DOCUMENT RESUME

ED 270 458 TM 860 307

AUTHOR Cason, Carolyn L.; And Others
TITLE Reviewer Standards in Division I Program Selection.
PUB DATE Apr 86
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Associaton (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Conference Papers; Data Analysis; Educational Research; *Evaluation Criteria; *Examiners; *Interrater Reliability; Measurement Techniques; Models; *Quality Control; *Rating Scales; Reliability; Research Reports; Standards; Test Validity
IDENTIFIERS *Reviewers

ABSTRACT

Cason and Cason's model of performance rating was used to determine the extent to which variation in reviewer standards affected the reliability and validity of the program review process used to select papers for inclusion in the annual program. Data analyzed were the overall recommendation for acceptance and ratings on seven quality criteria from each reviewer on each paper proposal in 1983, 1985, and 1986. The Casons' model fit each year's data. Significant rater stringency variance was found for each of the three years. Rater stringency persisted up to three years providing strong construct validation for the model. Removing the rater stringency effect improved reliabilities from .768, .722, and .739 to .813, .790 and .790. Construct validities also improved. Had adjusted ratings been used in 1986, up to 6 of the 35 papers accepted would have been rejected. There were no significant differences in mean rater standards year to year; however, mean paper proposal quality was sharply lower in 1985. In all years, mean paper quality of accepted proposals was significantly better than that of rejected proposals. Access to adjusted ratings at the time of the selection decision would ease the committee's task and probably improve the quality of its decisions. (Author/PN)

Reviewer Standards in Division I Program Selection*

Carolyn L. Cason
Gerald J. Cason
University of Arkansas for Medical Sciences
and
Frank T. Stritter
University of North Carolina Chapel Hill

Address correspondence to:

Carolyn L. Cason
UAMS-CON-529
4301 West Markham
Little Rock, Arkansas 72205
(501) 661-5163

*Presented at the annual meeting of the American Educational Research Association, San Francisco, April 1986.

We wish to thank the chairs and members of the 1983, 1985, and 1986 Program Committees for making the data available. Providing this kind of data for external analysis requires a rare and courageous commitment to scholarly and research cannons.

Dick Calkins of Technology, Inc., Houston, Texas provided useful statistical assistance.

The staff of the Academic Support Center, UAMS provided special programming and data processing support.

SHORT ABSTRACT

Cason and Cason's model of performance rating was used to determine the extent to which variation in reviewer standards affected the reliability and validity of the program review process used to select papers for inclusion in the annual program. Data analyzed were the overall recommendation for acceptance and ratings on seven quality criteria from each reviewer on each paper proposal in 1983 (NR=87, NP=120), 1985 (NR=86, NP=100), and 1986 (NR=82, NP=115). The Casons' model fit each year's data: $R > .756$; $p < .00001$). Significant rater stringency variance was found for each of the three years. Rater stringency persisted up to three years providing strong construct validation for the model. Removing the rater stringency effect improved reliabilities from .768, .722, and .739 to .813, .790 and .790. Construct validities also improved. Had adjusted ratings been used in 1986, up to 6 of the 35 papers accepted would have been rejected. There were no significant differences in mean rater standards year to year; however, mean paper proposal quality was sharply lower in 1985. In all years, mean paper quality of accepted proposals was significantly better than that of rejected proposals. Access to adjusted ratings at the time of the selection decision would ease the committee's task and probably improve the quality of its decisions.

Reviewer Standards in Division I Program Selection

Carolyn L. Cason
Gerald J. Cason
University of Arkansas for Medical Sciences
and
Frank T. Stritter
University of North Carolina Chapel Hill

Paper proposals selected for presentation at the Division I program of the annual AERA meeting are intended to address research issues of interest to the division membership and reflect careful and sound application of scientific method. These presentations communicate new scientific knowledge while at the same time providing a mechanism to formally acknowledge those who made the contribution. Thus what is selected for presentation becomes a matter of importance to both the body of scientific information and the individual researcher's professional career.

In Division I the Program Committee decides which paper proposals will be accepted for presentation. In general terms, the objective of the Program Committee is to accept the best proposed papers from among those meeting at least minimal standards of scholarship. The committee is aided in its decision making by multiple reviews of each paper proposal completed by Division I members who have volunteered to do so. To the extent that reviewers use the same or similar standards in making their reviews, such reviews aid the Program Committee in selecting quality paper proposals. However, the assumption of similar standards is suspect as reviewers have different backgrounds, different experiences, and different levels and areas of expertise in research, although use of multiple reviews of a single paper proposal may attenuate the effects of variation in standards among reviewers (Ebel, 1951; Stanley, 1961). The impact of such variation among reviewers on paper proposal selection has been largely unevaluated.

