

Variability of Referees' Ratings of Conference Papers

by Weichao Ma, Dorene Mboya, and Henry M. Walker

Department of Mathematics and Computer Science, Grinnell College, Grinnell, IA 50112

Draft 1.3: January 5, 2001

Abstract

The authors developed an on-line system to automate the paper submission and reviewing process for SIGCSE 2000. This system provided a mechanism for a statistical study of the ratings of papers by different referees. To compile sufficient sample data, 10 papers were sent to about 100 referees each, in addition to the normal reviewing of the 220 submitted papers. This paper reports the results of this work, based on 1917 reviews completed by 482 referees. The data support correlations of a paper's overall rating with internal factors (technical content, writing quality, originality, significance) and external factors (referee gender, nationality, familiarity with topic, etc.). The study also shows the effect of group reviews (similar to NSF panels), and the divergence of opinions between authors and referees concerning a paper's subject.

1. Context and Motivation for Project

As with many many conferences, SIGCSE Symposia depend upon contributed papers for various technical sessions. The selection of these papers depends upon a reviewing process, during which the papers are sent to multiple referees for reviews and rankings. Such a process raises natural questions about the likely variance of such rankings from different reviewers.

1. To what extent do some reviewers give consistently high or consistently low ratings?
2. Are ratings affected by such factors as referee gender, nationality, or expertise, or by the paper's format?
3. Do authors of one paper give consistently high or low ratings to other papers?
4. How does an overall rating correlate with subscores in areas of technical content, writing quality, originality, and significance?
5. How does a joint review by a large committee working collaboratively affect the ratings given by that committee?

For statistical validity, a study addressing such questions requires gathering ratings by numerous referees for diverse papers.

SIGCSE 2000 provided the opportunity for such data collection, when Symposium Co-Chairs, Nell Dale and Boots Cassel, and Program Chair (and co-author), Henry Walker, decided to expand an on-line review-reporting process begun for SIGCSE 1999 by Program Chair, Robert Noonan. The desired system would allow on-line paper submissions in html format (providing world-wide access to referees), restricted access to these papers by authors and authorized referees, and collection of reviews through a Web-based form. Unfortunately, as the new system also should interface appropriately with an existing MS Access database, it was concluded that existing software did not meet these needs. Thus, the co-authors undertook to develop a new system, with the support of the Fund for Excellence for Capstone Projects at Grinnell College.

2. Three Phases of Development

Work on this project proceeded in three phases:

- I. *Summer, 1999*: Software development and logical design for the variability study,
- II. *Fall, 1999*: Reviewing of papers and data collection, and
- III. *Winter, Spring 2000*: Statistical analysis.

This multi-phase schedule allowed authors and referees to use software as needed, while also providing time for the development and review of multiple prototypes for later parts of the process. For example, authors could submit papers once paper-submission forms were completed, although development continued on the reviewing form and the design of the statistical study.

3. Phase I: Development of a Web-based System and Design of Variability Study

For electronically-submitted papers, the on-line system records needed author, title, and subject information, downloads needed files (in html format with .gif or .jpeg figures), and adds a security header to prevent unauthorized access. For hardcopy submissions, the software maintains records of authors, titles, and subjects. The system also provided automated tools to assign papers to referees, distribute on-line papers, and receive reviews electronically.

Table 1: Review Ratings Definitions and Actual Ratings Received

Rating	Descriptor or Anchor	Quality Range	Commentary	Actual Ratings	
				Number	%
6	Exceptional	Top 5%	Likely to be among top 10 papers at conference	64	3.3%
5	Outstanding	Next 15%	Above average for symposium papers	357	18.6%
4	Very Good	Next 20%	Comparable to many symposium papers	506	31.6%
3	Average	Middle 20%	Average symposium papers	458	23.9%
2	Below Average	Lower 30%	Correct but not too interesting	423	22.1%
1	Deficient	Bottom 10%	Contains serious errors or deficiencies	109	5.7%

Work on a revised design for the reviewing form included the development and review of several prototypes. The left four columns of Table 1 show the final reviewing scale, as presented to referees. The 6-point scale prevented referees from being completely neutral about a paper. The form requested an “overall rating” as well as ratings for “technical content”, “organization and writing style”, “originality” and “significance”. The form also provided boxes for referee comments to amplify ratings in each category, and a mechanism for referees to indicate their familiarity with a paper’s subject.

