality of the submitted work.

From Data to Measurement

Measurement is a process engaged in so frequently that one often fails to remember that it is rare to directly measure the quality of interest. For example, in measuring the temperature, one may instead look at the level of mercury in a thermometer. We trust that the mercury responds to the heat energy in the environment. When measuring the quality of scientific work, we look instead at the answers reviewers have given to a series of questions. The responses to the questions are trusted to reveal an underlying quality inherent in the work. Because there is no direct access to the quality being measured, it is said that the quality is a "latent trait." Like a latent fingerprint, or a latent photographic image, the latent trait exists unseen, until revealed through the use of

other materials, techniques, or analysis. (Bond and Fox, 2001).

Through the peer review process, the research organization is seeking to make a determination about whether to accept or reject proposals. The process of seeking numeric scores from multiple reviewers results in a representation of the underlying quality they are seeking to measure. In this way the construct of acceptability (what will be called the acceptability quotient, or 'AQ') is operationalized through the reviewers' responses to the questions.

The task in analyzing these data is to understand how the multiple criteria combine to create a single decision of acceptance or rejection. It is tempting to think that because a decision is reached about acceptance or rejection through a consideration of the scores given by each reviewer for each criterion, that these scores contribute equally (or at least separately) to the decision. However, other possibilities can be explored through the use of a variety of analytic techniques.

Two facts must obtain to have a legitimate measurement process to determine the acceptability of proposals. The first requirement is that reviewers agree with one another. The second requirement is that the decision maker abides by

the scoring of the reviewers. Those two requirements guide the analysis of the data in determining the effectiveness of the peer review process in the large research organization.

Analytic Techniques Addressing Reviewer Agreement

Analysis of Variance (ANOVA) and G Study

The reliability and efficiency of the peer review process depend on the degree to which peers know and can detect the qualities of a manuscript that earn that manuscript acceptance or rejection. Further, because the process involves multiple reviewers for each item being evaluated, the process assumes that the several reviewers share a common knowledge of the qualities, and a common ability to discern those qualities. The goal of this study is to test those assumptions.

Before discussing the specifics of the data under analysis in this study, consider the general ideas behind what kind of evidence would support a conclusion that measurements are reliable. Imagine an exercise in measuring apples. Reviewers are assigned to assess apples for size. The reviewers are given a collection of apples that are assumed to range from small to large. Each apple will be

assessed by at least three reviewers, each of whom will judge the apple as small (1), medium (2) or large (3). Further, each reviewer will rate at least three apples. At the end of the measurement process, there will be a set of data that can be looked at from two perspectives: one consisting of the ratings given by each reviewer; the other consisting of the ratings received by each apple. Because each apple has received at least three ratings, each apple's score can be expressed as a mean and variance. If the reviewers were perfectly consistent, each apple would have a variance of zero, and a mean of exactly one, two or three.

Perfect Agreement

Table 2 shows the results of a hypothetical apple measuring activity, where apple reviewers are in perfect agreement. Each of the rows (a, b and c) contains the scores given to each apple. The reviewers are indicated by columns x, y and z. Intuitively, there is confidence in scores like this. The fact that the reviewers are all in agreement with one another leads to a belief that they are making their decisions based on common criteria, and lends confidence that the mean for each apple reliably represents its size. In this ideal situation, the variance for each

apple is zero, resulting in a mean variance across all apples of zero. The variance for each reviewer's scores (across the apples) is greater than the variance of scores within each apple. The ratio of the mean variance within apples to the mean variance across apples is shown in the lower right corner of Table 2. Where there is perfect agreement among reviewers, this ratio is always zero.

Table 2

Perfect Agreement

	X	Y	Z	Mean	Var	
A	1	1	1	1	0	
B	2	2	2	2	0	0
C	3	3	3	3	0	
Mean	2	2	2			
Var	0.67	0.67	0.67			
		0.67			Ratio	0

Harsh/Lenient Rater

Table 3 shows a situation where there is not perfect agreement among reviewers. In this case reviewer x tends to judge apples as larger, while reviewer z judges apples as smaller (reviewer x might be called a lenient reviewer, while z is a harsh reviewer). Again, there is still

intuitive sense in this scoring pattern, even though there is disagreement among the reviewers. The mean values for each apple can be used in understanding its size, even though only one apple achieves a score of exactly 1, 2 or 3. But the disagreement among reviewers shows up in the ratio figure, as it is now greater than zero. In this case there is still more variance between apples than within apples, but as the ratio approaches one, confidence in the reliability of the reviewers' scores declines.

Table 3

Harsh and Lenient Rating

	X	Y	Z	Mean	Var	
A	2	1	1	1.33	0.22	
B	3	2	2	2.33	0.67	0.33
C	1	1	1	1.00	0.22	
D	3	3	2	2.67	0.22	
Mean	2.25	1.75	1.50			
Var	0.25	0.69	0.19			
		0.375			Ratio	0.89

Complete Disagreement

Table 4 shows a scoring situation with no agreement among reviewers. Each apple receives a mean score of two, and a variance resulting from the complete lack of

agreement. The variance across the scores given by each reviewer shows less variance than the variance of the scores received by each apple. The resulting ratio is greater than one. It is impossible to infer an apple's size with such data, and there is no confidence in the reliability of the scores achieved by each apple.

Table 4

Complete Disagreement among Raters

	X	Y	Z	Mean	Var	
A	3	2	1	2	0.67	
B	2	1	3	2	0.67	0.67
C	2	3	1	2	0.67	
D	1	3	2	2	0.67	
Mean	2	2.25	1.75			
Var	0.5	0.69	0.69			
		0.63			Ratio	1.07

To take the example further, imagine a situation where reviewer x always rates apples as a 1, reviewer y always rates apples as a 2 and reviewer z rates apples as a 3. In such a situation, all of the variance resides across reviewers, and there is no variance across apples. The implication of such a measurement result is that

measurement is completely determined by the rater and is in no way affected by the item being rated.

The important idea behind Table 2, Table 3, and Table 4 is that the ratio of mean variances gives an indication of the reliability of the scores received. If the ratio is one or greater, there is no basis for believing that the scores achieved have any relationship to the quality being measured, (i.e., the scores have more to do with the reviewers than with the apple). Ratio values near zero give greater confidence in the reliability of the scores, although that alone is not sufficient to be confident in the quality of the measurement. A ratio less than one is necessary, but not sufficient in establishing confidence in the reliability of measurement.

Exactly what the ratio of variances should be to have confidence in a decision is traditionally captured in the F statistic. The F statistic is based on the ratio of variances along with the number of factors and the values they may assume. With that information, along with assuming that the data involved conform to a normal distribution, a reliable inference can be made about the likelihood that the two groups' differences are greater than would be expected by chance alone.

Apples with Multiple Criteria Resulting in a Single Latent Trait

The apple-measuring example can be expanded to approach the kind of information that is in the research organization's data. In the earlier example, each apple received three scores, and each rater measured three apples. In the research organization, each rater offers 10 scores. With the addition of multiple judging criteria comes the possibility for greater precision in determining the underlying trait. In the earlier example the trait being measured was size. This can be extended to a more complex example by assuming measurement of an apple's marketability. In this case reviewers could be asked to judge apples on several criteria that combined will determine how likely it is to sell an apple in a store.

The criteria might include size, color, shape, firmness, tartness, sweetness, lack of blemishes, etc. In combining these criteria to determine an underlying trait of, say, "marketability" some careful thinking will be required. It is likely that apples of more than one color will be sold. So the color criterion will have to be expressed as something like "color appropriate to variety." Additionally, some criteria might be relatively more or less difficult for an apple to achieve. All apples might be

required to be of a minimum size. An apple being larger than that size may not increase its "marketability." In that case there is a threshold criterion where a score below a certain level causes rejection, and above permits selection. But the likelihood of selection is not increased as the score rises above the threshold.

The measurement task is complicated as the number of criteria and number of raters increases. When the measurement task takes into account multiple criteria from multiple raters, additional analytic techniques are useful in disentangling the multiple possible interactions among the criteria.

G Study

Generalizability theory offers a conceptual framework and a series of statistical techniques to more completely explore multiple contributors to an item's score. As explicated in the previous examples, in a measurement situation the rater and the rater's tendency (harshness or lenience) in scoring items must be taken into account. Similarly, where multiple criteria are used, differences among the criteria in their functioning must also be taken into account.

In a measurement task, generalizability theory regards the reviewers and the criteria as "facets" of the measurement. A facet is a variable that can contribute to the score achieved in a measurement. In the case of reviewers and proposals, the reviewer is a facet, as is the proposal. Similarly, each criterion is considered a facet as well. In generalizability theory, the item being measured is not usually referred to as a facet, but this is only a matter of terminology. The computations employed in a G study make no such distinction and treat the item being measured as a facet.

A generalizability study (G study) seeks to quantify the effect each facet has on the score achieved by the object under measurement. The G study does this through the variance components found in the measurement. That is, there is a comparison of variance across proposals (which is construed as contributing to the "true" measure), with the variance across reviewers (which are seen as contributing to "error"). Additional error is also indicated through an interaction of proposal and reviewer.

In an ideal measurement situation, all of the variance occurs across proposals. The assumption under such circumstances is that the differences in scores achieved by the proposals are a result of actual differences in

proposals, as opposed to differences in the reviewers. In a real measurement situation, there is some variation attributable to reviewers.

This variation in reviewers' ratings can happen owing to several factors. Certain reviewers might simply be less precise in their ratings, resulting in random variation around the proposal's "true" score. That is, a reviewer in evaluating three proposals of equal quality, might give one a 3, one a 4, and one a 5. Another possibility is that a reviewer might tend to give all proposals a similar score. Finally, a reviewer might be harsh or lenient, consistently rating proposals lower or higher than the proposal's "true" score.

These tendencies for reviewers to deviate from a proposal's "true" score is captured in the reviewer facet variance component. In a given measurement situation there is a certain amount of overall variation in the scores. In the process of performing a G study the overall variance is said to be "partitioned." Portions of the overall variance are assigned to each of the facets. For that reason, as the variance increases for one facet, it decreases in other facets. Ultimately, it is the ratio of variance occurring in the item under measurement with the variance occurring

in other facets that allows an assessment of the quality of a measurement.

The information obtained through a G study permits a comparison between the variance components of the various facets of measurement. This comparison gives a sense of the distribution (or partitioning) of the overall variance to the other sources of variance. Table 2 can be used to demonstrate the way a G-study is applied to such a situation. Recall that the ANOVA involves the mean variance across apples and the mean variance across raters. The ratio of those means is the measure of the degree to which there is confidence that the raters are responding to traits in the apple as opposed to their own idiosyncrasies.

The G-study is said to "partition" the variance among the facets of a measurement. The first step in performing a G-study involves estimating the total variance. The total variance is often called the "sum of squares." "Sum of squares" in this case actually means "the sum of the square of the difference between the grand mean and an individual score." ($\sum(\bar{X}-X)^2$, where $\bar{X}$ is the grand mean of all responses and X is an individual response.) In the simple example in Table 2, the grand mean of all scores is 2. Calculating the sum of squares is simple. For apple b there

is no contribution to the sum of squares, as each of its scores is equal to the grand mean ($\overline{X} - X = 0$). The difference between the scores on apples a and c, and the grand mean are all 1. 1 squared is one, so the sum of squares is 6. This captures the total variance in this example.

The second step is to calculate the variance for apples and reviewers (in the terminology of generalizability theory, this is called "estimating the variance" of the individual facets). Again the calculations are simple. For apples a, b and c the mean scores are 1, 2, and 3 respectively. For reviewers, each has a mean score across apples of 2. The sum of squares across the apples is 0, while the sum of squares multiplied by the number of apples is 6. Recall that the total sum of squares was 6. In this case all of the variance is accounted for in the apples, and there is no variance attributable to the reviewers.

The G-study proceeds to account for degrees of freedom, and converts the sum of squares to mean squares by dividing the sum of squares by the degrees of freedom. Finally, the variance component is calculated for each of the facets and their interaction. For the interaction effect the variance is simply the mean square calculated in the previous step (i.e., the variance divided by the

degrees of freedom). To calculate the "apple variance" the interaction mean square is subtracted from the proposal mean square and divided by the number of reviewers. In a parallel fashion, the variance for reviewers is calculated as the interaction effect subtracted from the reviewer mean square and the result divided by the number of apples. (Brennan, 2001)

When this procedure is performed using the values in the example in Table 2, a value of one is calculated for what is termed the proposal effect, and a value of 0 for both reviewer effect and interaction effect. It is immediately apparent that all of the variance is accounted for in the apples, and reviewers contribute nothing to the variance. This is an ideal situation and leads to confidence in the scoring.

When a similar exercise is done using the numbers in Table 3 there is some variance attributable to reviewers. In this case, one reviewer tends to be harsh, while another reviewer tends to be lenient. When the calculations are performed, the results are as shown in Table 5. We can see that while there is variance attributable to both the reviewer and the interaction of reviewer and apple, the majority of the variance is still attributable to the apple.

Table 5.

Variance Components for Rating with Harsh and Lenient Raters

Effect	Var. Component
Apple	.5833
Reviewer	.1733
Interaction	.1389

Monte Carlo Simulation

While the ratio of variance is an indicator of quality of peer-review, it cannot be used directly to speak to the worthwhileness of peer-review. This is because the specific ratio found in the data from the research organization lacks a standard against which to compare the ratio. The question is: What is a good ratio? It is clear from the apples example that some disagreement among reviewers can occur without compromising confidence in the outcome. But how much disagreement can be tolerated?

When a statistical technique is employed to analyze data, it essentially addresses the question: What is the likelihood that these results occurred by chance? Because of the many complicating factors around the analysis of this particular data set, it is very difficult to establish a priori what ratios should be expected to provide confidence that the measurement process has yielded results

different from chance. A method used to explore such situations is the Monte Carlo simulation.

A Monte Carlo simulation relies upon the power of modern computers to rapidly generate large numbers of random data sets. Each data set can be described with some statistical measures, and the group of data sets can be described with the resulting collection of statistical measures. That collection of statistical measures can, in turn, be described with statistical measures that can be used to help understand what the range of values might reasonably be. That distribution of statistical measures allows a determination of what might be described as the true population statistics. When the statistic of interest in the sample data set is compared with the distribution of that statistic from the thousands of randomly generated data sets, an estimate can be made of the likelihood that the sample statistic was the result of chance.

Marble Draw Example

An example may make this clearer. Suppose there is a class of 30 students. Each student is asked to draw one marble four times from a bag containing 10 black and 10 white marbles, replacing the drawn marble after each draw. After all students have performed this exercise it is

discovered that 3 of the 30 students drew 4 black marbles. We would like to know how likely it is for that to occur. We can calculate the likelihood assuming that each color has a 50/50 chance of being selected. (.5 * .5 * .5 * .5 = .0625) meaning that 625 times out of 10000 there is the expectation of drawing four black marbles. Using a Monte Carlo simulation, the computer can simulate drawing four marbles 30 times, and count the number of times that result in four black marbles being drawn. The computer can repeat the simulated drawing of marbles many times, and count the number of draws that result in all black marbles. Eventually, this will result in a collection of counts that reveals what the range of the frequency is of drawing four black marbles.

In the case of the peer review data contained in the research organization data set, it is much more difficult to calculate theoretical random values. Conventions of scoring may not equally weight each of the scores. That is, it may be that the scores of 1-5 are not equally assigned. In this case it would be inappropriate to generate random scores where all scores have equal likelihood. Instead, the random assignment should be "weighted" in such a way that the actual distribution of scores is approximated. What is really sought is a test of whether the reviewer's scores

are influenced by the proposal (i.e., are reviewers agreeing with one another.) A Monte Carlo simulation in this context could be created that takes into account the frequency of the existing scores in the sample data. In this way, the distribution of values remains the same, but there is confidence that there is no proposal influence on the scores. By generating this random assignment thousands of times, a distribution of statistics is revealed that can be used to establish confidence intervals for the statistics that we see in the sample data.

The Monte Carlo model of the data can help establish confidence intervals, but it does nothing to put the data into a measurement construct. That is, from the Monte Carlo simulation it is possible to learn that the sampled data are significantly different from a random outcome, but one cannot infer from that "how different" the groups (or scores are). To move from a statement of difference to a statement of how much difference requires that much more than just the means of scores is taken into account.

Analytic Techniques Addressing Decision Making

There are two measures of editorial influence that will be considered. The first is one that will be called "editorial choice." The second is one that will be termed

"editorial reach." Each will be described in detail in the following sections.

Editorial Choice

A "division chair" (acting much like an editor) is charged with making the decision about which proposals are accepted and which rejected. A central question about the process is: What is the extent to which the decision makers adhere to the ratings of the reviewers? Multiple raters each offer scores, which are combined into a single mean score. If the ratings are meaningful, those proposals achieving higher mean scores should be more likely to be accepted than those with lower scores.

But, even when adhering perfectly to the raters' scores, some degree of editorial choice may be inevitable; for example when forced to select 50 of 100 proposals where proposals 41 through 60 (when rank ordered) have the same score. An editor in this position must select 10 of the 20 proposals. This is forced editorial choice. Other editorial choice is more arbitrary, based on a number of different possible factors.

Arbitrary editorial choice occurs when a proposal's score is above or below the cut score, and a decision is made contrary to the indication of the score (i.e.,

choosing to accept a proposal whose score is below the cut score, or rejecting a proposal whose score is above the cut score). Of course it is important to remember that making either choice compels an instance of the opposite choice, although that instance may be "hidden" in the group of proposals at the cut score. The following example illustrates these phenomena.

Imagine an editor forced to select 5 of the 10 proposals represented in Table 6. Proposals 4, 5, and 6 have the same score. The editor, in order to select 5 proposals, is forced to make an editorial choice among them. Given the need to select 5 proposals with three proposals at the cut score, it will be termed that the editor has three "forced" editorial choices, as the choice among the three cannot be made based on differences in the score.

Table 6

Example of Possible Proposal Scoring

Proposal	Score
1	4.0
2	4.0
3	3.5
4	3.0
5	3.0
6	3.0
7	2.8
8	2.2
9	1.2
10	1.0

Now assume that the editor decides to accept proposal 7 as one of the five. This means the editor is now compelled to reject a compensating proposal at or above the cut score. Such a selection will be called an "arbitrary" editorial choice. Further, because the scoring was intended to identify the proposals to be accepted, proposal seven has been "misscored." If a proposal is rejected with a score above the cut score, or accepted with a score below the cut score, that proposal will be said to have been misscored. To make the terminology clear, consider the two options that the editor has after selecting proposal seven.