Even though peer review is used as the basis for making many highly important decisions about scientific products (e.g., in promotion, funding, publishing), it has received only very limited attention as a topic of research. Marsh and Ball's (1981) study of manuscripts submitted to the Journal of Educational Psychology is the only one which has examined reviewers' ratings for variation in reviewer standards and the impact of corrections for reviewer bias on reviewer agreement of manuscript quality. Marsh and Ball's study is particularly relevant to the issue of variation of rater standards and their impact in the Division I review process because their study deals with a similar substantive research area, of equal breadth and complexity, where the same types and varieties of research methods are used, and a similar peer review process exists. They found significant systematic variation in mean ratings of different reviewers with single reviewer reliabilities of .34; somewhat above the middle of those reported in the psychological and sociological literature which ranged from .08 to .54. Corrections for response bias (variation in reviewer standards) did not yield statisically significant improvements in single reviewer reliabilty (.35). Although Marsh and Ball concluded that the observed variation in ratings was primarily a function of manuscript quality rather than reviewer bias, their results suggest that correction for reviewer bias was inadequate: manuscript quality may have been confounded with reviewer bias. Had Marsh and Ball used a more powerful method to test and correct for the presence of reviewer bias, they might have found both a significant reviewer effect and a significant improvement in inter-reviewer agreement arising from correction for reviewer bias. A major obstacle which they confronted was having very little data for the analysis: only two

external reviews per manuscript. The small amount of data available per manuscript may explain why their expectation of a significant reviewer effect (personal communication from Marsh) was not supported by their results.

The major objective of this study was to determine the extent to which variation in standards exists among Division I program proposal reviewers and the extent of the impact of such variation (if any) on the acceptance of papers proposed for the program. A second objective was to determine the degree to which Cason and Cason's (1984) model was appropriate to this kind of rating data and facilitated reaching the primary objective.

Theory

Cason and Cason (1984; G. Cason et al, 1983; C. Cason et al, 1983) have found that use of their model on rating data of a similar type achieved good fit and resulted in improved reliability and validity when ratings were adjusted for variation in rater standards. Their results suggest that their model may be appropriate for examining variation in rater standards in other settings when the performance to be evaluated is complex. Cason and Cason's (1984) model provides the basis for answering the questions (a) are there differences in standards used by different raters, and (b) do such differences (if they exist) produce significantly different average observed ratings from those expected were there no differences in standards. As illustrated in Figure 1, Cason and Cason's theory posits that "the rating received by a subject is a function of that subject's true ability and the rater's characteristics including the rater's resolving power, sensitivity, stringency, and effective rating floor and ceiling" (Cason & Cason, 1984, p. 223). The Casons' simplified model of their performance rating theory accounts for all systematic variation in performance ratings exclusively by variation in rater stringency and subject ability as illustrated in Figure 2. In the Casons' model the expected subject rating (ESR), measured as a percent of the maximum rating, is a function of the difference, z, between the rater's stringency (i.e., value associated with the Rater Reference Point or RRP) and the subject's ability (i.e., value associated with the Subject Ability Point or SAP). See Figure 3 for an illustration. In previous research this relationship was modified by an arbitrary scaling factor (SF=100): $z = (SAP - RRP)/SF$. The theoretically postulated curvilinear relationship between z and the expected subject rating (ESR) has been stipulated as the unit-normal ogive. Thus, the ESR (in percent) for a given z is equal to the proportion of area under the normal curve below z; that is, p(z) times 100: $ESR = p(z) \times 100$. This is a deterministic, not probabilistic relationship.

Data Source and Method

The ratings given by individual reviewers to paper proposals submitted to the AERA Division I Program Committees for the 1983, 1985, and 1986 programs constitute the data. The ratings used were the overall recommendation for acceptance. The original scale was defined as 1=accept; 2=accept with reservation; 3=accept only if room permits; and 4=reject. For the purposes of this analysis the scale was reversed and converted to percent units: 100% = accept, 67% = accept with reservation, 33% = accept only if room permits, and 0% = reject. Analyses were completed on reviewer acceptability ratings of 120 (1983), 100 (1985), and 115 (1986) paper proposals. Although each and every proposal was sent for review to 4 raters, both 1985 and 1986 had missing data, i.e, no rating was noted on the work sheets provided to the researchers by the Program Committee on some proposal-reviewer combinations. In 1983, Program Committee members each reviewed from 19 to 23 proposals; other reviewers from 2 to 6. In 1985, Program Committee members each reviewed from 10 to

18; others reviewed from 1 to 9 (median = 3). In 1986, Program Committee members each reviewed from 13 to 17 proposals, other reviewers from 1 to 9 (median = 4).

On each proposal reviewed for the 1986 program, reviewer evaluations of proposal quality on each of the seven criteria, as well as their recommendation on disposition (acceptability rating) were analyzed. To conduct item level quantitative analyses, the extreme left end of the scale (see Figure 4 for an illustration of the reviewer inventory) was assigned a value of 1 and the right most end a value of 5. The "+" marks on the scale were associated with the consecutive numbers 2, 3, and 4. For the criterion "Clarity of Summary" obsure=1 and clear=5. In those instances in which the reviewer's mark fell between "+"s, the numeric value assigned was that which represented the "+" closest to the reviewer's mark, in the judgement of the researcher recoding the data onto machine scannable answer sheets. See Figure 5 for an example of the machine scannable answer sheets. ith the recoding, the behaviorally anchored scale was treated as a Likert one to facilitate the analysis and reporting process. This item level data was processed through the UAMS Objective Test Scoring and Performance Rating (OTS-PR) system. A full set of standard reports was obtained including those providing the rated quality of the proposals and the quality of the rating inventory as reflected in intra-class correlation measures of inter-rater consistency.