4. Phase II: Reviewing/Data Collection

Expanding somewhat the reviewing process for past SIGCSE Symposia, each of the 219 papers was sent to 5 referees (a withdrawn paper was not reviewed), and 10 selected papers were sent to about 100 referees each. While not all assigned reviews were completed, all papers had at least 3 completed reviews; 46 papers had exactly 3 reviews; 96 papers had exactly 4 reviews; and 77 papers had exactly 5 reviews; and 1009 reviews were received for the 10 selected papers. Also, 6 groups of 4-8 people volunteered to collaborate in reviewing papers (following a format not unlike an NSF panel). Altogether, 482 referees submitted 1917 reviews.

5. Phase III: Analysis

Ratings’ analysis includes frequency and variability measures, internal factors, external factors, and subject classification. Since variability measures require many reviews for a single paper, those results are based on the 100 or so reviews for each of 10 papers. The other factors involve correlations and counts across papers, and the entire reviewing database provides an appropriate basis for these results.

Overall Frequencies

The review form provided some guidance concerning the meaning of values in the ratings scale, and the right-hand two columns of Table 1 indicate the number and percentage of overall ratings in each category. While referees gave somewhat more 4’s and fewer 2’s than originally anticipated, the number of ratings for the other categories is within 4% of the rough guidelines.

Variability

Table 2 shows ratings’ counts, means, medians, and standard deviations for the overall rating of each of the 10 sample papers. Based on a ratings scale from 1 (low) to 6 (high), the papers are listed in order of decreasing mean to aid subsequent analysis.

Paper	Number of Responses	Mean	Median	Std.Dev.	Scores					
					1	2	3	4	5	6
1	95	4.50	5	0.90	0	3	7	35	40	10
2	110	4.48	5	0.97	0	4	12	34	47	13
3	99	3.89	4	1.12	3	8	21	37	25	5
4	107	3.76	4	1.16	3	14	23	37	26	4
5	102	3.57	4	1.22	4	18	24	33	18	5
6	111	3.35	3	1.07	1	27	34	31	17	1
7	104	3.25	3	1.10	5	21	36	28	13	1
8	116	2.84	3	1.02	7	40	42	21	4	2
9	95	2.78	3	1.19	11	36	21	17	10	0
10	109	1.89	2	0.82	37	53	13	6	0	0

Table 2: Frequency, Means, and Variability

Table 3: Correlation Coefficients for Internal Factors

Paper	Number of Responses	Regression Constant	Technical Content	Organization and Writing	Originality	Significance
All	1917	-0.50	0.31	0.20	0.18	0.42
1	95	-0.18	0.23	0.23	<i>0.14</i>	0.46
2	110	-0.01	0.39	0.16	0.10	0.42
3	99	-0.23	0.23	0.23	0.22	0.36
4	107	-0.44	0.31	0.26	0.04	0.52
5	102	-0.48	0.33	<i>0.11</i>	0.30	0.39
6	111	-0.16	0.23	0.03	0.26	0.53
7	104	-0.42	0.34	0.06	0.27	0.46
8	116	-0.36	0.34	0.21	0.15	0.36
9	95	-0.34	0.32	0.15	0.08	0.49
10	109	-0.09	0.22	0.29	<i>0.13</i>	0.28

Table 2 suggests several variability results.

- Standard deviation of overall ratings for each sample paper was between 0.8 and 1.2 (for the 6-point ratings scale),
- Standard deviation was somewhat higher (1.0-1.2) for papers with middle scores,
- Standard deviation was lower (about 0.8) for papers with high or low scores,
- While scores may range rather widely, papers scoring at the extremes have either no low or no high scores.
- The six papers with intermediate median scores had both a few 1's and a few 6's; however, the combined number of extreme low (1) and high (6) scores never exceeded 9% for these middle-scoring papers.