The editor can choose to reject a proposal at the cut score or above the cut score. In either case this will be

considered an arbitrary choice. If the editor chooses to reject a proposal above the cut score, that proposal will be said to have been both misscored and an arbitrary choice. If the editor chooses to reject one of the proposals at the cut score the proposal will not be regarded as misscored, but will be described as an arbitrary choice. The overall editorial choice will be summarized in the following way.

Assume the editor chooses to select proposals 1,2,3,6, and 7. In such a case, the editor had 3 forced editorial choices (proposals 4, 5, and 6) imposed by the situation, made two arbitrary choices (proposal 7 and one of the proposals 4, 5 and 6), and decided that one proposal (proposal 7) was misscored.

Cut Score

Now consider the cut score. In general, when making a decision based on a score there are two methods used depending on the circumstances. In one circumstance (e.g., a driving test for a driver's license) there is an established cut score above which one is included (receives a driver's license), and below which one is excluded (is denied a license). In the case of a driver's test, there is

no limit to the number of driver licenses available, and all who meet or exceed the cut score receive a license.

Another situation occurs where only a limited number of applicants may be accepted (e.g., admission to a competitive academic program). Where only some of the applicants are accepted only those with the highest scores will be included. This is accomplished by putting applicants in rank order according to their scores, and selecting the top applicants until the number of available admissions is met. In this case the cut score is established after a decision is made, and is the lowest score achieved in the group of those accepted.

In both methods, the selection process is based on the score achieved. In the case of a selection process like that of a drivers license, a pre-defined cut score is established and selection is based on only the score. In the latter case both the score and the rank order are involved. In the competitive situation (assuming the decision is based entirely on the score) the cut score is determined after the selection process and becomes the lowest score of those accepted. However, many times a combination of the two can occur, as when there is both a minimum acceptable score and a limited number to be accepted.

In the large research organization we can assume that a combination approach is used. That is, there are a limited number of proposals that can be accepted, but any accepted proposal must meet some minimum standard. While no explicit cut score may be established, an implicit cut score is created once the decisions are made.

In the large research organization the decision is not based exclusively on the score, but also on the basis of editorial choice. That is, the decision-maker has the prerogative to consider factors other than raters' scores to make the decision to accept or reject. This somewhat complicates the determination of a cut score, as some accepted proposals might have scores lower than some that are rejected.

For the purposes of this study a hypothetical cut score will be computed based on the number accepted and the rank order of the scores. That is, once the number (N) of accepted proposals has been determined a cut score will be calculated by "counting down" the rank ordered scores by the number of accepted proposals, and taking that Nth score as the cut score.

It is useful to express some measure of editorial influence over the acceptance of proposals. Editorial influence involves: accepting proposals with scores that

are below the cut point, rejecting proposals with scores above the cut point, and selecting a portion of the proposals at the cut point. Therefore, a measure of total editorial influence is the percentage of those proposals falling into those three categories. The sum of the high-scoring rejected, low-scoring accepted, and the number of proposals at the cut score (if not all of those proposals were accepted), taken as the percentage of the total number of proposals is the measure of editorial choice.

Editorial Reach

"Editorial reach" is a measure of how far an editor "reaches" in accepting low-scoring proposals or rejecting high-scoring proposals. An editor who accepts very low-scoring proposals exhibits greater reach than does an editor who accepts proposals very near (but below) the cut-score.

Further, an editor displays greater reach in identifying a proposal as misscored where the reviewers were in agreement as opposed to where reviewers disagree. For example, accepting a proposal one point below the cut-score where the reviewers are in perfect agreement (variance of 0) represents greater reach than accepting a

proposal of the same score but with a higher variance (i.e., greater disagreement).

An index of overall editorial reach can be calculated. Such a calculation must take into account three factors: the difference in score between cut score and the misscored proposal, the agreement index (variance) of the misscored proposal, and the total number of proposals. A formula for such an index is $\frac{1}{N}\sum_{i=1}^{N_m}\frac{\sqrt{(S_c - X_i)^2}}{1+A_i}$. Where N_m is the number of misscored proposals, S_c is the cut score, X_i is the score of the ith misscored proposal, A_i is the agreement index of the ith proposal, and N is the total number of proposals.

Questions/Hypotheses

The questions to answer are of several kinds. One kind of question at the simplest level is about the consistency or fairness of the decision making process. That is, are those who received higher ratings more likely to be selected than those with lower ratings? This question is relatively simple to answer by computing mean scores achieved and plotting them against the decision made. If

the highest scores are selected, then it is a fair inference that the decision made is based on the score.

Another indication of consistency is the extent to which reviewers agree with one another in their decisions. This becomes a more complicated question requiring more careful analytical methods to answer. It is reasonable to expect that reviewers will display some degree of agreement with one another. This will be indicated by ANOVA and G study results, as well as by comparison with agreement found in the Monte Carlo simulation.

The efficiency of the process is another important consideration. Questions about the efficiency will explore how the 10 criteria combine to create a score. Are 10 criteria sufficient? Do they reveal multiple factors that contribute to the AQ of the proposal? Efficiency is also involved in the number of reviews each proposal receives. Is that number sufficient? Is that number excessive?

In addition to the above questions, this large data set may permit the testing of some suggestions of previous research and theory of scientific knowledge. Are there differences between the divisions of the organization? Can those differences be seen as reflecting the differences seen between the hard and soft sciences in peer review acceptance? Additionally, these data can be analyzed to

test the assumption that peer review can be seen as a measurement process.

The most basic of statistical tests are likely to show that those proposals that received the highest scores are more likely to be accepted than those with lower scores. Beyond that, the data are likely to show that there is sufficient agreement among reviewers to justify the conclusion that they are responding to differences in the proposals.

The research organization divides its submission process by sub-topics of the organization's larger field. So the proposals were submitted to address issues of a particular sub-topic. The divisions of the organization may have differing ways of approaching the review process. The literature of scientific paradigms suggests that areas of different focus will reveal different approaches. Specifically, those areas that have a more established paradigm should show more consistency in the decision making process, while fields with a developing paradigm should show more variation in responses. This data set should be able to address this question: Do the decisions of these sub-groups vary?

The people who rate the proposals and the people who submit the proposals have varying amounts of experience and

status. The results of different status or experience can be compared. Do high-status reviewers make decisions different than those of lower-status reviewers? Do high-status authors receive different decisions than low-status authors? Where reviews differ on a given proposal, is the decision of the higher status reviewer more likely to be accepted?

The research organization and many others engaged in the peer review process, take great care in evaluating proposals over what are considered as separate and distinct criteria. In other measurement contexts it is common for a "halo effect" to develop. The halo effect occurs when an overall impression of the quality of a work influences the judgment of individual criteria. When such an effect occurs, the use of individual criteria loses its effectiveness, and instead each criterion becomes a proxy for a more general assessment of the work. Will such a halo effect be evident in the research organization's reviews?

The literature suggests that those who have published will tend to be published in the future. This makes a certain amount of intuitive sense. Those who have established the ability to produce works that are passed by the peer review process might reasonably expect to enjoy a greater degree of success than those who have never

published. If this is true then a good hypothesis is that those with greater experience and/or status should score higher and receive higher acceptance rates.

CHAPTER 4
FINDINGS

The Broad Picture

With an understanding of the key terms involved in the data set, and with an understanding of how they relate to one another, what follows is an examination of the specifics of the data under analysis. The data were collected over three years, and in general the description and analysis will be carried out independently for each of the three years, and where appropriate on a division by division basis.

Table 7 summarizes the overall number of proposals, reviews, and reviewers for the three years, along with the mean number of reviews that each proposal received, and the mean number of reviews given by each reviewer.

Table 7

Summary of Proposals, Reviews and Reviewers by Year

	Year			
	2001	2002	2003	Total
Proposals	3206	3900	3969	11075
Reviews	10248	12310	11382	33940
M[a] (SD)	3.20 (0.74)	3.16 (0.76)	2.87 (0.69)	3.06(0.74)
Reviewers	1860	2641	2179	6680
M[b] (SD)	5.51 (4.67)	4.66 (3.89)	5.23 (4.43)	5.08(4.31)

[a]Mean reviews per proposal. [b]Mean reviews per reviewer.

Table 7 shows that over the three years there was an increase in the number of proposals submitted and a large change in the number of reviewers. A casual look at the figures in Table 7 shows differences in the means across the years, and ANOVAs computed comparing the means show the differences to be statistically significant ($p < .0001$). However, because of the large sample size, statistical significance is fairly easy to achieve. Next will be discussed the characteristics of the reviewers across the divisions and through the years.

Reviewer Summaries by Division and Year

Table 8 sub-divides the information presented in Table 7. While means range from a low of 2.9 reviews per reviewer in Division 8 in 2001, to a high of 6.3 (Division 9, 2002 and 2003), what are notable about this are the maximum values and the resulting high SD. This is a result of a relatively few reviewers submitting many times the average number of reviews. For example, in year 2 one reviewer contributed to the reviews of 41% of the proposals submitted to Division 2.

Table 8
Proposals per Reviewer

Division	2001	2002	2003
1			
M (SD)	8.9 (6.800)	4.8 (4.756)	5.6 (4.6)
N[a] (Max[b])	199 (38)	204 (29)	148 (28)
2			
M (SD)	6.8 (5.684)	5.1 (7.053)	5.4 (4.9)
N[a] (Max[b])	138 (28)	131 (54)	134 (31)
3			
M (SD)	5.8 (4.399)	3.9 (2.660)	5.0 (3.7)
N[a] (Max[b])	496 (36)	609 (24)	512 (28)
4			
M (SD)	5.1 (3.391)	4.2 (2.801)	3.8 (2.8)
N[a] (Max[b])	216 (22)	293 (23)	237 (21)
5			
M (SD)	2.5 (1.526)	4.4 (2.788)	4.6 (3.2)
N[a] (Max[b])	80 (8)	86 (18)	71 (17)
6			
M (SD)	6.2 (6.575)	4.7 (2.958)	5.7 (3.4)
N[a] (Max[b])	27 (35)	49 (15)	30 (17)
7			
M (SD)	5.7 (4.528)	5.0 (3.716)	4.7 (3.7)
N[a] (Max[b])	112 (34)	204 (24)	179 (26)
8			
M (SD)	2.9 (1.616)	4.8 (3.443)	4.7 (2.9)
N[a] (Max[b])	95 (9)	130 (16)	103 (15)
9			
M (SD)	5.9 (5.757)	6.3 (6.024)	6.3 (5.1)
N[a] (Max[b])	75 (22)	67 (21)	71 (20)
10			
M (SD)	4.9 (2.897)	4.7 (3.483)	5.6 (3.9)
N[a] (Max[b])	144 (14)	195 (23)	200 (24)
11			
M (SD)	4.4 (3.422)	5.4 (4.310)	6.1 (6.3)
N[a] (Max[b])	227 (25)	526 (49)	385 (76)
12			
M (SD)	3.3 (2.390)	4.2 (3.111)	5.8 (4.7)
N[a] (Max[b])	51 (13)	147 (16)	109 (28)

[a]Number of proposals. [b]Maximum proposals per reviewer.

Proposal Summaries by Division and Year

11,075 proposals were submitted for review across the three years included in this study. Table 9 shows how these proposals were distributed across the divisions for each year, and the percentage that each division accepted, as well as the overall acceptance rate.

Table 9
Proposals Submitted and Accepted

	Year					
	2001		2002		2003	
Div.	N	Accepted	N	Accepted	N	Accepted
1	226	58%	252	63%	302	44%
2	136	44%	200	54%	212	62%
3	640	58%	712	74%	830	60%
4	227	44%	318	55%	274	59%
5	87	68%	126	71%	115	68%
6	55	55%	76	42%	62	47%
7	273	65%	317	51%	443	60%
8	182	67%	211	64%	252	54%
9	106	64%	84	63%	92	49%
10	312	52%	341	60%	351	43%
11	748	43%	990	45%	829	52%
12	214	43%	273	51%	207	50%
Total	3206	53%	2454	58%	3969	55%

Table 10 shows information comparable to that shown in Table 9, but shows reviews per proposal instead of proposals per reviewer. In contrast to the distribution of proposals per reviewer, the number of reviews received by each proposal is constrained to a narrow range.

Table 10
Summary of Reviews per Proposal

Division	2001	2002	2003
1			
M (SD)	3.07 (.830)	3.45 (.941)	2.79 (.446)
N	226	252	302
2			
M (SD)	2.84 (.408)	2.94 (.670)	3.13 (.590)
N	136	200	212
3			
M (SD)	3.31 (.748)	3.40 (.764)	3.16 (.564)
N	640	712	830
4			
M (SD)	3.92 (.970)	3.66 (.756)	3.08 (.364)
N	227	318	274
5			
M (SD)	2.67 (.543)	2.91 (.456)	2.72 (.669)
N	87	126	115
6			
M (SD)	3.13 (.336)	2.97 (.431)	3.27 (.632)
N	55	76	62
7			
M (SD)	2.89 (.577)	2.89 (.415)	2.28 (.458)
N	273	317	443
8			
M (SD)	2.65 (.627)	3.06 (.303)	2.10 (.307)
N	182	211	252
9			
M (SD)	4.89 (.318)	5.05 (.344)	4.96 (.205)
N	106	84	92
10			
M (SD)	3.20 (.482)	3.00 (.675)	3.06 (.333)
N	312	341	351
11			
M (SD)	3.12 (.445)	2.92 (.648)	2.72 (.559)
N	748	990	829
12			
M (SD)	2.94 (.316)	2.85 (.378)	2.74 (.659)
N	214	273	207

Table 11 shows the number of reviewers for each of the divisions for each of the years. No reviewer rated proposals from more than one division, so the totals shown at the bottom of the table indicate the actual number of reviewers for each year. Across years there is likely to be considerable duplication of reviewers, so summing across the columns will over-estimate the number of reviewers for the three years. Because of continuing development and changes of the electronic system, along with user unfamiliarity, accurate tracking of people across years was not achieved.

Table 11

Reviewers per Division by Year

	2001		2002		2003	
Division	N	%	N	%	N	%
1	199	10.7%	204	7.7%	148	6.8%
2	138	7.4%	131	5.0%	134	6.1%
3	496	26.7%	609	23.1%	512	23.5%
4	216	11.6%	293	11.1%	237	10.9%
5	80	4.3%	86	3.3%	71	3.3%
6	27	1.5%	49	1.9%	30	1.4%
7	112	6.0%	204	7.7%	179	8.2%
8	95	5.1%	130	4.9%	103	4.7%
9	75	4.0%	67	2.5%	71	3.3%
10	144	7.7%	195	7.4%	200	9.2%
11	227	12.2%	526	19.9%	385	17.7%
12	51	2.7%	147	5.6%	109	5.0%
Total	1860	100.0%	2641	100.0%	2179	100.0%

Table 12 shows the comparable figures for authors by division across the years. Unlike reviewers, some authors submitted proposals to more than one division. Therefore, the totals at the bottom of the table are greater than the number of individual authors.

Table 12

Authors per Division by Year

Division	2001	2002	2003
1	129	184	119
2	65	128	100
3	520	536	341
4	278	234	27
5	3	105	2
6	43	62	42
7	145	212	58
8	48	160	11
9	98	63	52
10	222	219	138
11	519	654	308
12	155	167	60
Total	2225	2724	1258

Accepted proposals were assigned a presentation format. Examination of the cross tabulation of requested format and assigned format in Table 13 shows that about one third of the accepted proposals were assigned a format other than the one requested.

Table 13

Cross Tabulation of Requested Format and Accepted Format

Requested Format	Accepted As: Paper	Round Table	Poster	New Member Poster	Other	Total
Paper	3097	798	420	107	13	4435
Round Table	173	558	45	28	0	804
Poster	95	62	349	27	2	535
New Member Poster	9	7	11	123	0	150
Other	52	54	19	3	41	169
Total	3426	1479	844	288	56	6093

As mentioned, the reviews were made electronically and consisted of a one to five response to each of 10 criteria. Each criterion was presented as a statement or question with a negative anchor (1) and a positive anchor (5), and radio buttons for the choices 1, 2, 3, 4, and 5. The web-based system did not require reviewers to provide a response to each criterion. If no radio button was clicked, the system recorded a zero as the response. Zero responses were coded as missing in this analysis, and also in summaries of data that the system provided to decision makers.

The system was programmed with default questions that represented the questions asked of reviewers in the

organization before the adoption of the electronic system. The ten criteria and their negative and positive anchors are shown in Table 14.

Each person in charge of a division of the organization had the opportunity to customize the "questions" as well as the anchors. However, all responses to the criteria were restricted to a 1 to 5 value, and 1 was always interpreted as least favorable and 5 as most favorable. In each of the three years only one of the divisions took the opportunity to significantly alter the default criteria. In 2001 and 2002 the changes made were minor changes in phrasing or ordering of the criteria. For those years the order of the criteria was adjusted for this analysis.

In 2003, Division 11 altered the criteria such that direct comparisons of responses to the criteria including that division are impossible. However, a factor analysis of response patterns shows that the "q10" or "overall" question can serve as a proxy for the other questions. The division with the customized criteria showed no difference in the factor analysis (i.e., all criteria loaded on a single factor and q10 loaded most heavily on that factor).

Table 14

Proposal Review Criteria Statements and Anchors

Criterion	Negative Anchor	Positive Anchor
Choice of problem/topic	Insignificant	Highly significant
Theoretical Framework	Not articulated	Well articulated
Methods	Not well executed	Well executed
Data source(s)	Inappropriate	Appropriate
Conclusions/ Interpretations	Ungrounded	Well grounded
Quality of writing/ organization	Unclear/Unorganized	Clear/Well organized
Contribution to field	Routine	Highly Original
Membership appeal	Small audience	Large audience
Would you attend this session	No	Yes
Overall Recommendation	Not acceptable	Outstanding Proposal, Definitely accept

Author and Reviewer Characteristics

Included with the reviews and proposals were information about the author and reviewer. Table 15, Table 16, Table 17, and Table 18 show the distribution of author and reviewer characteristics across all divisions and all three years.