Using regression analyses (Ward & Jennings, 1973) which were based on an improved version of the procedures described in technical detail in Cason and Cason (1985), RRPs and SAPs were estimated using ratings given by individual reviewers to paper proposals. The data from each year were analyzed separately. The improved method used here differed from that reported in 1985 in the following ways. Estimation of parameters for a given data set involved two successive regression analyses upon the same data. In each, the model followed the same general form given in 1985. However, in the first the criterion contained percent scores. In the second, the criterion contained the z transforms of the expected values from the first regression analysis. This approach provided better approximations of least squares solutions for the theoretical model.

Results

Descriptive statistics on observed ratings are given in Table 1. On the face of it, the observed ratings are so consistent with respect to mean and standard deviation it would seem tempting to assume that both average proposal quality and average rater standards remained consistent between 1983 and 1986. This turns out not to be the case.

Table 1. Observed Acceptability Ratings of 1983, 1985, and 1986 Proposals

	1983	1985	1986
Mean	48.24	50.10	48.73
Standard Deviation	27.43	27.56	27.47

As can be seen in Table 2, Cason and Cason's model obtained a non-chance fit to the 1983, 1985, and 1986 paper proposal reviewer acceptablity ratings. While there was no global, significant rater "main effect" in 1983 ($F=0.79; df=86,275; p=0.89$), there were significant rater effects in 1985 ($F=1.48$; $df=85,210$; $p=0.013$) and 1986 ($F=1.42$; $df=81, 258$; $p=0.021$). For details on the way in which these effects were tested, see the description of the statistical models provided in Cason and Cason

(1984). Even though a significant rater stringency effect was not observed in the 1983 data there could still be variation in standards used by individual reviewers. The absence of a significant rater effect indicated that the mean stringency of the group of reviewers who rated each proposal was statistically equal to (i.e., not different from) the mean across all reviewers. Statistically significant overall fit of the model is prerequisite to establishment of the presence and importance of both of the formal model constructs: stringency and proposal quality. The significant rater effects in 1985 and 1986, considered in conjunction with proportions of variance accounted for by rater stringency and proposal quality, clearly validate both constructs in these data.

Table 2. Fit of Cason and Cason' Model to 1983, 1985, and 1986 AERA Division I Program Review Data

	Year		
	1983	1985	1986
Multiple R[a]	.759	.786	.776
Components of Variance			
Reviewers	.117	.189[c]	.144[c]
Proposal Quality[b]	.459	.393	.415
Error ($1-R^2$)	.424	.418	.441
Number of Reviewers	87	86	82
Number of Proposals	120	100	115
Number of Observations	480	394	453

[a] All Rs are significant at $p<.00001$.
[b] Same as r_{ic} or single rater r_{xx}.
[c] Significant rater effect; $p < .025$.

Table 2 also shows the relative contribution of reviewer standards (stringency), proposal quality, and random error in the reviewers' acceptability ratings in each of the 3 years. Components of variance in Table 2 were estimated as a sum of the products of the respective standardized weights (Beta$_i$) and correlations (r_{iy}) between predictor variables and the criterion in the regression analysis:

(Equation 1)

$$\text{Proportion of Variance} = (\text{Beta}_i * r_{iy})$$

where i= 1 to n proposals; or, 1 to k reviewers.

The summation of products is across the set of either reviewer or paper proposals (Hays, 1963). The proportion of variance contributed by variation in reviewer standards/stringency ranged from 12 to 20%. This is an important, although modest, amount of total variance to be removed from the error term where it would be placed in an analysis making no provision for variance in rater standards as an explicit measurement design variable.

The relatively modest amount of total variance attributable to rater stringency is very misleading with respect to magnitude of the impact of an individual rater's standards upon the rating given to individual proposals of different levels of quality. Table 3 illustrates this point using the 1986 data. The table contains the expected rating for combinations of high, average, and low quality proposals (high

and low being defined as ± 1 S.D. from the mean SAP) and high, average and low stringency raters (where high and low is defined as ± 1 S.D. from the mean RRP). As the distribution of SAPs and RRPs in these data are approximately normal, these stipulative definitions of high and low avoid potentially misleading extreme outlyer cases. For example, it can be seen from the values in Table 3 that a paper proposal of mean intrinsic quality could have received a rating either near outright rejection or acceptance depending on whether a rater of high or low stringency had reviewed it.

Table 3. Rating (in %) Expected from Raters with High (+1SD), Mean, and Low (-1SD) Stringencies

	Stringency		
	High (+1SD) RRP=593	Mean RRP=496	Low (-1SD) RRP=399
Proposal Quality			
High (+1SD) SAP=606	56	86	98
Mean SAP=487	15	46	81
Low (-1SD) SAP=368	1	10	38

According to Hays (1963, p. 424), the intra-class correlation (r_{ic}) is a function of the variance attributable to an effect (σ_a) as a proportion of total variance.