Internal Factors Affecting Ratings

Referees gave each paper an overall rating plus separate ratings for technical content, organization and writing, originality, and significance. A regression analysis of overall ratings versus category ratings yields the coefficients shown in Table 3. In the table, **values in bold face** are statistically significant (P -value under 0.05), and *values in italics* have marginal significance (P -value between 0.05 and 0.1). Other values are not significant (P -value above 0.1).

Table 3, together with analysis of standard deviations, suggests several conclusions:

- Paper significance is the most important factor contributing to the overall rating (regression coefficient about 0.4),
- Technical content is next most important (regression coefficient about 0.3),

- Originality and writing have about equal importance (coefficient about 0.2 each).
- Differences from significance to content to originality and writing are statistically significant.
- Differences between originality and writing are not statistically significant.

External Factors Affecting Ratings

Unbiased reviewing is vital in a process where paper acceptance depends in part on referee ratings. Thus, this analysis of SIGCSE 2000 reviews also investigated the effect on ratings of such external factors as the gender of the referee, whether the referee lived in the United States or outside the United States, whether the referee was also the author of another paper submitted to the Symposium, whether the referee worked individually or worked in a group (following the NSF panel model), and how familiar the referee was with the material covered in the paper. In addition, the study considered the extent to which the form of the submitted paper (hardcopy or electronic) affected ratings.

To test the potential effect of each factor, a separate variable was defined for that factor. For example, to study possible gender bias, a field was added to each referee record, with value 1 for females and 0 for males. A similar process was followed for other factors, except that variables were added for both high and low familiarity with the subject (as reported by the referee). Table 4 shows the resulting correlation, where a coefficient indicates the effect of the corresponding factor.

Table 4: Possible External Factors

	N	Constant	Female (Male)	International (Domestic)	Ref.Is Not Author	Author Ref. in Group (not in group)	Familiarity High Low	Hardcopy (E-copy)
All	1917	3.47	0.15	-0.21	-0.03	-0.42	-0.10 -0.23	-0.10
1	95	4.40	0.42	-0.77	-0.31	-0.3	0.33 0.08	N/A
2	110	4.40	0.01	-0.43	-0.04	0.79	0.11 0.03	N/A
3	99	4.18	0.03	-0.27	0.13	-0.74	-0.29 -0.24	N/A
4	107	3.77	0.24	-0.06	0.28	-0.20	-0.23 0.28	N/A
5	102	3.39	0.14	-0.52	-0.18	0.08	0.07 0.69	N/A
6	111	3.53	-0.06	-0.02	0.08	-0.67	-0.07 —	N/A
7	104	3.49	0.07	0.09	0.08	-0.77	-0.25 0.77	N/A
8	116	3.00	0.37	-0.053	0.37	-0.62	-0.28 -0.25	N/A
9	95	3.23	0.27	-0.88	-0.62	-1.13	-0.34 0.17	N/A
10	109	2.06	0.07	0.08	-0.40	-0.51	-0.10 0.20	N/A

The column headers for Table 4 indicate the factor value coded as 1 (with the value 0 in parentheses). Thus, a female referee had a code 1 for gender, while a male had code 0. For this variable, the coefficient shows the extent to which females gave higher ratings than males. As in the previous tables, numbers in bold face are statistically significant (with a t-test smaller than 0.05). Statistical significance for all papers combined is likely due to the large numbers.