Table 15

Author Status Percentage across Divisions and Years

Status	Freq.	Percent
Professor	1057	9.5
Associate Professor	1349	12.2
Assistant Professor	2921	26.4
Graduate Student	2534	22.9
Educator	454	4.1
Other	1252	11.3
Total	9567	86.4
No answer	1508	13.6
Total	11075	100.0

Table 16

Summary of Author Years

Years	Freq.	Percent
<1	2485	22.4
1-4	2438	22.0
5-10	1748	15.8
11-20	667	6.0
>20	257	2.3
Not a Member	760	6.9
Total	8355	75.4
No answer	2720	24.6
Total	11075	100.0

Table 17

Summary of Reviewer Status

Status	Freq.	Percent
Professor	981	14.7
Assc. Professor	1127	16.9
Asst. Professor	1995	29.9
Grad. Student	820	12.3
Educator	435	6.5
Other	949	14.2
Total	6307	94.4
Missing	373	5.6
Total	6680	100.0

Table 18

Summary of Reviewer Years

Years	Freq.	Percent
<1	581	8.7
1-4	1732	25.9
5-10	2053	30.7
11-20	1143	17.1
>20	598	9.0
Not a Member	229	3.4
Total	6336	94.9
Missing	344	5.1
Total	6680	100.0

The Criteria

As mentioned, a factor analysis was done of the 10 criteria. The factor analysis allows an understanding of what underlying qualities might be influencing the rating of proposals and how the individual criteria might differentially reflect those qualities. Across all the years and in each division the findings were nearly identical. All of the criteria loaded primarily on a single factor, and in each case (division/year combination) the criterion that most heavily loaded on the single factor was the "Overall" criterion.

The only exceptions to the extraction of a single factor were the 2002/Division 9, and the 2003/Division 4. In both of those cases a second factor accounted for 10% of the variance (compared with 61% of variance accounted for by the primary component in the 2002 case, and 65% in the 2003 case).

The correlation matrix for each of the year/division combinations shows a high degree of correlation among all of the criteria, with significance ($p < .0001$) for every combination.

This predominance of a single factor and the high correlation among the criteria suggest that there is only a single underlying factor which the answers to the questions

are revealing. Further, the fact that in all cases the "Overall" (q10) response most heavily loads on that factor permits a simplification of the analysis by focusing on that response as a proxy for the other responses.

A central issue of the analysis is the degree to which reviewers agree with one another. For this reason, it is useful to establish some figure to serve as an index of agreement. Further, two thirds of the proposals received exactly three reviews. This fact is used to limit the analysis to those proposals, thus reducing the number of possible combinations of reviewers' scores.

A measure of agreement among reviewers is the variance of the responses given. This variance can act as the index of agreement. Limiting the analysis to proposals with three reviews results in 9 possible values for the index of agreement. But even by reducing the data in this fashion, there is a challenge in visualizing the remaining data.

A variety of aids to visualize and interpret the data are used to approach a solution to this problem. A scatter plot can allow a number of characteristics to be displayed on a single diagram. Figure 1 is a scatter plot showing the relation among three factors: the mean score (q10mean), the agreement index of the three scores received (q10var) and the decision (accept and reject). Owing to the fact that

all proposals in the plot received exactly three scores for the overall criterion, there are relatively few discrete mean scores and standard deviations possible.

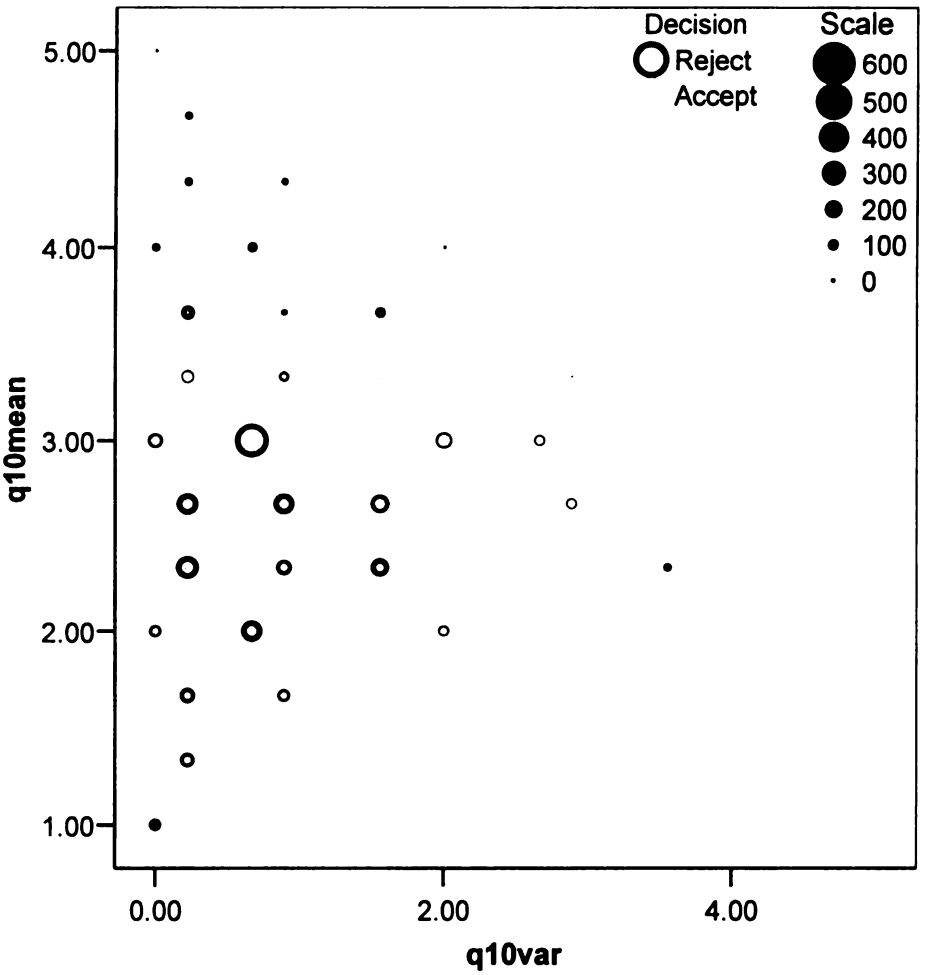

Figure 1. Accepted and rejected proposals.

Note that along the y-axis is the mean score (across the three reviewers) achieved by a proposal in the "overall" criterion (q10). The x-axis shows the agreement index resulting from the scores received. An agreement

index of zero indicates perfect agreement among the reviewers. Notice that above the 0.00 point on the y-axis are exactly five centers for scores. These are the 1-5 mean scores. Obviously, if there is perfect agreement among reviewers, the mean score will be the same as the score each reviewer gave.

The next variance higher than 0.00 is 0.22. Two clusters of proposals can be seen at that level of variance between each of the adjacent "complete agreement" scores. These represent scores with two raters in agreement and the third rater differing by only one point (e.g., 4-4-3, 2-2-1, etc.).[1] To the right on the plot are indications of greater disagreement among the reviewers. The size of the circles on the scatter plot shows an indication of the number of proposals with the given mean and rater agreement index.

Notice also, that for proposals with extreme means (i.e., means that are near 1 or 5), the variance is relatively low. This simply reflects the fact that for a proposal to achieve a high (or low) mean, it is necessary for the reviewers to agree on a relatively high (or low)

---

[1] Table 19 presents all of the possible combinations of three ratings and the associated agreement index.

rating. Conversely, the points on the scatter plot that indicate the highest variance are those points with the mid-range means.

Table 19

Ratings Associated with Each Agreement Index

Index	Possible Ratings
0.0	111, 222, 333, 444, 555
0.2	112,122,223,233,334,344,445,455
0.7	123, 234, 345
0.9	113, 133, 224, 244, 335, 355
1.6	124, 134, 235, 245
2.0	114, 144, 225, 255
2.7	135
2.9	125, 145
3.6	115, 155

Figure 2 is a bar graph showing the frequency of each agreement index in the sample data. Each bar represents the number of proposals achieving a particular level of agreement. 63% of the proposals had an index among the reviewer scores in the smallest three agreement indexes. A casual examination of Figure 2 might cause one to believe that there is considerable agreement among reviewers.

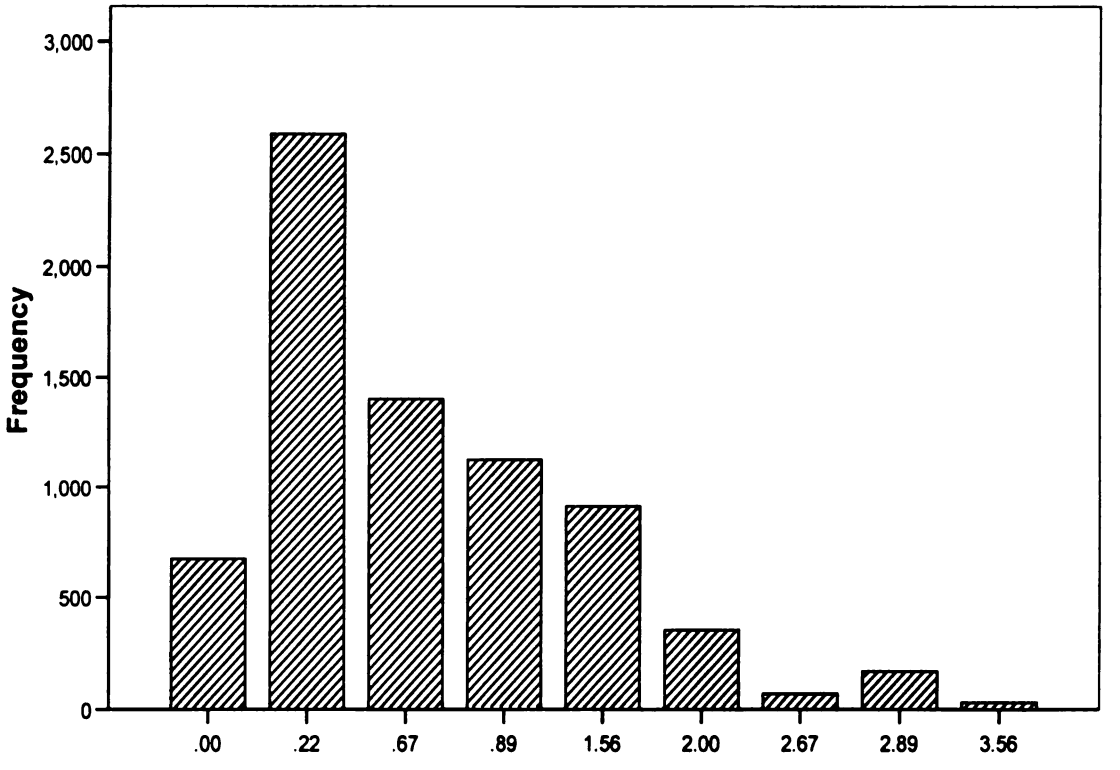

Figure 2. Frequency of each possible variance of three scores.

Figures 3 and 4 contain the same information as Figure 1, but Figure 3 contains only the accepted proposals, while Figure 4 contains the rejected proposals. Again the relationship between variance and mean is shown, and especially the high variance occurring in the mid-range scores. The separated scatter plots make it easier to see the difference in the distribution between the accepted proposals and the rejected proposals.

Note that in the plot of the rejected proposals there is only a slight weighting to the lower portion (lower mean score), while in the accepted proposals there is a clear preponderance of proposals with higher means.

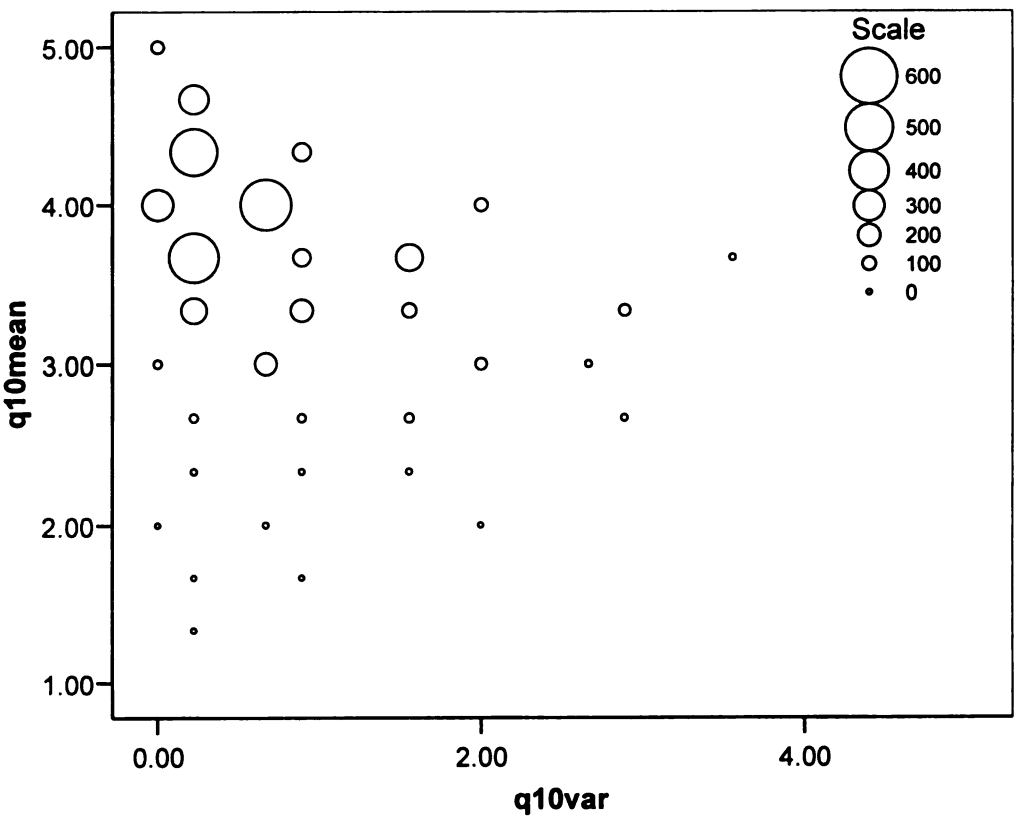

Figure 3. Mean versus variance of accepted proposals.

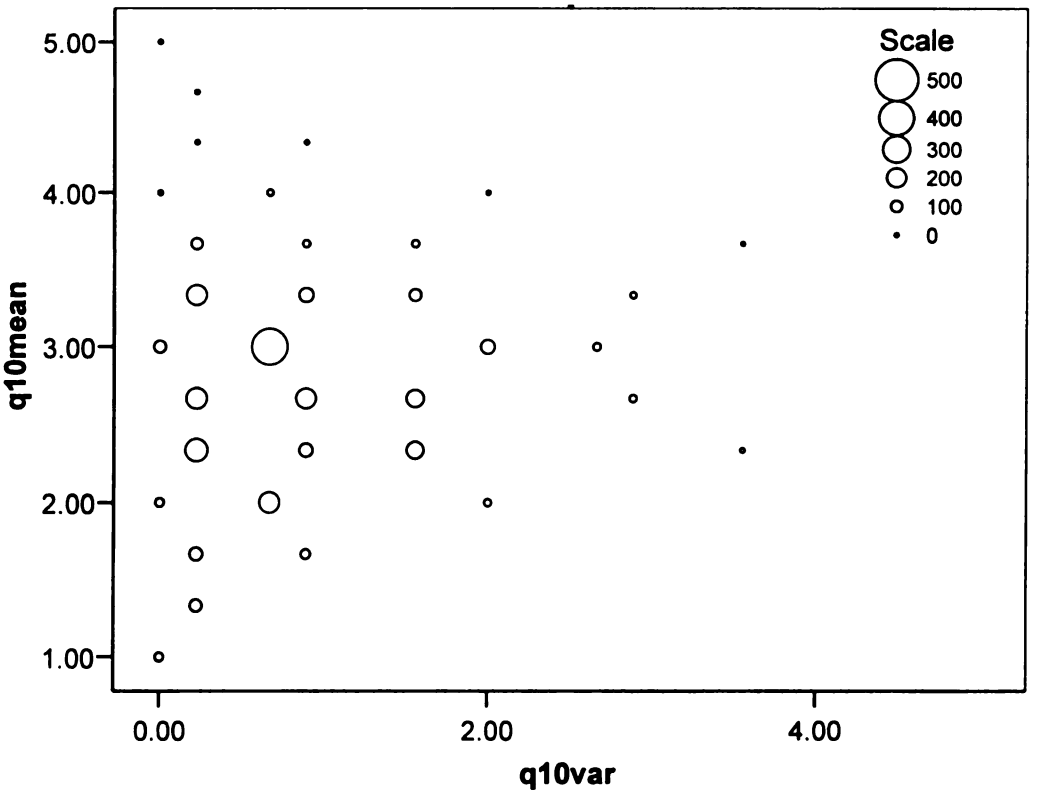

Figure 4. Mean versus variance of rejected proposals.

G Study

ANOVA relies upon assumptions about the data to provide meaningful results. Specifically, ANOVA assumes the design to be fully balanced. Referring back to the apple examples, each apple was rated by each reviewer, thus providing for a fully balanced design. In contrast, the data under analysis are examined in a matrix that is very sparsely populated. This is owing to the fact that there are many hundreds of reviewers and many hundreds of

proposals, but each reviewer rated relatively few proposals and about three reviewers rated each proposal.

Generalizability theory provides methods for analyzing some sparse matrices and was applied to the data here. A G study performed on the data resulted in variance components for the facets that were impossible to distinguish from the "noise." This can be interpreted as a failure not of the measurement process, but only of the analytical technique. The data simply are such that a G study offers no insight.

Results of Monte Carlo Simulation

With the failure of the G study, more importance is placed on the Monte Carlo simulation. Recall that the Monte Carlo simulation is employed to understand the agreement and the mean scores that would obtain from a purely random distribution of scores by reviewers. The basis of the random scoring is the pattern of scoring that was found in the original data. In the sample data 8.8% of the scores given were one, 18.1% two, 24.8% three, 32.3% of scores given were four, and 15.9% five. In performing the Monte Carlo simulation it is important to reflect this distribution in the generation of the random scores. To accomplish that in the random generating of scores each score is produced a corresponding fraction of the time.

The importance of properly weighting the generating of the random numbers is illustrated in Figure 5 and Figure 6. Figure 5 shows the result of a Monte Carlo simulation where 7314 simulated proposals were each given simulated ratings of three randomly generated scores. Each randomly generated score had equal likelihood of one, two, three, four, or five. This process was repeated 1000 times. Each simulated proposal received one combination of mean and variance among the scores. The median number of times a mean/variance combination occurred across the 1000 iterations (for each possible mean/variance combination) is shown in Figure 5. Note the symmetry across the mean scores and the predominance of variances indicating disagreement in the reviews.