(Equation 2)

$$r_{ic} = \sigma_a / (\sigma_a + \sigma_e)$$

The proportion of variance attributable to proposal quality reported in Table 2 can thus be interpreted as the intra-class correlation of reviewers with respect to their observed acceptability ratings of the proposals. As Hays points out, this is equivalent to the reliability of a single reviewer's observed acceptability rating. Alternatively, this value may be interpreted as the expected correlation between the ratings given by randomly chosen pairs of reviewers. The reliability of a mean of several reviewers' ratings, as is available in these data (where number of reviewers = k), is given by the Spearman-Brown expansion formula:

(Equation 3)

$$r_k = (k * r)/(1 + ((k - 1) * r))$$

where r = the reliability of a unit length measure, in this case a single reviewer; and,
k = number of reviewers.

Table 4 shows the impact of adjusting acceptability ratings on the reliability of both single reviewer and aggregate ratings obtained from 4 reviewers. The values for the single reviewer adjusted ratings were obtained by including only the sum of the error and proposal variances in the denominator of Equation 2. The unadjusted (observed) acceptability ratings must include the variance associated with reviewers in addition to that associated with proposals and error (Ebel, 1951). Thus, so long

as variance attributable to reviewers is greater than zero (regardless of the presence of a significant reviewer effect), adjusted acceptability ratings must have higher reliabilities than unadjusted ones. Therefore, as can be seen in Table 4 the reliability of adjusted ratings for each Division I year analyzed is greater than the reliability computed upon the observed ratings. For purposes of comparison, the reliability for observed and adjusted ratings of a single and an aggregate of four raters in Marsh and Ball's study are given. The value obtained by Marsh and Ball in each of these cases is systematically lower than the lowest comparable value obtained by our analysis of Division I review data.

Table 4. Reliability of Ratings
Intra-class Correlations

	Single Rater k=1		Aggregate of Raters k=4	
	Observed	Adjusted	Observed	Adjusted
Marsh and Ball Division I	.340	.350	.670	.683
1983	.459	.520	.768	.813
1985	.393	.485	.722	.790
1986	.415	.485	.739	.790

While the fit of the model to the data reported in Table 2 and the presence of significant rater effects in the 1985 and 1986 data support the validity of the model, its constructs, and its appropriateness to the kind of rating data under consideration here, further, stronger support for the model is available in the results reported in Table 5. Table 5 contains the correlations between reviewer stringencies (RRPs) estimated on reviewers who participated in program review in more than one year. For those who reviewed in both 1983 and 1985 and those who reviewed in both 1985 and 1986 there was a low but statistically significant correlation in their RRPs. The 1983-1986 correlation failed to reach statistical significance. These data clearly show that stringencies reflect some substantive characteristic of the reviewer which persists over a period of up to two years. The significant correlation between 83-85 reviewer stringencies emphasizes that while a significant rater "main" effect was absent, true differences in reviewer standards were measured. As there were real differences in reviewer standards in each year, adjustments for variation in rater standards produced real (i.e., statistically significant) improvements in reliability.

Table 5. Stability of Reviewer Standards over Time

	1985	1986
1983		
r	.33	.14
n	41	32
p	.02	.23
1985		
r		.27
n		40
p		.05

The results in Table 5 showing the persistence of consistent reviewer standards over time (as reflected in RRPs) are stronger than and therefore provide greater support for the theoretical model underlying these analyses than the only previously published comparable results (Cason & Cason, 1984, Table 3 p. 240). Although consistency among raters represented by an intra-class correlation (Ebel, 1951; Stanley, 1961) is frequently interpreted as a measure of reliability, it may also be interpreted as a measure of validity. Stanley (1961) observed that each rater may be considered a different method of measuring a given construct (e.g., paper proposal quality). The appropriateness of this interpretation is supported in the present context by its equivalent use in the analysis of reviews of manuscripts submitted to the Journal of Educational Psychology (Marsh & Ball, 1981). This interpretation seems particularly appropriate with respect to a global measure of proposal quality (i.e., acceptabilty rating) in light of the report by Littlefield and Troendle (1986). Therefore, the single rater reliabilities (intra-class correlations) reported in Table 4 may be equally well interpreted as both single rater construct reliability coefficients and single rater validity coefficients. However, reliability and validity do not expand in the same manner with increased numbers of independent observations. The increase in reliability is directly proportional to the number of observations; the increase in validity is approximately proportional to the square root of the number of observations as shown in Equation 4 (Gulliksen, 1950).

(Equation 4)

$$r_{xy,k} = r_{xy}(k^{1/2})/((1+(k-1)r_{xx})^{1/2})$$

where $r_{xy,k}$ is the validity based on k independent raters;
r_{xy} is the validity of a single rater;
r_{xx} is the reliabilty of a single rater; and,
k is the number of independent raters/ratings.

Table 6 reports the validity of a single rater and the aggregate of four raters as measures of global acceptability. As discussed above, the single rater observed and adjusted validities are equal to the corresponding single rater observed and adjusted reliabilities reported in Table 4. As with reliability, a non-trivial improvement in convergent construct validity was obtained by adjusted ratings when contrasted with observed ratings. By the same logic as was applied to reliabilities, the improvements in validity are real; that is, statistically significant.