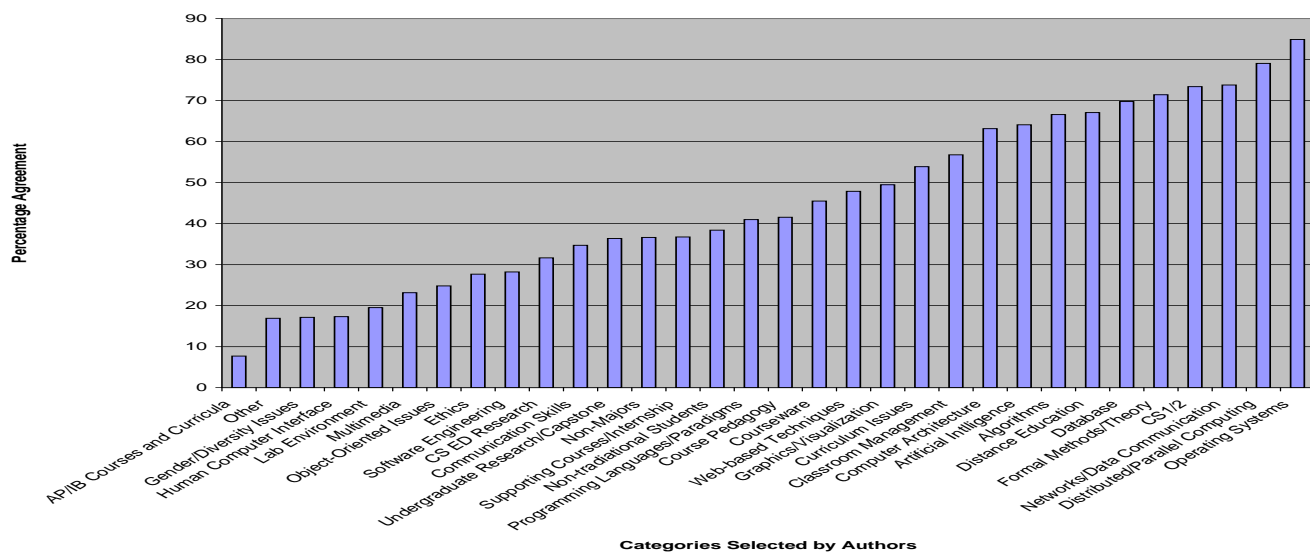
These results suggest that few systemic referee factors affect paper ratings. Specifically, the following factors are **NOT** statistically significant in contributing to overall paper ratings: referee gender, referee’s country (U. S. versus non-U.S.), familiarity of referee with subject, or paper format (electronic submission versus hardcopy). However, ratings from referees in groups were lower (by about 0.5) than ratings

from those who were not in groups, and this difference is statistically significant.

Subject Classification Of Papers

As part of the paper-submission process, authors indicate the subject(s) of their papers, and this categorization helps guide the assignment of referees. Referees then also indicate the subject(s) they think are appropriate for papers. For SIGCSE 2000, authors identified between 1 and 9 subjects for their papers, and on average, authors placed a paper in 4 categories. For each of these identified subjects, Figure 1 shows the percentage of times a referee identifies that same subject. Thus, in the cases when authors indicated that a paper related to Advanced Placement or International Baccalaureate, referees agreed with categorization only about 9% of the time.

Figure 1: Percentage of Referees Agreeing with Author on a Paper’s Subject



A second approach for determining to what extent authors and referees agree upon the subject(s) of a paper is to look at the various categories specified for a paper by an author and then to determine the percentage of these categories also specified by the referees. Figure 2 gives the results of this analysis.

Both Figures 1 and 2 indicate that the categorization process often depends significantly upon who does it; authors and referees often disagree about the appropriate categories for a paper. Careful review of these data yield the following results:

- a. Categories with under 30% agreement between author and referee: AP/IB, Other, Gender/Diversity Issues, HCI, Lab Environment, Multimedia, OO Issues, Ethics, Software Engineering
- b. Categories with over 70% agreement: CS 1/2, Formal Methods/Theory, Networks/Data Communication, Distributed/Parallel Computing, Operating Systems
- c. About 20% of referees agreed completely with all subject categories given by authors, but the majority agreed with fewer than half of the subject categories; 31.2% of referees agreed with fewer than 30% of those categories.