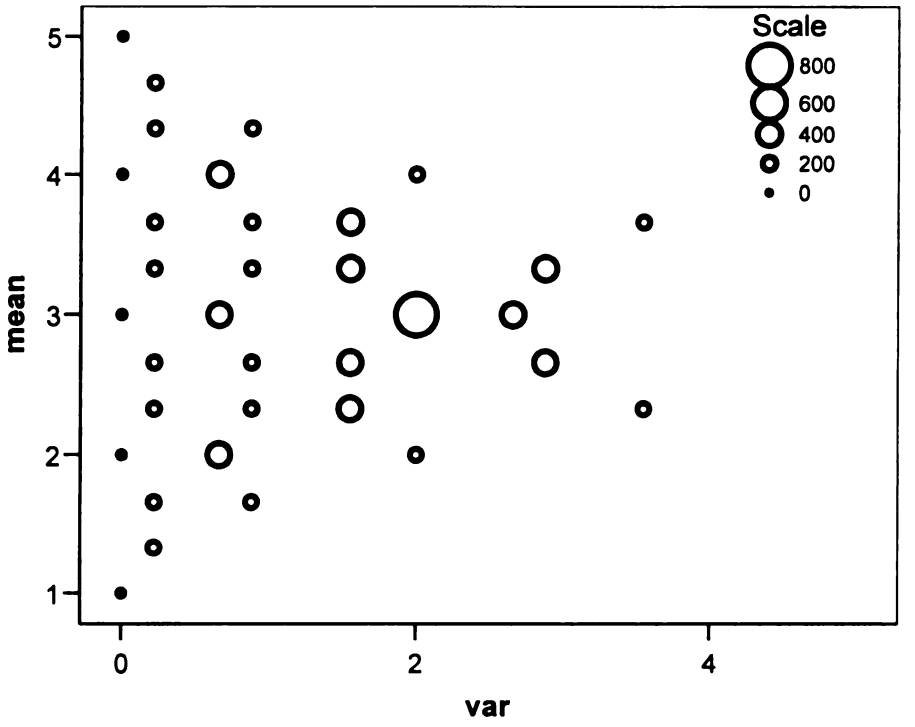

Figure 5. Median mean/variance outcome of unweighted Monte Carlo simulation.

Figure 6 shows the corresponding results of a Monte Carlo simulation where scores were generated according to the frequencies found in the sample data. The simulated reviews were made randomly, but the results were weighted to reflect the tendency of reviewers' differential use of the rating scores. Note the differences between Figure 6 and Figure 5. Figure 6 shows much more of a tendency for agreement, as well as an asymmetry with a predominance of mean scores around three and four.

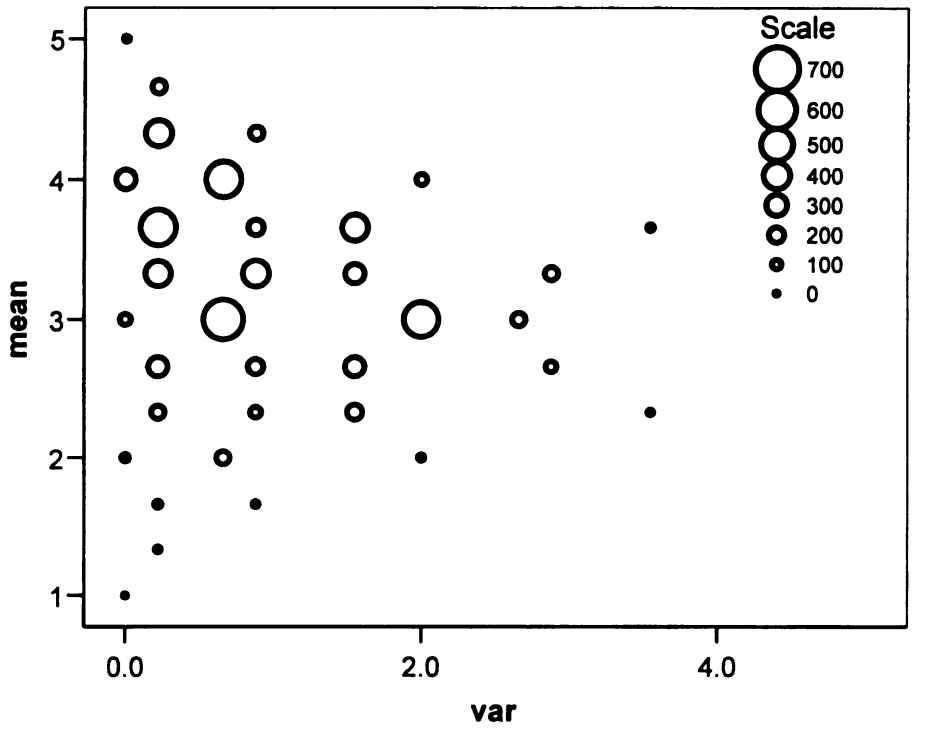

Figure 6. Random, weighted mean/variance distribution.

Figure 7 shows the mean/variance frequency from the sample data. The distribution of mean/variance found in the sample data is very similar to that found in the random, weighted Monte Carlo simulation. This indicates that a large amount of the tendency toward a particular mean score as well as reviewer agreement can be attributed to the tendency of reviewers to center their scores on four.

These three figures show the general trends for the distribution of mean/variance combinations. But to be specific and quantitative about the distribution of these scores it is useful to look at the distribution of mean

scores and the distribution of agreement indexes independently. Toward that end, a different view of the data can be helpful.

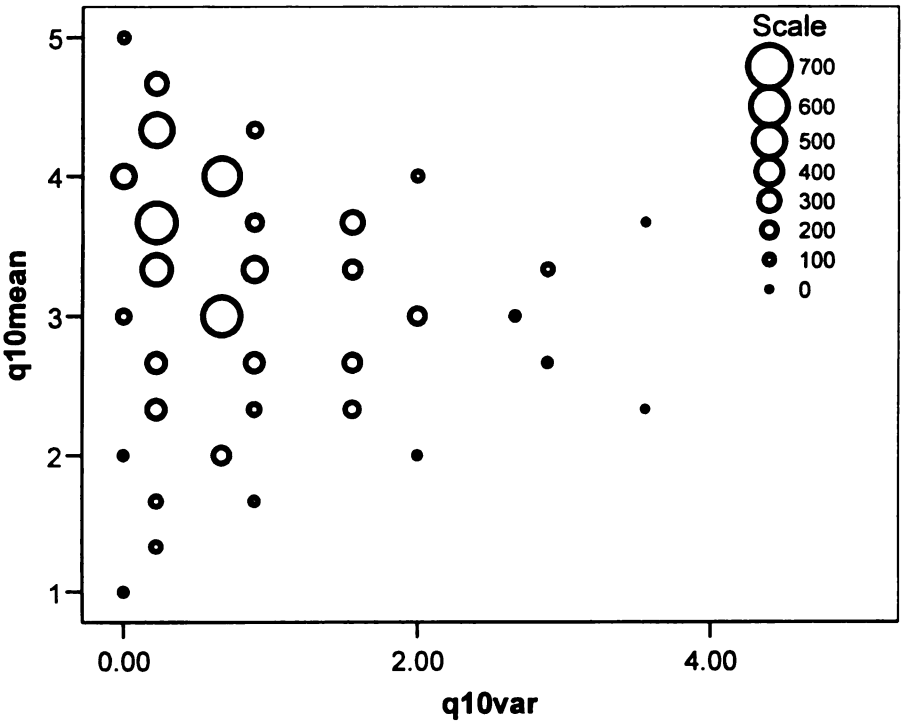

Figure 7. Sample data mean and variance frequency.

Figure 8 displays both the results of the Monte Carlo simulation as well as the sample data for all scores of 1,1,1. The histogram to the left of the figure shows the frequency across the 1000 repetitions for a particular number of randomly generated scores of 1,1,1. In nearly 200 of the 1000 iterations the score of 1,1,1 occurred four times. The most times that 1,1,1 occurred was 13 times. In the sample data 1,1,1 occurred 61 times. This results in a

Z score of 24.4, meaning that it is extremely unlikely ($p < .00001$) that the number of occurrences of 1,1,1 in the sample data occurred by chance.

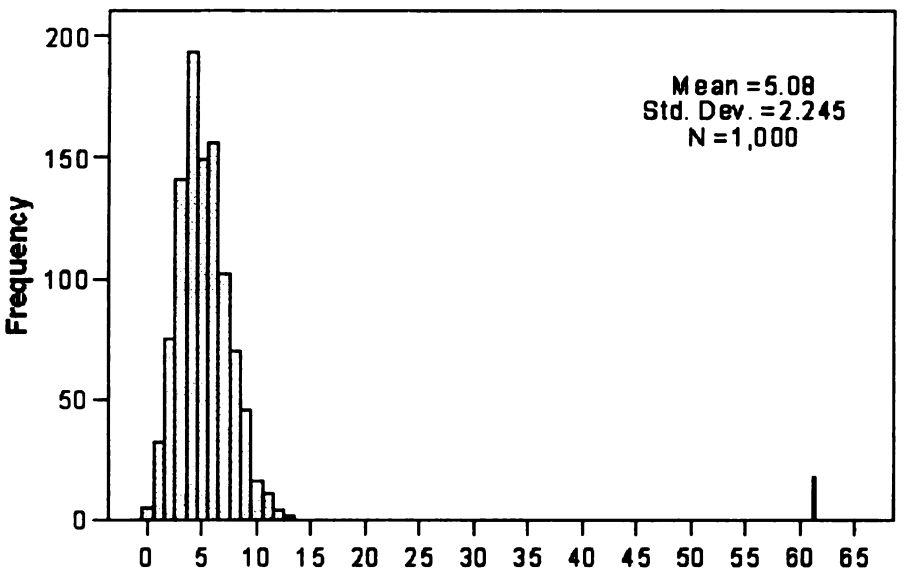

Figure 8. Histogram of frequency of 1,1,1 in simulated data compared with actual.

In contrast with the findings displayed in Figure 8, Figure 9 shows the corresponding results of the Monte Carlo simulation for a mean score of 3.33 and an agreement index of 0.2. In the sample data this result occurred 447 times, resulting in a Z score of 0.52. This indicates that there is no statistically significant difference between this result and what one could expect by chance.

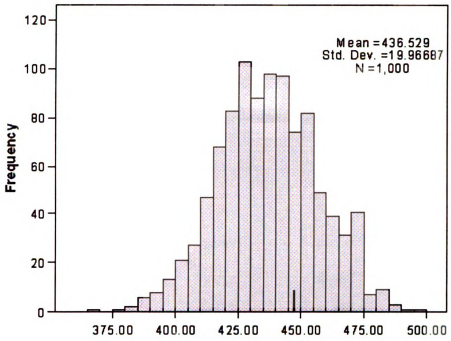

Figure 9. Histogram of frequency of 3,3,4 in simulated data compared with actual.

Table 21 presents the corresponding comparison between the Monte Carlo simulation for each mean/variance combination and the Z score of the sample data. The pattern is consistent across the years, indicating that agreement is more than expected by chance toward the extremes (high-scoring proposals and low-scoring proposals), and very close to chance results in the middle of the scoring range.

Table 20

Z Scores for Each Mean/Index

		Z score		
Mean	Index	2001	2002	2003
1.0	0.0	15.11***	8.82***	22.08***
1.3	0.2	7.73***	8.19***	8.46***
1.6	0.2	2.11*	5.34***	4.61***
2.0	0.0	2.85**	2.53**	0.20
2.3	0.2	4.13***	2.62**	2.77**
2.6	0.2	-0.75	1.23	0.33
3.0	0.0	1 20	1.15	2.46**
3.3	0.2	1.06	-0.68	0.44
3.6	0.2	0.11	1.25*	2.37*
4.0	0.0	2.06*	2.67**	2.74**
4.3	0.2	4.13***	2.85**	5.42***
4.6	0.2	4.55***	4.63***	5.21***
5.0	0.0	5.16***	7.75***	6.77***

*p<.05 **p<.001 ***p<.000001

Editorial Choice

Editorial choice is the measure of the portion of proposals (within a division) that were accepted with scores below the cut-score or rejected with scores above the cut-score.

Table 21 presents the editorial choice of each division by year. In divisions with a substantial number of proposals, the range is from nearly 3.5% to about 38% of proposal decisions made by the editor.

Table 21

Editorial Choice by Division and Year

	2001		2002		2003	
Div	N	Choice	N	Choice	N	Choice
1	127	19.69%	153	23.53%	224	17.86%
2	108	30.56%	130	12.31%	179	13.41%
3	439	24.15%	410	11.95%	600	10.83%
4	97	18.56%	119	19.33%	236	11.02%
5	61	16.39%	102	10.78%	86	3.49%
6	47	17.02%	62	14.52%	46	13.04%
7	187	24.06%	267	27.72%	119	37.82%
8	115	15.65%	196	8.67%	21	14.29%
9	2	0.00%	-	-	-	-
10	232	5.17%	246	18.70%	308	3.57%
11	619	22.29%	690	15.51%	543	11.05%
12	195	16.92%	229	9.61%	119	9.24%

Editorial Reach

Another measure of editorial influence is editorial "reach." Reach takes into account not only the editorial choice described above, but also the distance from the cut score as well as the index of agreement among the reviewers. Table 23 details the range of editorial reach across the divisions and through the years.

Table 22

Editorial Reach across Divisions and through Years

	2001		2002		2003	
Division	N	Reach	N	Reach	N	Reach
1	127	0.055	153	0.031	224	0.029
2	108	0.124	130	0.022	179	0.034
3	439	0.034	410	0.015	600	0.019
4	97	0.023	119	0.035	236	0.005
5	61	0.031	102	0.015	86	0.007
6	47	0.048	62	0.027	46	0.020
7	187	0.037	267	0.029	119	0.148
8	115	0.037	196	0.005	21	0.030
9	2	0.000	-	-	-	-
10	232	0.007	246	0.016	308	0.007
11	619	0.020	690	0.010	543	0.023
12	195	0.019	229	0.021	119	0.012

Reviewer Characteristics

There is no difference among categories of reviewer in their tendency to rate proposals. The tendencies seen across the categories of reviewer (both professional status and years of membership) are not different within the groups. Regardless of years of membership or professional status, all categories of reviewers use four about one third of the time, and use three and four over 50% of the time.

Author Characteristics

Author characteristics have some influence on the likelihood of being accepted. Authors vary by professional status and years of membership. Except for "Educator," each of the categories of professional status has equal likelihood of acceptance (55% are accepted), while the "Educator" status has a 45% likelihood of acceptance.

Author membership is also a significant predictor of acceptance. All member groups are more likely to be accepted than the non-member, and more years of membership increases the likelihood of acceptance, until the years of membership reaches greater than 20 years. This is generally consistent with the literature that suggests that those who

have published in the past are more likely to have their work accepted.

CHAPTER 5
DISCUSSION

In a data set of this size it is especially easy to make the error of conflating statistical significance with substantive significance. Statistical significance can be achieved with a small effect in such a sample size. But to make a sensible interpretation it is necessary to attend both to the probability figures (statistical significance) as well as the magnitude of any effect (substantive significance). The difference between statistical and substantive effect is clearly illustrated in the findings around reviewer agreement.

Answers to Questions

Reviewer Agreement

The goal of the research organization's peer review process, put most bluntly, is to accept good proposals and to reject bad proposals. Reviewers are thought to know what qualities make one proposal better than another and to be able to detect those qualities. For the process to be meaningful and effective reviewers must agree with one another. If reviewers disagree it is impossible to glean from their reviews anything other than their idiosyncratic

opinion. On the other hand, when reviewers agree, there can be confidence that they are responding to some quality within the proposal. If reviewers do not respond to qualities that inhere in the proposal, then peer review is an exercise in futility. Therefore, a central question is: Do reviewers agree?

The findings show that the answer to the question is more complicated than a simple yes or no. The data show that agreement among reviewers occurs more frequently than expected by chance. This implies that reviewers do have a shared idea of what makes a proposal acceptable and unacceptable, and are able to discern that quality. In other words, reviewers are responding to something in the proposal. But the findings also show that agreement of reviewers is not consistent across the rating scale. Reviewers are in greatest agreement about very low-scoring and very high-scoring proposals. There is strong evidence that in the middle range of scores agreement among reviewers is no greater than chance. This is because a large amount of the agreement found in the mid-range scores can be attributed to the fact that 57% of all scores given by reviewers were three or four on the five-point scale.

The agreement that was found at the extremes scores is much greater than expected by chance, but it still amounts

to a small fraction of the total number of proposals. Over half of the proposals received scores from reviewers with substantial disagreement. From this it is reasonable to conclude that many of the decisions to reject or accept are the result of chance.

Further, much of the apparent agreement among reviewers is attributable to the combination of two common rater problems. The reviewer pool tends to be lenient (i.e., reluctant to use scores of one and two), and to have a central tendency (i.e., a preference to rate proposals four). The centering of scores around four necessarily results in the appearance of agreement. But that agreement is an artifact of the distribution of scores and is not indicative of reviewers responding to the individual qualities of proposals.

While it is the case that agreement among reviewers occurs more than expected by chance, it does not occur more than reasonably expected by those relying on the outcome of the process. It was found through the Monte Carlo simulation that one could expect 35% of the proposals to receive agreement in their ratings without any influence caused by the content of the proposal. That means that "greater than chance" implies only that more than 35% of the ratings show agreement. The fact that 55% of the actual

reviews showed significant disagreement casts into doubt the substantive fact of agreement. This is an example of statistical significance failing to translate into substantive significance.

Efficiency

Another question addressed was that of the efficiency of the peer review process. It was found that the 10 criteria were not being used efficiently. That is, there was no discrimination among the various criteria. There are at least two possible reasons for this. One is reviewers' tendency to center their scores on four and vary from that score relatively rarely. Not only was this tendency seen across proposals, it was also found across criteria within proposals.

Another possible cause of the lack of discrimination among the criteria is that of a "halo effect" in the reviewers. A halo effect occurs when the reading of a proposal causes an overall impression that results in a similar score in all criteria. In such a case an "overall" score is taken as the score for each of the other criteria.

The question remains unanswered as to whether the number of reviewers is an efficient number in determining the quality of the review process. Questions about the

efficiency of the number of reviewers were unable to be answered owing to the sparseness of the matrix and subsequent inability to run a valid G study analysis.

Differences among Divisions and Years

The analysis also focused on the comparison across years and across divisions. This permitted a test of the robustness of the findings, as well of the possible variations among areas of the organization. Across the divisions and through the years, reviewers were found to agree to a similar extent and to make their ratings in a similar range of scores. No significant differences in reviewer behavior were found between divisions or from year to year.

Editorial influence varied widely across divisions and through the years. This points out the importance of the editorial role played by the division chair. Some chairs adhered closely to the reviewers' scoring, while others made liberal use of their editorial prerogative. It is difficult to judge which editorial behavior is more appropriate. Among the proposal scores where most of the editorial choice was made, there is relatively little reviewer agreement, and a large amount of the agreement that is there is likely to be the result of chance. An

editor who makes a decision based on the scores is basing it on unreliable scores.

Decision Based on Score

Not surprisingly, there is a tendency for editors to accept higher scoring proposals and to reject lower scoring proposals. But there is evidence that this is based primarily on rejecting lower scored proposals and making extensive editorial choice about proposals with a mid-range score. As mentioned, there is considerable variation among the editors in their adherence to the reviewers' scores. Some editors only rarely exercise editorial choice, while others make much more liberal use of it.