Table 6. Validity of Ratings

	Single Rater k=1		Aggregate of Raters k=4	
	Observed	Adjusted	Observed	Adjusted
1983	.459	.520	.595	.650
1985	.393	.485	.532	.619
1986	.415	.485	.554	.619

The origin (i.e., the zero point) on the ability and stringency scale is arbitrary. In the actual estimation of RRPs one rater's RRP is chosen to anchor the scale and arbitrarily set equal to 500. This process is carried out independently on each set of data. In the present case, separate analyses were completed on each year's program proposals.

We made the plausible assumption that mean rater stringency remained constant for those raters who participated in reviews for both 1983 and 1985. There were 41 raters in common between 1983 and 1985. The mean RRP of these 41 raters on the original uncalibrated scales were 515.54 and 557.30 for 1983 and 1985 respectively. The two scales were calibrated by adjusting all the RRPs such that the 41 common raters had a mean RRP of 500. The original 1983 RRPs were adjusted by the additive constant -15.54. The original 1985 RRPs were adjusted by the additive constant -57.30.

The 40 raters in common between 1985 and 1986 had mean RRPs on the original scales of 568.48 and 514.76, respectively. When the original values of these 40 raters' RRPs on the 1985 scale were adjusted by -57.30 to fall on the calibrated scale for 1983-85, their resulting mean on the calibrated scale was 511.78. To obtain the same mean for these 40 raters' RRPs on the 1986 data required an adjustment of -3.58 for values on the original, uncalibrated 1986 scale. Within a given year, SAPs and RRPs are determined on the same scale; therefore, adjustments to achieve calibration were the same within each year for both RRPs and SAPs. This process is analogous to the calibration of exam scores when latent-trait models are used and calibration is achieved through equating item difficulties for linking items, i.e., sub-sets of exam items in common between exams. However, because item difficulties have much larger standard errors than do RRPs and SAPs, far less data is required in the rating case. All further information on and discussion of SAPs and RRPs is in terms of values on the calibrated scale defined above.

Table 7 provides summary information on reviewer standards (RRPs) and proposal quality (SAPs) in calibrated scale values for all three programs. In each year the mean stringency of program committee members was slightly greater than non-committee member reviewers (although as indicated by the standard errors, not significantly so). There was a slight increase in committee member stringency between 1983 and 1985; followed in 1986 by a decline to approximately the 1983 level. These changes were also not statistically significant. Over the three years, non-committee members' average stringency fluctuated even less than did that of committee members. The differences between mean committee members' and mean non-committee members' stringencies within and across years were also not statistically significant. The absence of statistically different mean stringencies indicates that the observed differences could be attributed to chance fluctuations in rater stringencies arising from random sampling of reviewers from the same hypothetical pool of potential reviewers. Nevertheless, any difference in rater standards has the potential of making a practical difference with regard to the evaluation of an individual paper proposal.

Proposal quality as measured by mean SAPs of proposals accepted fluctuated significanctly between 1983 and 1985 and between 1985 and 1986; first declining then rising above the 1983 value. In each year the mean quality of the rejected proposals was significantly below that of the accepted proposals. Across the three years, the quality of rejected proposals declined substantially from 1983 to 1985 then returned in 1986 to near the 1983 value. Under the assumption that scale calibration across years was successful, the lower proposal quality of 1985 cannot be attributed to the concurrent, slightly higher stringencies of reviewers in that year.

Table 7. Reviewer Standards and Proposal Quality

	Year		
	1983	1985	1986
Reviewer Stringency (RRP)			
Committee Members			
N	6	8	8
Mean	499.5	505.7	499.4
Se	22.6	11.1	19.8
Non-Committee Member			
N	81	76	74
Mean	497.5	496.7	492.4
Se	7.9	13.1	11.6
Proposal Quality (SAP)			
Accepted			
N	42	33	35
Mean	571.5	528.7	587.4
Se	13.8	12.7	14.4
Rejected			
N	78	67	80
Mean	443.5	377.7	437.4
Se	9.8	12.5	11.4

Table 8 reports the correlations between disposition of proposals (i.e., acceptance = 1, or, rejection = 0 for the program) with the mean observed acceptability across 4 reviewers (including one program committee member), the adjusted acceptability rating of each proposal, and the acceptability rating given by the Program Committee member. The moderate values of the correlations between the mean observed acceptability ratings and disposition of proposals reflects, in part, the less than perfect reliability of this measure which was available at the time that the disposition decision was made. According to the informal account of Program Committee members, other factors contributing to the moderate correlation between mean observed acceptability rating and disposition included: Division I policy to encourage participation from professions previously under-represented in the program by defacto application of less stringent standards, committee members giving differential credibility to selected reviewers, accepting only a single paper from a given author that proposed two or more closely related papers each of which received high acceptability ratings, rejection of papers the same or highly similar to ones presented by the author elsewhere in spite of high acceptability ratings.