6. Conclusions and Future Work

This paper presents information on the natural variance of overall paper ratings and provides insights about both internal and external factors that may affect these ratings. Much to the relief of program committees, paper ratings do not seem to be influenced by external factors (e.g., gender, nationality, whether the referee also authored another paper, or even familiarity with the subject). Working in groups, however, does seem to have a definite negative effect on overall ratings.

While the results arose specifically from reviews for SIGCSE 2000, the conclusions may provide insights for authors and program committees more generally and suggest an approach for additional studies of ratings for other conferences. The authors hope that others will undertake similar studies of ratings variability for papers submitted to other conferences.

7. Acknowledgments

The success of this project depended upon the support and contributions of many individuals: Symposium Co-Chairs, Boots Cassel and Nell

Dale, provided steady inspiration, support, and guidance; Dean James Swartz provided funding for summer stipends through Grinnell College's Fund for Excellence; Grinnell College's Computer Services, especially William Francis and Mark Miller, provided hardware and software support; Wayne Twitchell and Theresa Walker provided extensive guidance and technical support for Web-design and software development; Carol Trosset and Laura Sinnett consulted on survey development; Thomas Moore served as statistical consultant; the CS Education Group at the University of Texas at Austin, various members of the SIGCSE 2000 Committee, and Joseph Kmoch provided extensive feedback on draft forms and scripts; Karen Thomson handled many reformatting tasks for paper publication; and most importantly, hundreds of referees for SIGCSE 2000 completed 1917 reviews of 220 papers as part of the reviewing process.

The authors are deeply indebted to all of these people for their efforts, insights, support, and helpful feedback. The project relied heavily on the work of each of these people, and all contributed in important ways to its success.

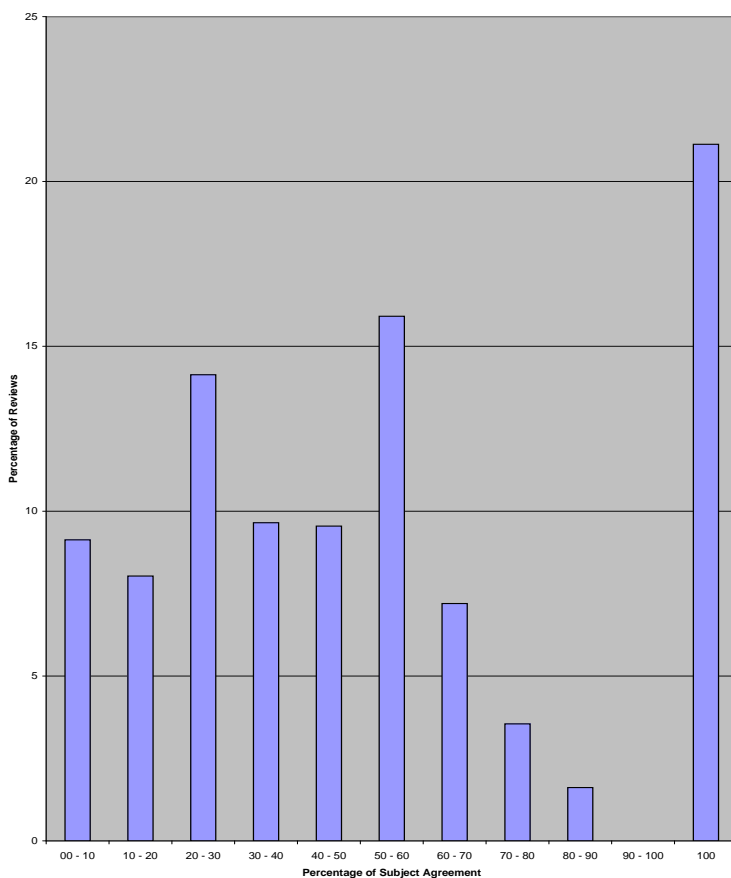


Figure 2: Percentage of Author-Designated Subjects Also Given By Referees