Implications/Recommendations

The implications of the findings depend to a great degree on the interpretations of the results and the desires of the research organization. The findings indicate that there are two main scoring problems: a central tendency of raters, and a failure to discriminate among the criteria. These scoring problems have at least two possible causes, each requiring a different solution.

Central Tendency

Over 55% of the scores given were 3 or 4. This is a problem because it results in a large number of proposals being given similar scores. Because of the similar scores, it is impossible to discriminate among those proposals. Such a scoring pattern is the result of one or both of two possible reasons.

One possible cause of the central tendency in scores is that proposals generally fall into that range of score. These might be accurate scores. If this is the case then the solution to the problem is to add precision to the scale by adding more steps, so that what had been scores of three and four are spread over a greater range.

The other possible cause of the central tendency is a preference by reviewers to give three and four regardless of the quality of the proposal. If that is the source of the problem, then the solution involves encouraging reviewers to properly use the range of the scale. This problem can be addressed by a combination of greater reviewer training along with a clearer scoring rubric. Instead of providing only negative and positive anchors (for one and five), scoring could be improved by offering descriptors for each of the intermediate scores, thus encouraging use of the entire scoring scale.

Deciding which of these solutions will address the problem requires pilot testing both methods with a representative sample of reviewers and proposals.

Halo Effect

A second problem is the halo effect evidenced by reviewers failing to discriminate among the ten criteria. This is more clearly a problem with reviewer behavior than is the central tendency, but again, the solution depends on the interpretation of the result of this reviewer behavior.

It is possible that the halo effect results only in the submission of an accurate overall score (the 10^{th} criterion). Because the ultimate goal of the scores is the binary choice of acceptance or rejection, it is reasonable to reduce the ten criteria to a single holistic score. If such is the case, then a more efficient scoring system might be to have the reviewer offer only the single score.

If, on the other hand, a separate and independent score for each of the criteria is desired, then additional training or instruction is required to achieve the desired discrimination among the criteria. This would result in a greater reliance on the editor to appropriately weight the various criteria to achieve the final decision of accept or reject.

Editorial Influence

In addition to the problems associated with reviewer disagreement, and lack of discrimination among the criteria, editorial choice has significant influence over the results of the peer review process. Editors should be made aware of the degree to which they are choosing to overrule the choices of their reviewers. Further, the organization could provide guidance as to how much editorial influence should be exercised. Instead of making decisions according to the judgment of the reviewers, some editors overrule the decisions of the reviewers in 25% or more of the cases, while other reviewers more diligently adhere to the reviewers' decisions.

Other Conceptions of Peer Review

This study has conceptualized peer review as a measurement process. As such it conceives peer review as intending to identify a quality within proposals that makes it worthy of acceptance. This is a reasonable conception given the formal process of collecting 30 numeric scores per proposal, and then using those numeric responses in selecting proposals for acceptance. But alternative conceptions are possible.

One alternative conception is that the process is as much for the reviewers and the organization as a whole, as it is for the authors and the selection of proposals. Reviewers who are involved in the process are engaged in reading their peer's work, and specifically attending to a set of criteria that are deemed important in scientific work. Reviewers are peers, and as peers are also engaged in the process of producing work that will pass through the peer review process. The act of reviewing others' work has the effect of reinforcing the standards that are expected of acceptable work, as well as broadening the reviewer by exposing him or her to the work of peers.

Limitations

While this data set holds a great amount of information capable of providing insight to the peer review process of this large research organization, it also is limited in its ability to illuminate the larger field of peer review. This is for several reasons.

Softer Science

The general scope of the research done by the large research organization falls within what would be described as a "soft science." That is, the research deals with

social and psychological issues as opposed to the "hard sciences" of physics, or mathematics. Because of that, all of the findings should be interpreted within that context. Peer review literature suggests that reviewer behavior may be considerably different in the hard sciences.

Proposals, Not Completed Works

This peer review process involves proposals, as opposed to completed works. Because it is understood that the proposals submitted represent work in progress, and because the submissions are made 6 to 8 months before the meeting is held, many proposals are submitted that may lack traits expected of completed work. This fact could lead to several effects.

There may be greater acceptance of a work that is known to be one in progress than one that is completed. Perhaps another reason to tend toward lenience is that in editorial peer review as it is commonly practiced, an author receiving negative reviews often has a chance to make revisions and re-submit the work. In the case of this large research organization, there is no such opportunity.

Future Research

Peer review is a field that calls for much more investigation. The difficulties in researching peer review are numerous, but the consequences of the peer review process are so broad reaching that it is important to better understand the process and to ensure that the process performs the task it is intended to perform. Future research suggested by this study might take several directions.

Text Analysis

One direction for future research is greater exploration of this study's data set. While this study looked only at the categorical data around the authors and reviewers, and the discrete scores of the criteria, the data set also contains text comments from each reviewer. Future text analysis may uncover relationships among qualities of the comments, the scores achieved, and the decision made to accept or reject.

Editorial Influence

There is much more to understand about the role of the editor. Toward that understanding, interviews of editors would be useful in gaining insight into the selection

process and their use of input from reviewers. How do editors view their role in the process? How does an editor make a decision about acceptance or rejection in the face of disparate reviews? What efforts are made at training reviewers?

Ultimate Publication

Many of the proposals submitted to the organization will result in papers that the authors will wish to publish in journals. Comparing the result of the journal peer review process with that of the research organization's process would provide another measure of the effectiveness of both. To accomplish this would require following up on the proposals to determine which ultimately resulted in publication.

Conclusion

The results of this study raise issues of serious concern to those who place the fate of their scientific and scholarly work into the hands of the peer review process. Unless a work is exceptionally good or exceptionally bad there is little reason to have confidence that the decision is reliable. Reviewers fail to discriminate among proposals with mid-range scores, and editorial choice governs much of

the decision making process. However, the problems are not uncommon in measurement practice and are amenable to solution. A more complete scoring rubric, greater training of reviewers, and common guidance for reviewers can make the process less random.

APPENDIX A
Tables of Reviews per Proposal

Tables A1, A2 and A3 detail the number of proposals submitted to each division, the mean number of reviews given per proposal, and the total number of reviews for each of the three years.

Table A1

Reviews per Proposal by Division for 2001

Division	Mean	N*	SD	Reviews
1	3.07	226	.830	694
2	2.84	136	.408	386
3	3.31	640	.748	2120
4	3.92	227	.970	890
5	2.67	87	.543	232
6	3.13	55	.336	172
7	2.89	273	.577	789
8	2.65	182	.627	483
9	4.89	106	.318	518
10	3.20	312	.482	999
11	3.12	748	.445	2335
12	2.94	214	.316	630
Total	3.20	3206	.742	10248

*Number of proposals submitted

Table A2
Reviews per Proposal by Division for 2002

Division	Mean	N*	SD	Reviews
1	3.45	252	.941	870
2	2.94	200	.670	588
3	3.40	712	.764	2421
4	3.66	318	.756	1164
5	2.91	126	.456	367
6	2.97	76	.431	226
7	2.89	317	.415	917
8	3.06	211	.303	645
9	5.05	84	.344	424
10	3.00	341	.675	1024
11	2.92	990	.648	2886
12	2.85	273	.378	778
Total	3.16	3900	.755	12310

*Number of proposals submitted

Table A3
Reviews per Proposal by Division for 2003

Division	Mean	N*	SD	Review
1	2.79	302	.446	843
2	3.13	212	.590	663
3	3.16	830	.564	2626
4	3.08	274	.364	844
5	2.72	115	.669	313
6	3.27	62	.632	203
7	2.28	443	.458	1009
8	2.10	252	.307	528
9	4.96	92	.205	456
10	3.06	351	.333	1073
11	2.72	829	.559	2256
12	2.74	207	.659	568
Total	2.87	3969	.686	11382

*Number of proposals submitted

APPENDIX B
Figures Showing Accept and Reject by Division and Year

Figures A1 through A36 compare the scores of rejected and accepted proposals for each division and each year.

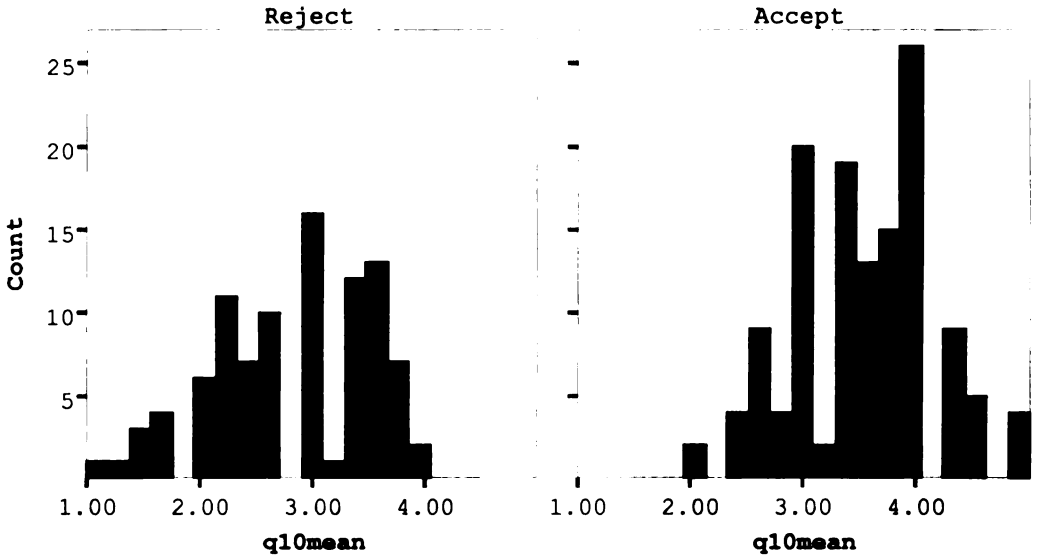

Figure A1. Accept and reject, Division 1, Year 1.

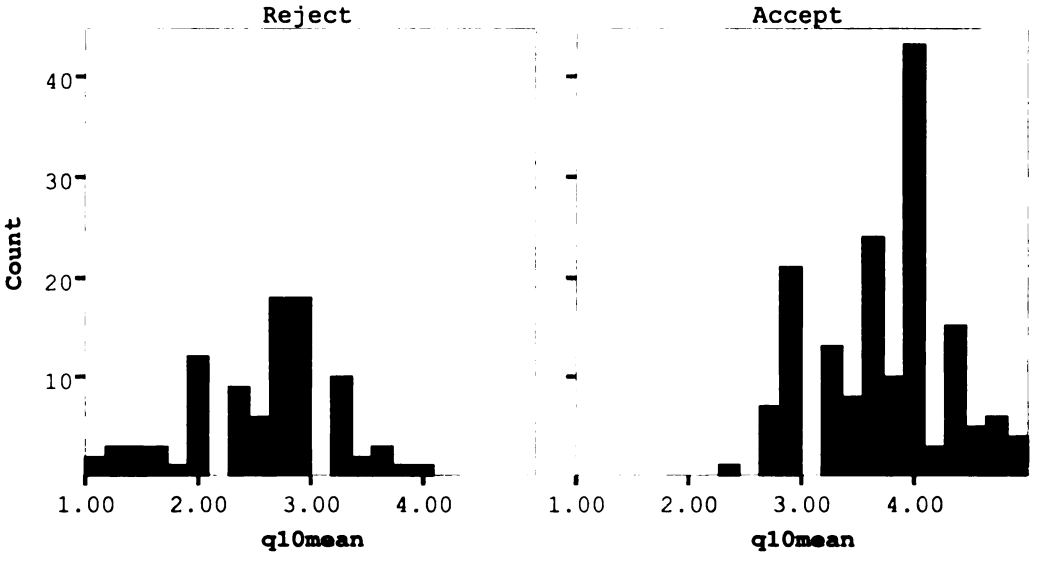

Figure A2. Accept and reject, Division 1, Year 2.

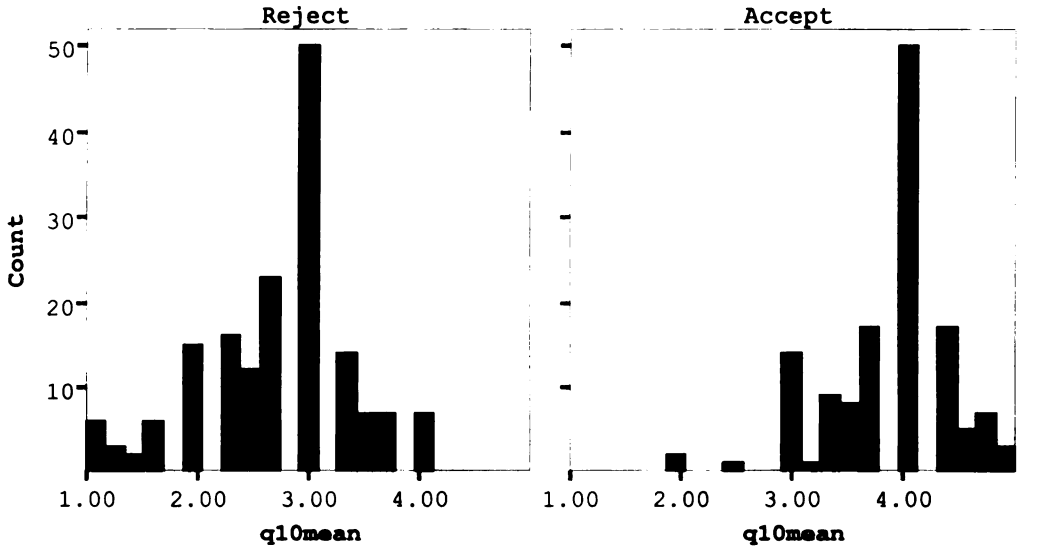

Figure A3. Accept and reject, Division 1, Year 3.

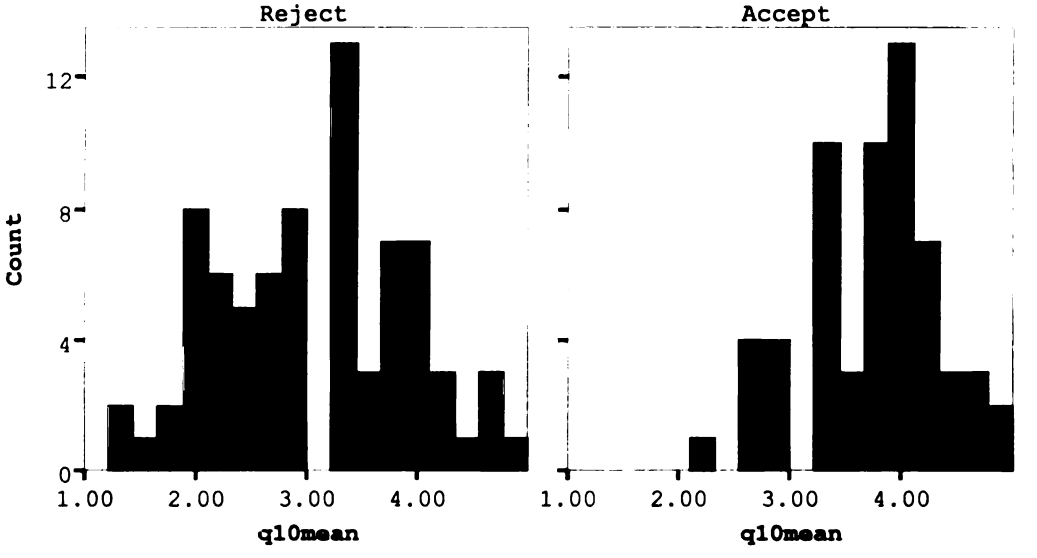

Figure A4. Accept and reject, Division 2, Year 1.

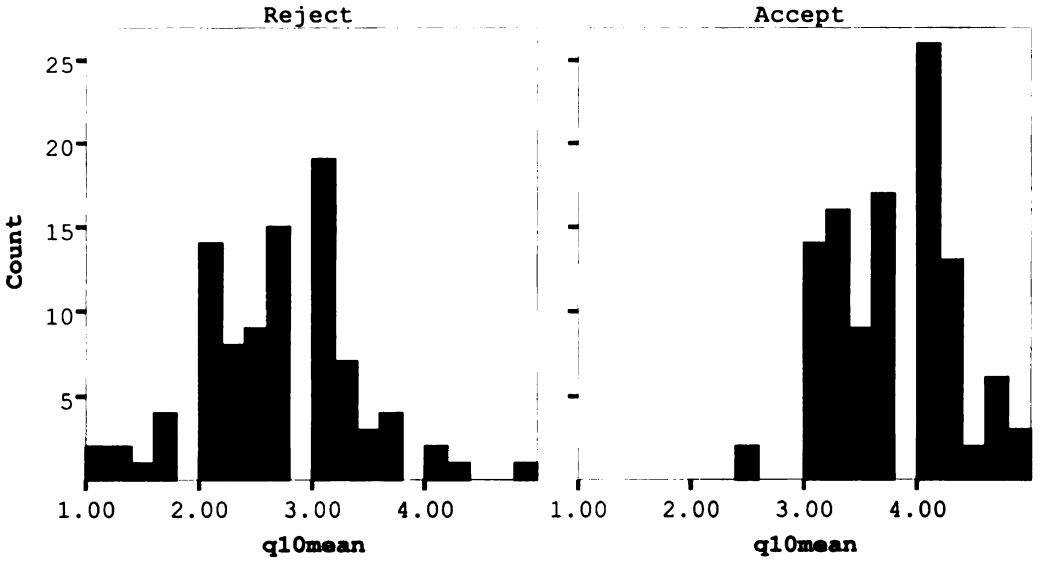

Figure A5. Accept and reject, Division 2, Year 2.