Table 8. Correlations: Disposition with Acceptability Ratings

	Year		
	1983 N=120	1985 N=100	1986 N=115
Disposition with			
Mean Observed Rating	.64	.53	.61
Adjusted Rating	.60	.47	.61
Program Committee Member's Rating	.56	.70	.49
Upper Limit of r	.84	.85	.87

In 1983 and 1986 the acceptability rating given by the Program Committee member was correlated only slightly less strongly with disposition than the mean acceptability rating across all reviewers. This suggests that in these cases the actual disposition was influenced by the reviews of non-committee members. The reversal of this pattern in 1985 may be an artifact of the process used to make the disposition decision and to record the Program Committee member's acceptability rating. According to one Program Committee member, the committee reached consensus on the disposition decision and the record of an individual Program Committee member's rating of a particular paper was changed to conform with this consensus. Presumably the reason for the correlation between disposition and Program Committee members acceptability rating being only .70 results from a failure to alter the recorded individual Program Committee member's acceptability rating to conform with the committee's decision.

The lower correlation between the mean observed acceptability rating with disposition and the adjusted with the disposition found in 1985 concurrent with the much higher correlation between the Program Committee member's rating and disposition is consistent with selective attention on the part of committee members to other reviewers' acceptability ratings that were more "credible". However, these apparently anomalous results are only suggestive of that hypothesis and are open to other interpretations.

Under Cason and Cason's model the best available (i.e., most reliable, valid) single measure of proposal acceptability is the adjusted aggregate-multirater acceptability rating. This measure was not available to any of the Program Committees at the time that disposition decisions were made. A measure of how well the committee managed to extract the best information from the observed ratings available to them is the correlation between disposition and adjusted ratings. By this interpretation the Program Committees in 1983 and 1986 did the best and about equally well. The lower correlation between disposition and adjusted ratings in 1985 suggests that this committee would be less likely to endorse adjusted ratings as "best" measures in spite of the fact that the Cason and Cason model achieved its best fit with the 1985 data; there was a significant rater stringency effect and the adjusted ratings were more construct valid (i.e., .53 for observed vs .62 for adjusted).

The correlations between the mean observed acceptability ratings and the adjusted acceptability ratings for 1983, 1985, and 1986 were respectively .92, .86, and .93. The lower correlation found in 1985 data reflects the lower proportion of variance attributable to proposal quality and higher proportion of variance attributable to reviewer standards in 1985 than in either 1983 or 1986. Similarly 1985 had the lowest reliability associated with observed acceptability ratings. Each of these findings represent related but different quantifications of the fact that in 1985 the Program Committee's task of extracting useful information from the observed ratings was more difficult than in the other two years. Furthermore, the correlation between the mean observed acceptability rating with disposition reported in Table 8 tends to exaggerate the departure between what the committee actually chose to accept and acceptance based on a simple selection of the top N papers each year as determined strictly on mean observed ratings (where N = number of papers accepted within a given year). The maximum value of this correlation is a function of the proportion of proposals to be accepted within any given year and in none of these years would it have been equal to 1.00 (McNemar, 1969). For example, in 1986 when 35 of 115 paper proposals could be accepted, the maximum correlation obtainable between mean observed ratings and disposition is .87.

In Table 9, the 1986 data are used to provide a contrast between the results of the Program Committees' actual selection policy and what the results would have looked like had they (a) chosen the top 35 proposals based on mean observed acceptability ratings or (b) chosen the top 35 proposals based on adjusted acceptability ratings. For reasons discussed above, the range of acceptability ratings for accepted and rejected proposals resulting from the Program Committee's actual disposition decisions overlaps substantially. At least one paper with a high rating (83%) was rejected while at least one with a moderately low rating (34%) was accepted. The other two decision rules prohibit overlaps of this kind between accepted and rejected proposals. Table 9 shows the second two decision rules result in higher mean ratings of accepted and lower mean ratings for rejected proposals with the rule based on adjusted scores giving the greatest differentiation between the means of the two groups. The correlation between the disposition of the proposals and the mean observed acceptability ratings for the actual program was .61. For disposition and strict ranking based on mean observed acceptability the correlation was .76. Correlation between the adjusted score and selection based strictly on ranking of the adjusted score was .77.

Table 9. Contrast Between Alternate Selection Policies

	Committee Selection	Outcome Based on Ranking of Mean Observed Scores	Ranking of Adjusted Scores
Accept (N=35)			
Mean	69.9	75.4	79.2
Se	2.6	1.6	1.7
Range	33.5-94.2	62.2-94.2	65.6-99.0
Reject (N=80)			
Mean	33.6	37.2	35.1
Se	2.2	1.9	2.1
Range	6.3-83.3	6.3-61.2	1.6-65.4

The correlation between the mean observed and the adjusted acceptability ratings of the 1986 proposals was .93. Given this and the relatively modest difference in the reliability of these two measures (.74 vs. .79) it might appear that little difference would arise from chosing to use one or the other. This turns out not to be the case.

Table 10 shows the impact of using either the observed mean or the adjusted rating for selecting proposals for inclusion in the program given the simplified decision rule of selecting the top 35 proposals. While this table does not directly show exactly what the choice between these two measures of acceptability would have produced in the context of the Program Committee's more complicated decision rule, it is probably highly suggestive and close to what would have happened. Of the 35 included in the program by the simplified decision rule based on either of these two measures, 6 (17%) included under one measure would be excluded under the other and visa versa. A 17% change in the specific proposals included in the program is a practically important difference. Even if the difference were only a single proposal, decisions based on the adjusted ratings would be superior because they are based on both a more reliable and a more construct valid measure (Stanley, 1961, Marsh & Ball, 1981; Cason & Cason, 1984).