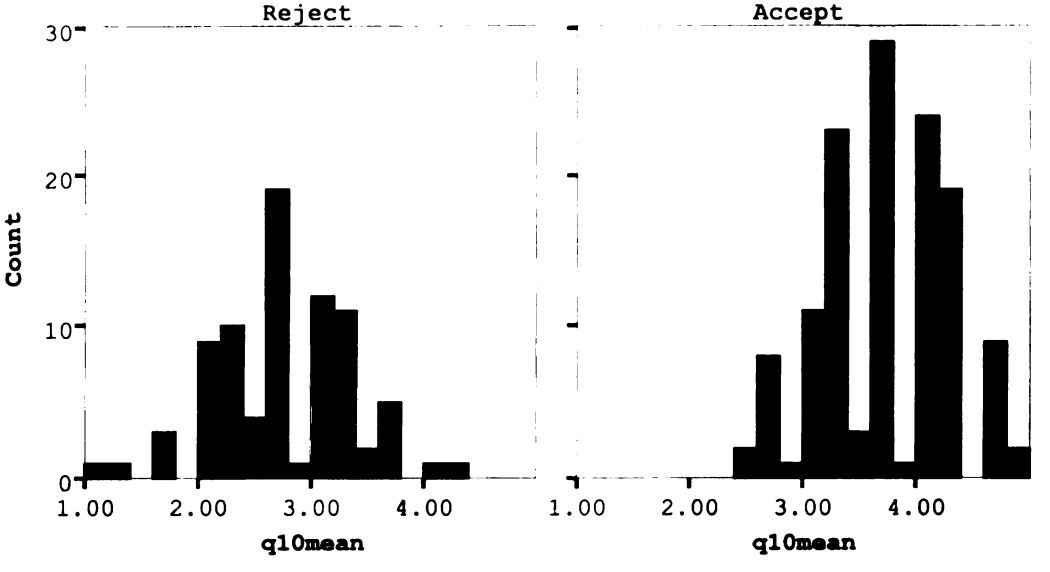

Figure A6. Accept and reject, Division 2, Year 3.

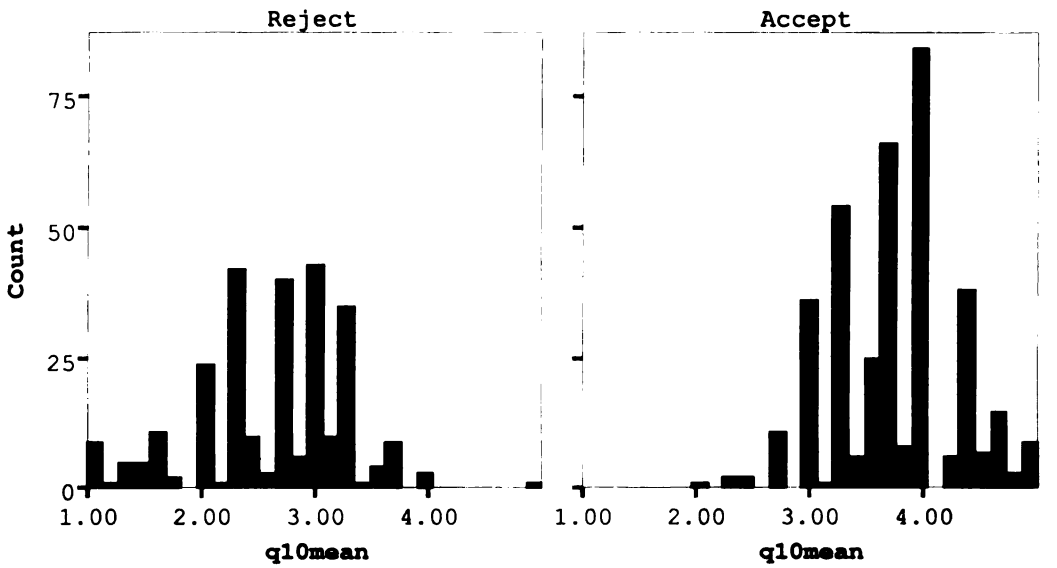

Figure A7. Accept and reject, Division 3, Year 1.

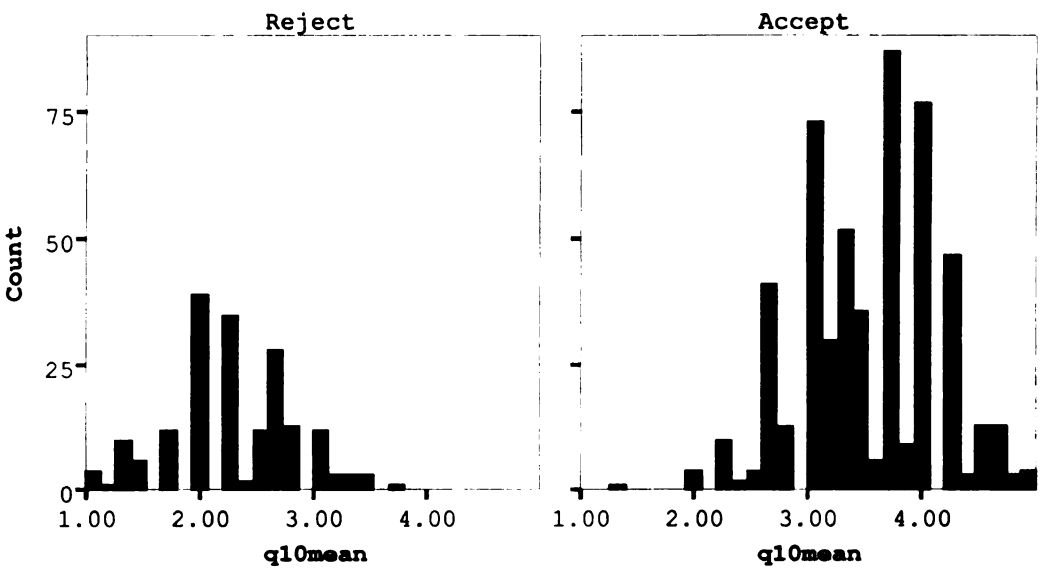

Figure A8. Accept and reject, Division 3, Year 2.

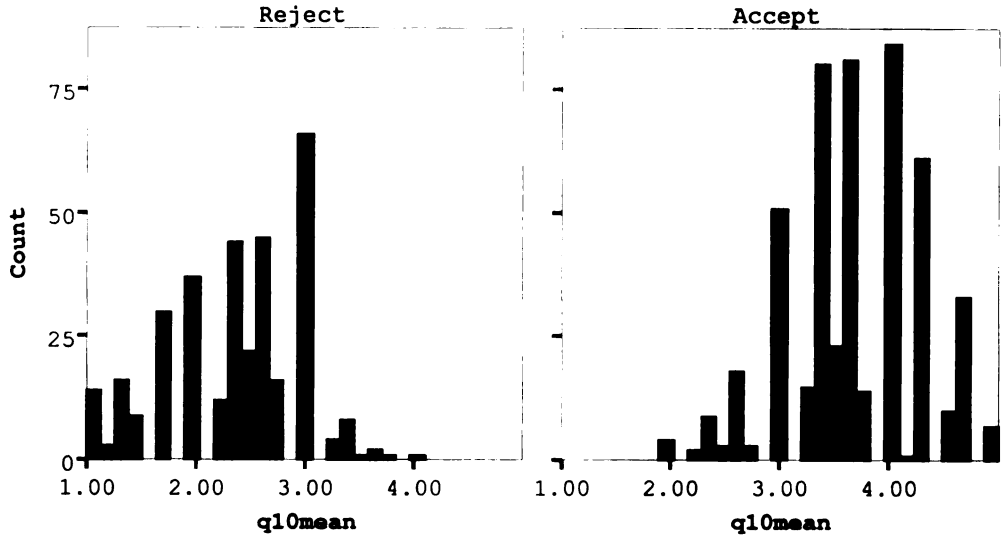

Figure A9. Accept and reject, Division 3, Year 3.

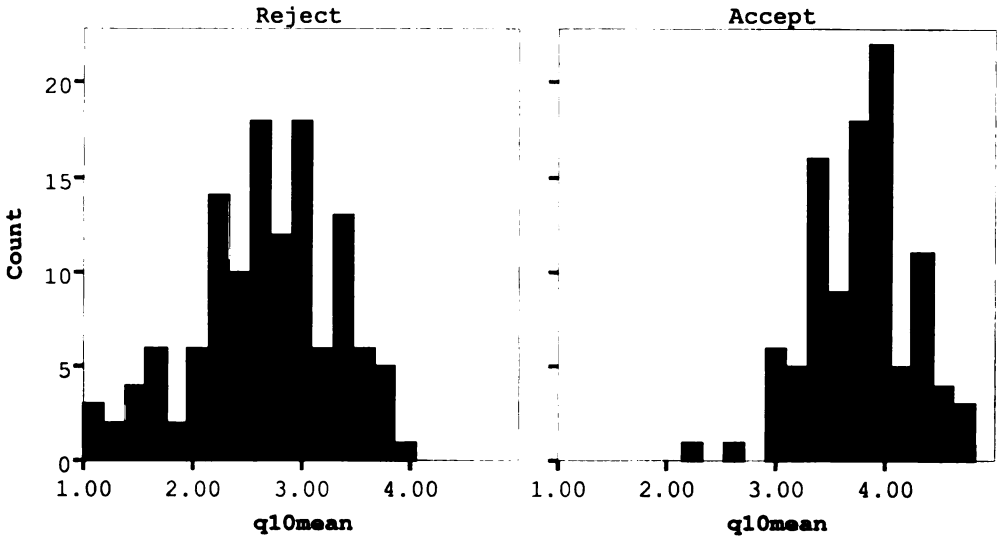

Figure A10. Accept and reject, Division 4, Year 1.

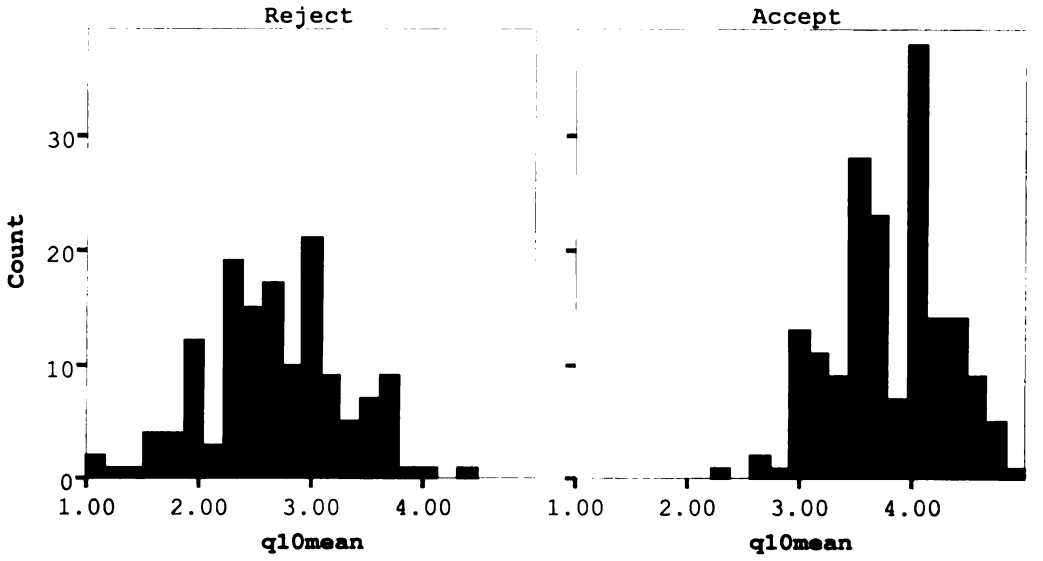

Figure A11. Accept and reject, Division 4, Year 2.

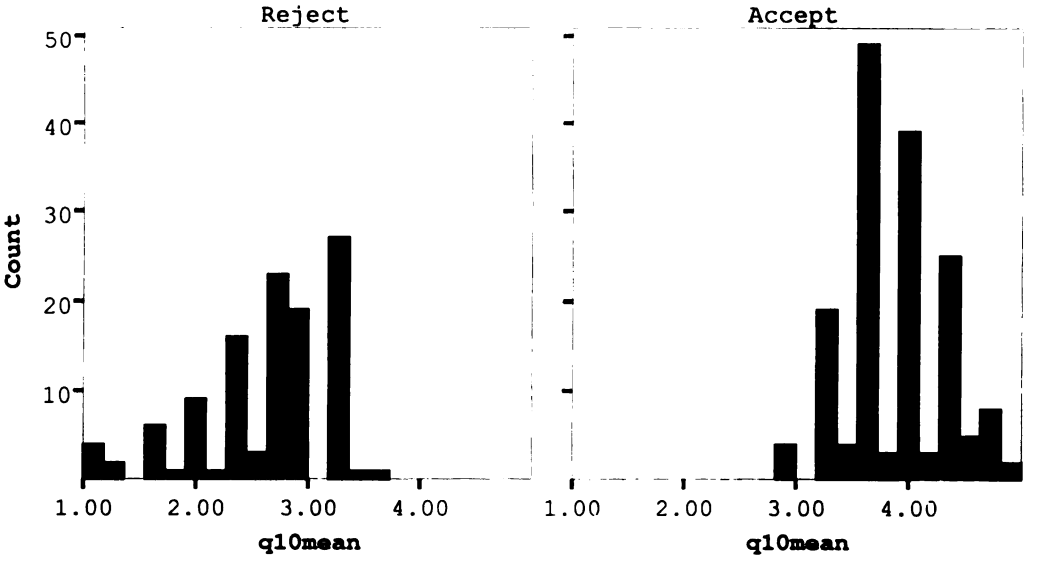

Figure A12. Accept and reject, Division 4, Year 3.

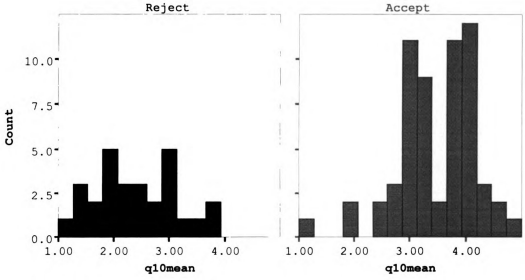

Figure A13. Accept and reject, Division 5, Year 1.

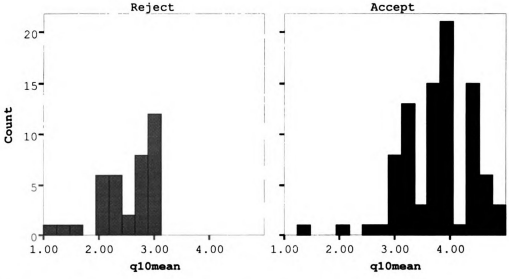

Figure A14. Accept and reject, Division 5, Year 2.

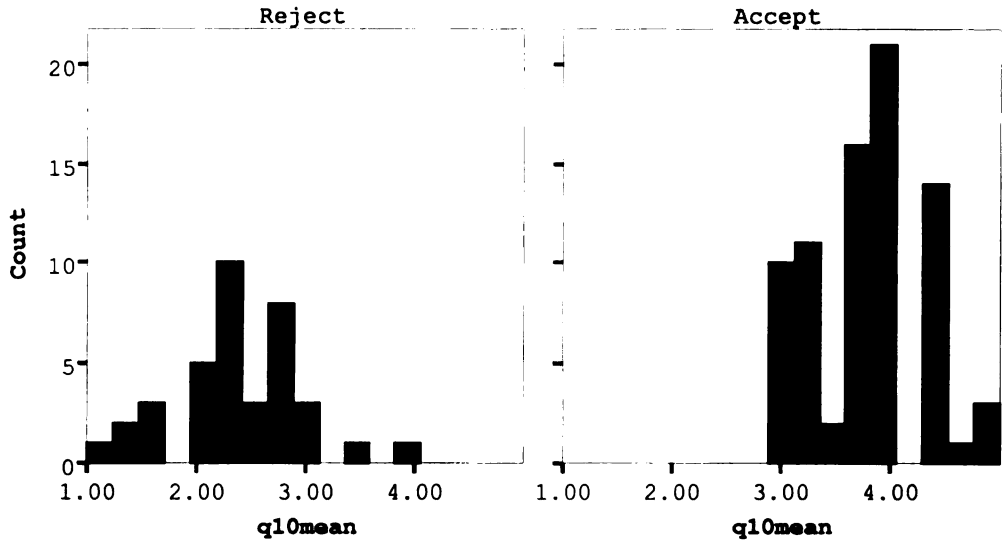

Figure A15. Accept and reject, Division 5, Year 3.

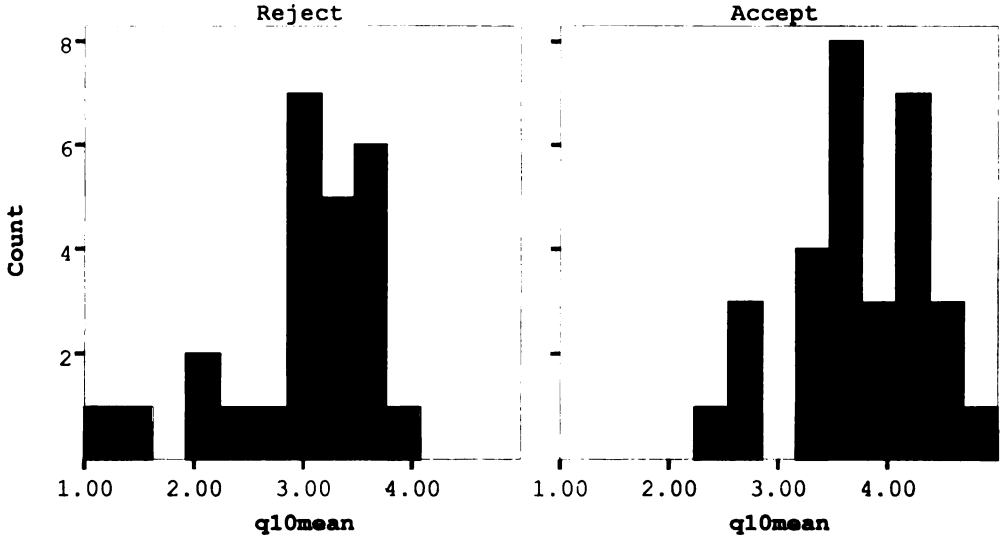

Figure A16. Accept and reject, Division 6, Year 1.

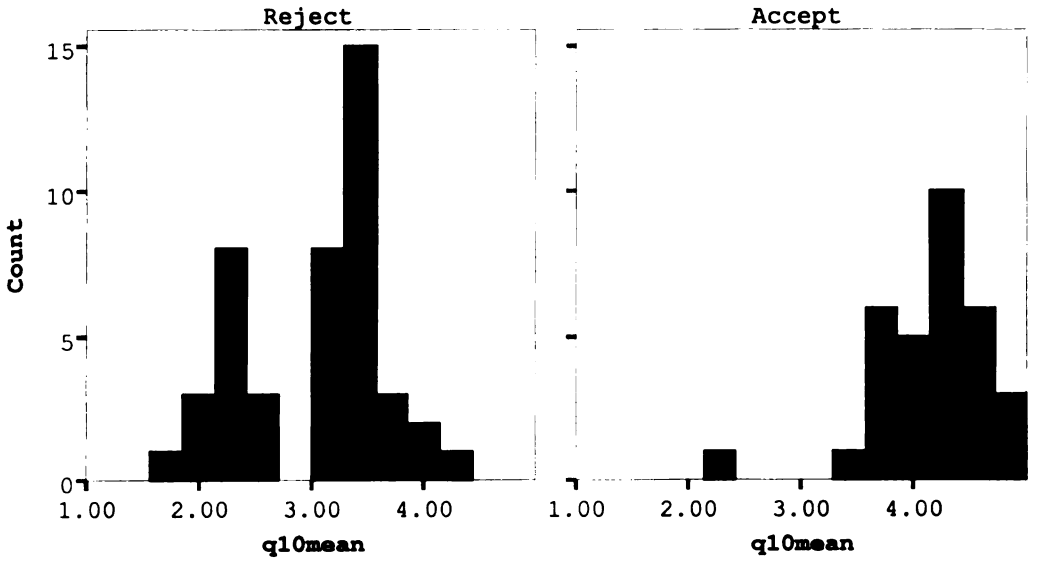

Figure A17. Accept and reject, Division 6, Year 2.