Table 10. Transitions in Accept/Reject Outcome Resulting from Using Adjusted or Observed Ratings

	Outcome Based on Adjusted Ratings		
	Accept	Reject	Total
Outcome Based on Observed Ratings			
Accept	29	6	35
Reject	6	74	80
Total	35	80	115

All of the analyses reported to this point have been based on the acceptability rating located at the top of the Division reviewer inventory (see Figure 4). That inventory also requests that the reviewer rate the proposal on seven quality criteria. Currently the information on the multiple forms completed on an individual proposal is not systematically, formally integrated into a composite report for use by the Program Committee nor feedback to the proposal authors or proposal reviewers. As the information contained in this section of the reviewer's inventory is potentially as useful as the global acceptability recommendations, cursory analyses of the proposal quality data were completed on the 1986 data (this is the only year for which these data were made available to the researchers).

The first step in the analysis of these data was the transfer of quality ratings from the reviewer inventory to OTS-PR machine scannable rating sheets (See Figure 5). These rating sheets were then scanned and processed by the UAMS OTS-PR system which generated summary reports including inventory analyses and proposal quality summaries. Figure 6 which is photo-reduced output from the UAMS OTS-PR system provides the single rater and k rater mean reliabilities (where k is the geometric mean number of raters per paper; Ebel, 1951) for each of the seven quality criteria, the mean observed rating, the standard deviation and standard error of measurement as well as the same statistics on the average across criteria (i.e., the average or overall proposal quality rating). The moderately low reliabilities for the average across multiple raters for each of the criteria and the average or overall quality rating leaves substantial room for improvement in the scale itself and the way in which it is used by reviewers. The single rater reliabilities reported in Figure 6 are the convergent validity coefficients for each of the quality criteria and the summative total across these criteria. These values were computed in the same manner (i.e., as single rater reliabilities/intra-class correlations) as that used by Marsh and Ball to compute the diagonal element (convergent validities) in their multi-method (rater 1, rater 2) multi-trait (manuscript review subscales) analysis in accord with Campbell and Fiske (1959). The note to Table 2 by Marsh and Ball explicitly states this equivalence between single rater reliabilities and convergent validity coefficients. The validities for Marsh and Ball's subscales ranged from .20 to .27. As can be seen from Figure 6, the analogous validities for the 1986 quality criteria (subscales) ranged from .17 to .24.

The validity found by Marsh and Ball for the overall recommendation for acceptance of a manuscript was .34. For the 1986 Division I data the comparable value was .415 (calculated as a single rater reliability because the single rater reliability is equivalent to the two rater intercorrelation given by the intra-class correlation). In actual fact both of these numbers represent underestimates of the true validity of the data. In Marsh and Ball's analysis two raters' data on each manuscript were available. In the 1986 Division I data four independent ratings were available. The single rater validities given above must be expanded using Equation 4

to determine the validity of the aggregate of multiple independent ratings. With respect to overall recommendation or acceptability rating, the Division I review process is more valid than that found in Marsh and Ball's study as a result of both (a) higher validity at the single rater level; and (b) more independent ratings per manuscript/paper proposal. In fact, in each of the three years analyzed the Division I review process had greater single rater validity and a greater number of independent ratings per manuscript/paper proposal than in the Marsh and Ball study.

The correlation between the average quality and the mean observed acceptability rating was .83. The moderately high correlation between the mean quality rating and the mean observed acceptability rating indicates that these two measures reflect substantially but not exactly the same thing (they share 60+% of the variance). This leaves unresolved the question of whether one or the other or some explicit combination of the two measures would best serve the Program Committee as a summative integration of the information provided by the reviewers. Littlefield and Troendle's (1986) results suggest including acceptability as a subscale within the list of quality criteria, preferably at the top of the list.

Figure 7 parts A, B, and C provide photo-reduced facsimiles of the individual performance reports (IPRs) generated by UAMS OTS-PR on the best, a near average, and the weakest proposals, as measured by mean quality rating, submitted to the 1986 Program Committee.

The principle use of the OTS-PR system with respect to rating data is processing ratings of students' performance in clinical settings as part of degree/credit granting courses and clerkships and formal training programs such as residency programs. For this reason, labeling of some aspects of the report is at variance with the current application: "students" are the subjects of the evaluation, "class" is the collective group of subjects upon whom evaluations were conducted (in this case all 1986 proposals). The IPR provides, in both graphic and tabular form, information on an individual's performance on each item and the performance of an average member of his comparison group. The standard error of measurement is provided on each item as well as on the subject's total score. The standard deviation of scores in the class is provided on each item and the total. The number of raters upon whom a given subject's average on an item was computed is provided in the right most column of the report. Note that the number of raters varies from report to report and item to item in these examples. The number given is the number of ratings in the valid range, i.e., 1 to 5, and excludes omissions, NAs, etc.

A glance at the graphic portions of these three reports quickly conveys the relative strengths and weaknesses by quality criteria of average proposals submitted in 1986: lower case "c" profiles in each graph. These reports also rapidly communicate the range of quality on each criterion from the best proposal to the weakest proposal: lower case "x" profiles.