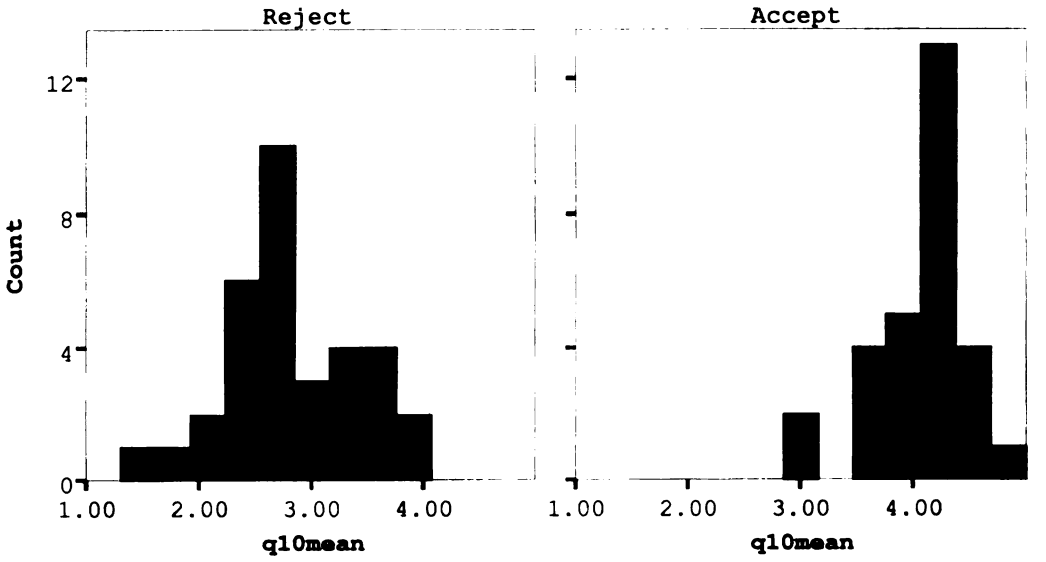

Figure A18. Accept and reject, Division 6, Year 3.

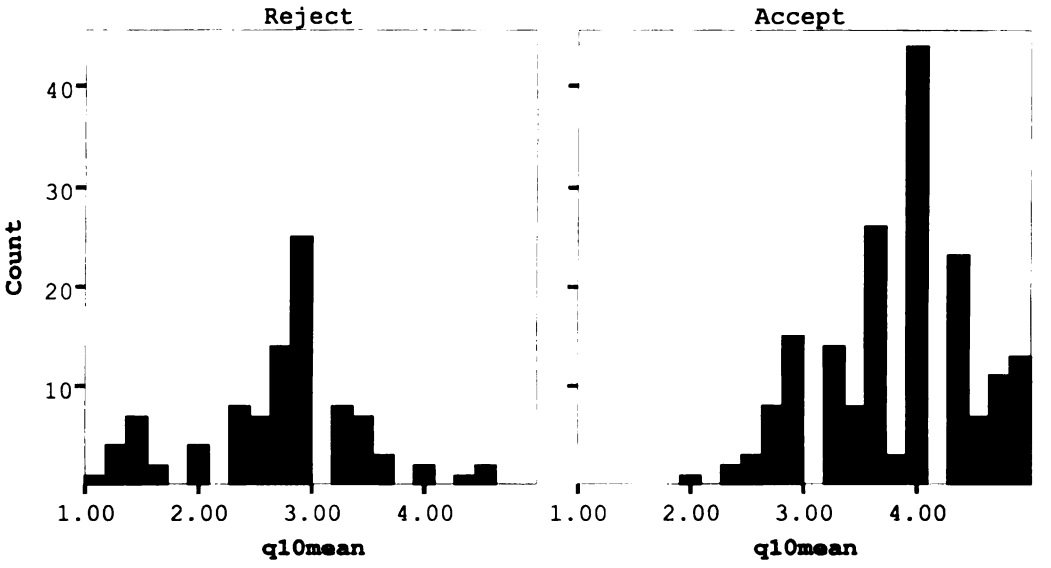

Figure A19. Accept and reject, Division 7, Year 1.

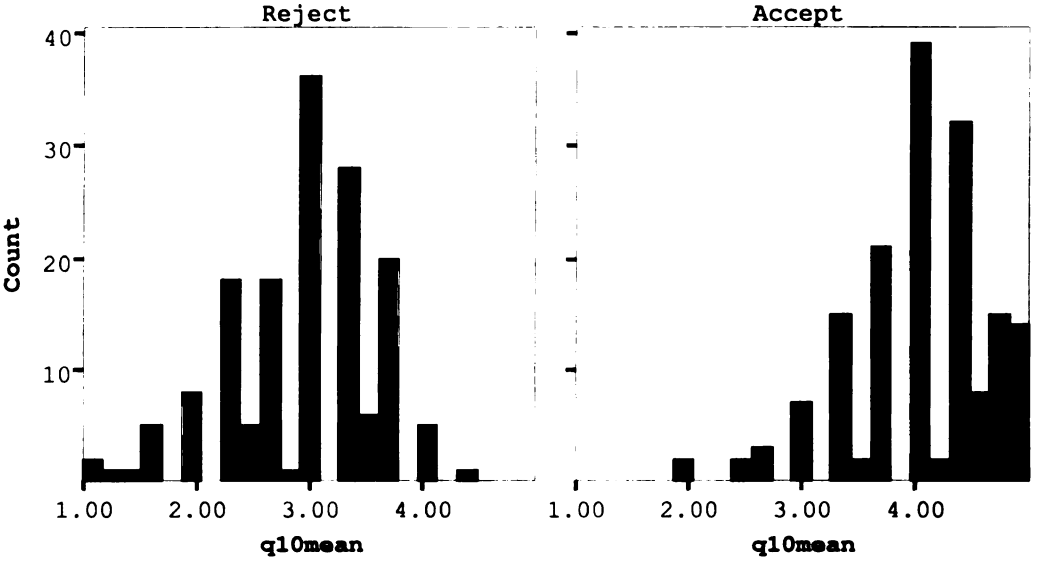

Figure A20. Accept and reject, Division 7, Year 2.

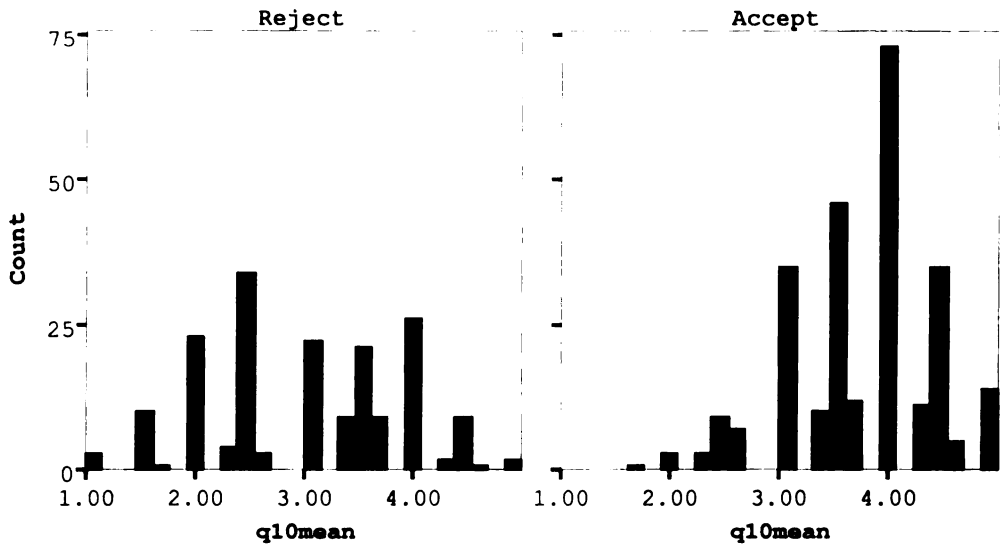

Figure A21. Accept and reject, Division 7, Year 3.

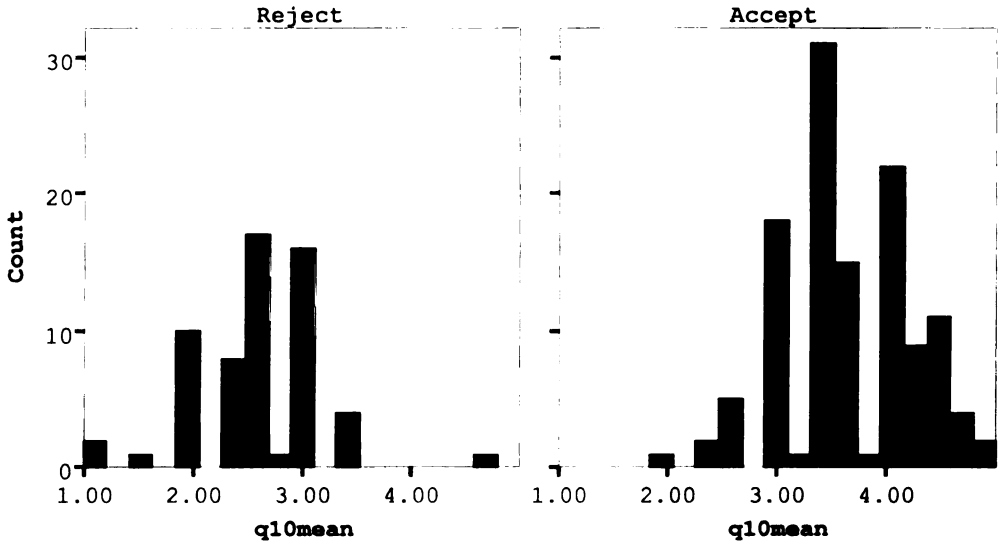

Figure A22. Accept and reject, Division 8, Year 1.

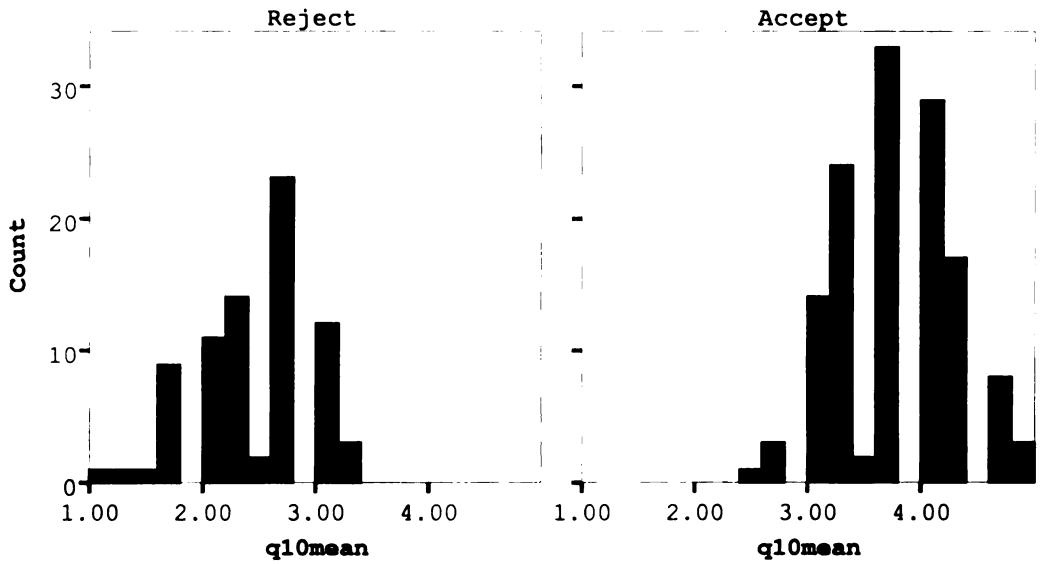

Figure A23. Accept and reject, Division 8, Year 2.

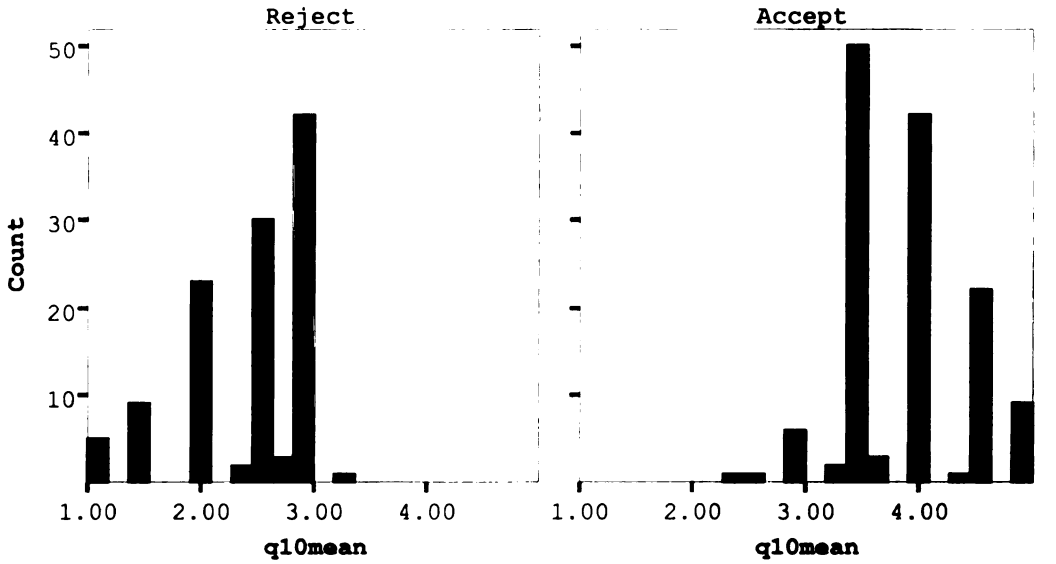

Figure A24. Accept and reject, Division 8, Year 3.

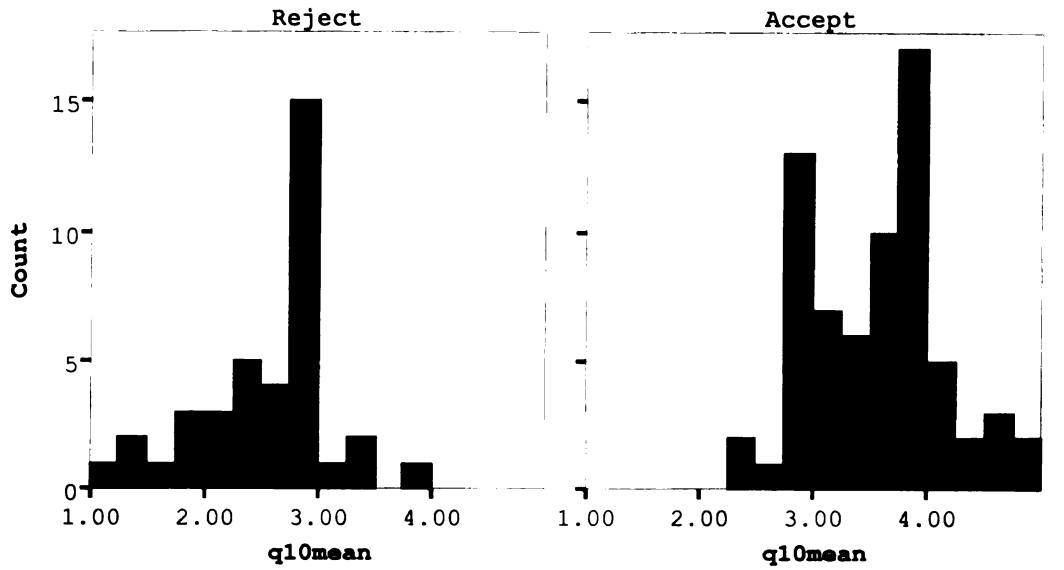

Figure A25. Accept and reject, Division 9, Year 1.

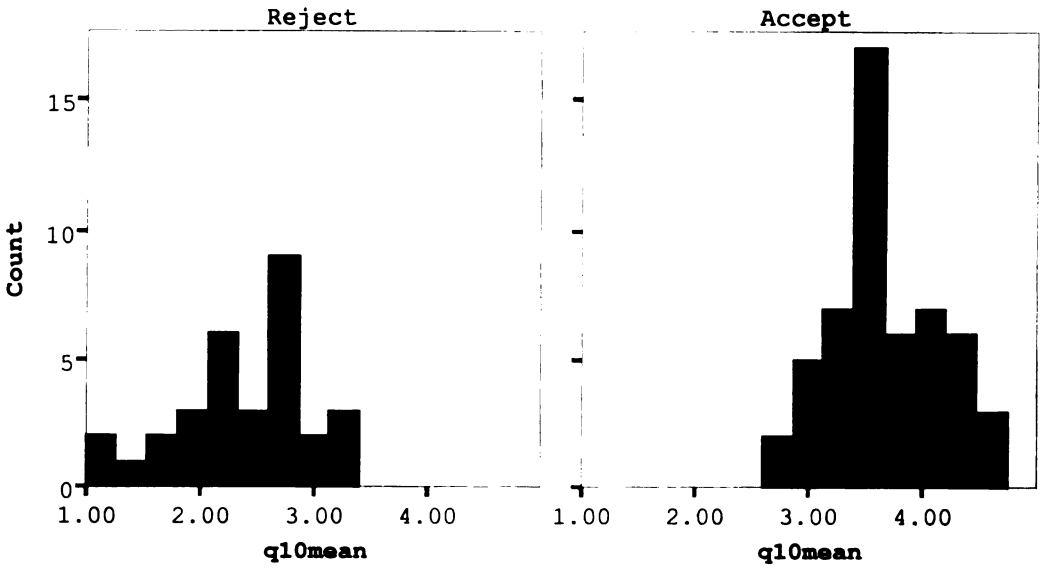

Figure A26. Accept and reject, Division 9, Year 2.

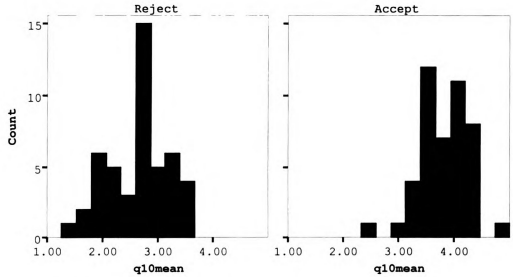

Figure A27. Accept and reject, Division 9, Year 3.

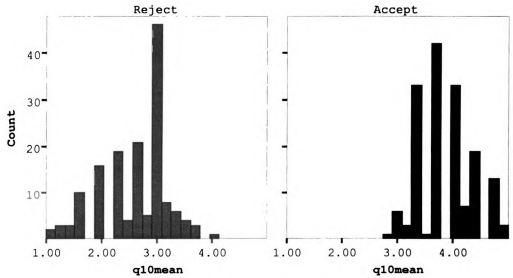

Figure A28. Accept and reject, Division 10, Year 1.

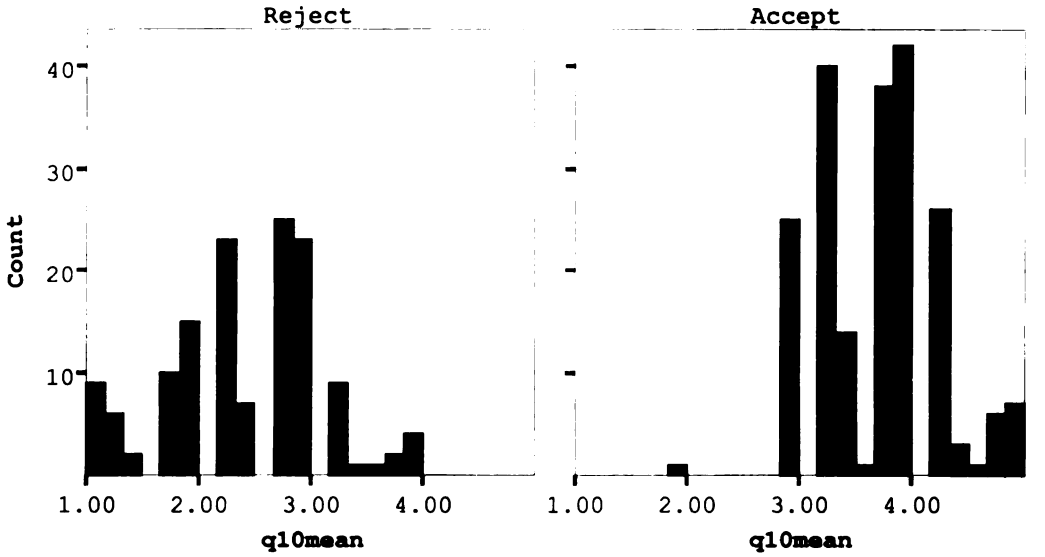

Figure A29. Accept and reject, Division 10, Year 2.

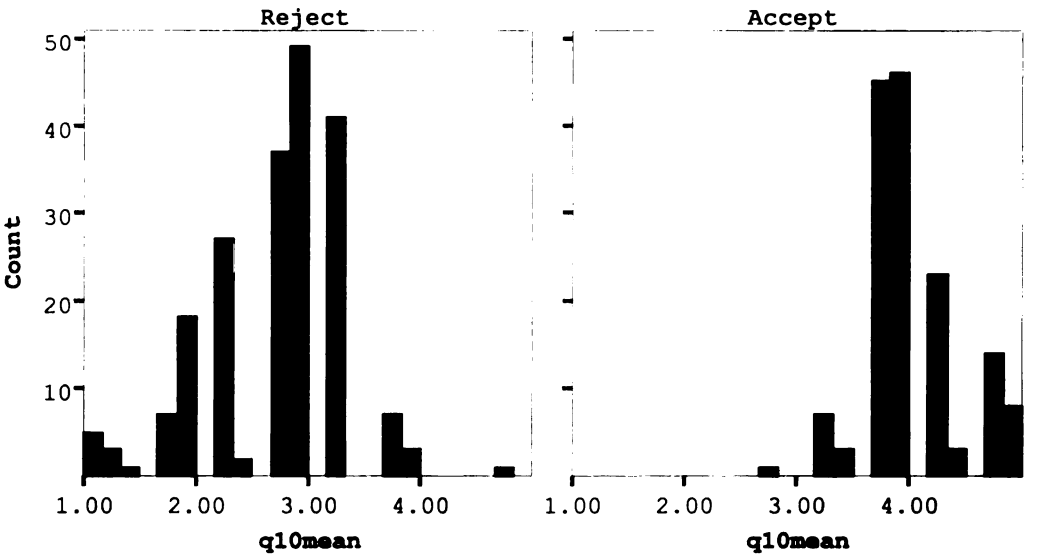

Figure A30. Accept and reject, Division 10, Year 3.

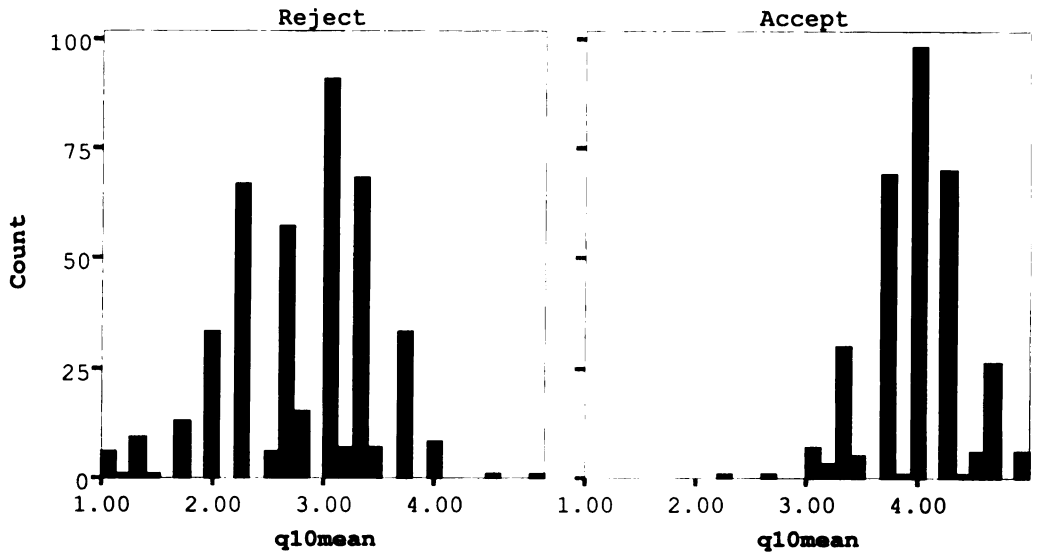

Figure A31. Accept and reject, Division 11, Year 1.

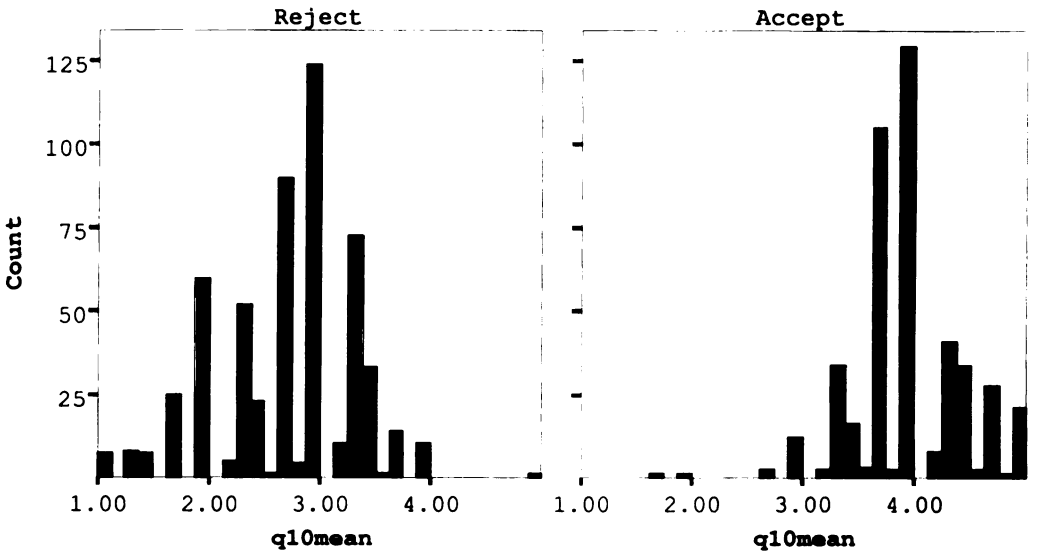

Figure A32. Accept and reject, Division 11, Year 2.

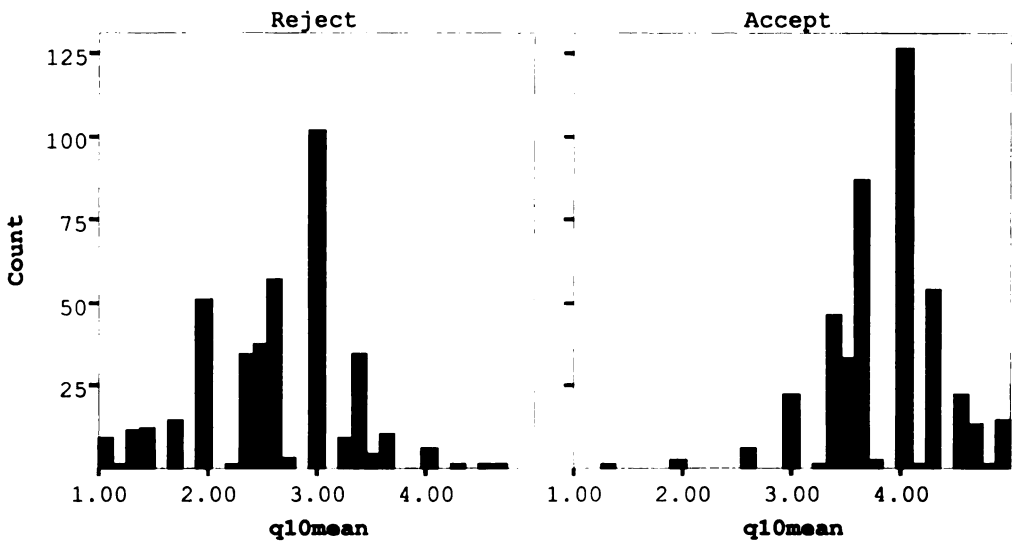

Figure A33. Accept and reject, Division 11, Year 3.

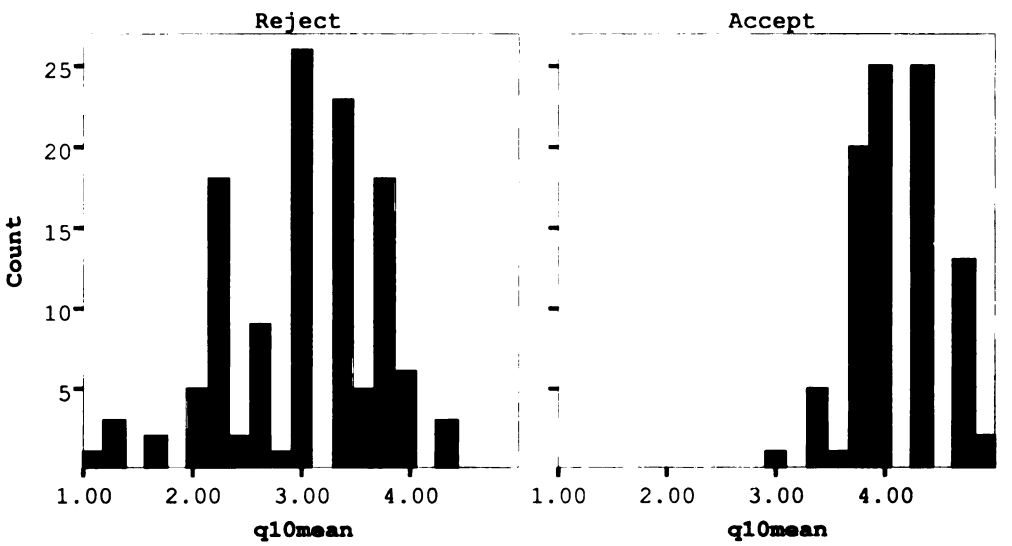

Figure A34. Accept and reject, Division 12, Year 1.

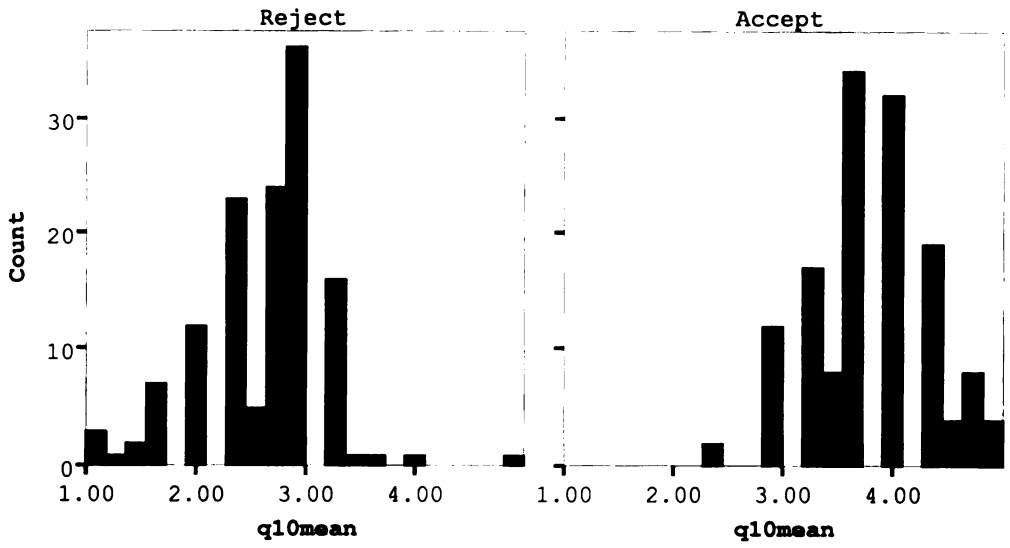

Figure A35. Accept and reject, Division 12, Year 2.

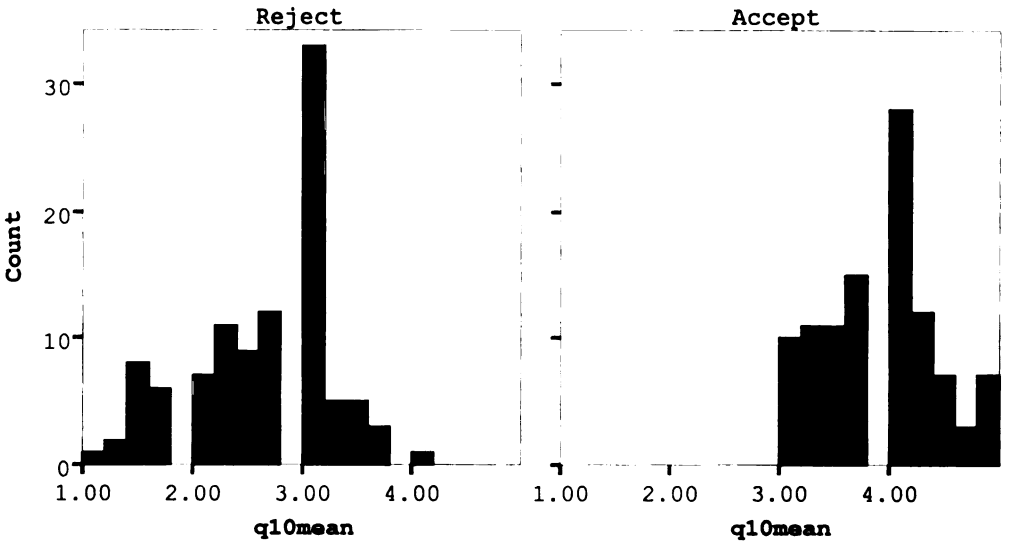

Figure A36. Accept and reject, Division 12, Year 3.

REFERENCES

Adair, R. (1982). "A Physics Editor Comments on Peters and Ceci's Peer-Review Study"." The Behavioral and Brain Sciences **5**(2): 196.

Bond, T. G. and C. M. Fox (2001). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, Lawrence Erlbaum Associates.

Brennan, R. L. (2001). Generalizability Theory. New York, Springer.

Strengthening Peer Review in Federal Agencies That Support Education Research (2004)
Center for Education.

Chubin, D. E. and E. J. Hackett (1990). Peerless Science. Albany, State University of New York.

Hackett, E. and D. Chubin (2003). Peer Review for the 21st Century. Washington, D.C., National Research Council.

Kaplan, D. (1995). "How to Fix Peer Review", The Scientist, Vol. 19, Issue 1, Jun. 6. p. 10.

Lagemann, E. (2002). An Elusive Science: The Troubling History of Education Research. Chicago, University of Chicago Press.

Lindsey, D. (1976). "Distinction, Achievement, and Editorial Board Membership." American Psychologist **31**(11): 799-804.

National Research Council. (2004). Strengthening Peer Review in Federal Agencies that Support Education Research. Committee on Research in Education. L. Towne, J. M. Fletcher, and L. L. Wise, Eds. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Newman, S. (1966). "Improving the evaluation of submitted manuscripts." The American Psychologist **21**(10): 980-981.

Peters, D. P. and S. J. Ceci (1982). "Peer-review practices of psychological journals: The fate of published articles, submitted again." Behavioral & Brain Sciences **5**(2): 187-255.

Rennie, Drummond, et al. (1989). "The International Congress on Peer Review in Biomedical Publication." The Journal of the American Medical Association **261**(5): 749.

Rennie, D. (2002). "Fourth International Congress on Peer Review in Biomedical Publication." JAMA **287**(21): 2759-2760.

Soffer, A. (1980). "The Unique Role of Peer Review Journals." Chest **78**(4): 547-48.

Speck, B. W. (1993). Publication Peer Review: An annotated bibliography. Westport, Greenwood Press.

Weller, A. C. (2002). Editorial Peer Review: Its strengths and weaknesses. Medford, Information Today.

Wilson, E. (1979). "Comments from a Servant of the Scattered Family." Contemporary Sociology **8**(6): 804-08.

Wolff, W. (1970). "A Study of Criteria for Journal Manuscripts." American Psychologist **25**(7): 636-39.

Zuckerman, H. and R. Merton (1971). "Patterns of Evaluation in Science." Minerva **9**: 66-100.