In passing, it is worth noting that the greatest weakness on average in the 1986 proposals is the credibility of findings and conclusions. On average, the greatest strength of the 1986 proposals was appropriateness to Division I. Do these results imply that those proposals highly appropriate to Division I lack credibility in their findings and conclusions? This question emphasizes some of the ambiguities and uncertainties in the intended meaning of the quality criteria.

Conclusions

The results relating disposition of proposals to reviewer rated acceptability when combined with the obtained reliabilities for observed mean acceptability ratings clearly indicate that in the three years studied, the Program Committee chair and members, reviewers, and the general Division I review process did an excellent job in selecting high quality programs for the Division. There is a clear distinction in the quality of papers, on average, accepted and rejected and reasonable policy explanations for why some relatively high rated proposals could be rejected and/or moderately low rated proposals might be accepted. The Division I review process was shown to be both more valid and reliable than that reported in an analysis of manuscripts submitted to a high quality peer review journal concerned with a domain of research problems in many ways similar to that of interest to Division I.

Cason and Cason's simplified model of performance rating fit each set of review data. Empirical support was found in all years for both major model constructs: stringency (rater standards) and ability (proposal quality). Even in that year (1983) where no significant stringency effect was directly observed, the assignment of part of the variance to stringency could not be discounted as capitalizing on chance. This follows from the significant correlation of stringencies estimated for raters participating in the review process in both 1983 and 1985. On average, the stringency of both committee and non-committee reviewers may be interpreted as drawn from the same hypothetical population of potential reviewers. It is reasonable to expect that Cason and Cason's model would fit future program review data unless the rater pool changed in some substantial way.

Application of the model permitted partitioning of the variance so that a more valid and reliable measure of proposal acceptability than represented by the mean observed acceptability rating was extracted from the data. Even though the increase in validity and reliability was modest, the adjusted ratings were nevertheless more valid and reliable.

It might appear easy to dismiss these results as trivial even though statistically significant because reliability was improved in 1986, for example, only from .74 to .79 and validity improved from .55 to .62. However, this improvement could result in up to a 17% change in the composition of the program. Even if the acceptance or rejection of only a single proposal were affected, the adjusted ratings provide the preferred criteria.

It might also be tempting to dismiss or undervalue the results because of the presumptively multi-dimensional nature of both the measurement of the quality of the proposals and the decision to include or exclude papers proposed for the program. Supporters of this view would likely take encouragement from the relatively large unexplained variance in each year's data; the Cason and Cason model accounted for only 56% of the variance in the 1986 data, leaving 44% as unexplained error. However, the Cason and Cason model can be applied in a multi-dimensional manner. If independent (i.e., orthogonal) evaluative criteria can be identified, Cason and Cason's model can be applied separately to each factor and achieve a reduction in the error term for variation in rater standards within each factor. Thus, while a multi-dimensional analysis might account for a greater proportion of variance than did the uni-dimensional analysis reported here, for each factor in that analysis it is likely that partialling out rater standards in each factor would produce a further incremental reduction in error variance for each factor. This effect has already been demonstrated in ratings of medical students' performance in a clinical practice setting (C. Cason, G. Cason, & Littlefield, 1983).

These results suggest that there are several areas in which changes might provide improvements in the validity and reliability of the review process. The results clearly show that there is significant variation in rater standards which affects the validity and reliability of the review process. The task of the Program Committee would be made less difficult were they provided adjusted acceptability ratings at the time the decision to include or exclude paper proposals is made. The Program Committee would also likely find it useful were they provided a summary quantitative report integrating the ratings provided by each reviewer of each separate proposal (e.g., similar to those illustrated in Figure 7). In addition there may be improvements that are possible with respect to the separate quality items or their definitions to be included in the proposal review inventory. According to Marsh and Ball improvements in review inventory content have indeed been accomplished by Gottfredson (1978). Furthermore, an analysis of the validity and reliability of the review and acceptance process should be a routine part of that process. This would provide the Program Committee a means for both monitoring the current quality of the process and movement toward the goal of improved reliability and validity of the process. Implementation of these suggested changes requires the availability and application of machine and/or computer based automation technologies for the collection, analysis, and reporting of rating data.

Importance

Progress in professions education research depends, in part, upon an efficient, effective, and believable process for selecting the best papers and articles for inclusion in professional meetings and journals. The approach presented here permits assessment of the current state and progress toward improving the peer review and program committee processes in Division I and has potential for use in other similar settings. Certainly, the results support a greater level of confidence in the selection process than some might have otherwise believed. Yet, there is clearly room and need for further improvement. The methods presented and suggestions made can provide part of the basis for making such improvements. The results reported above emphasize our need to take heed of Marsh and Ball's wry obsevation "It seems ironic that scientific method has scarecely been used to determine how best to evaluate the products of scientific research" (p. 880).

References

Campbell, D.T. & Fiske, D.C. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cason, C.L., Cason, G.J., & Littlefield, J.H. (1983). Variation of intra-rater stringency in cognitive-technical and affective-interpersonal clinical performance domains. Presented at AERA, Montreal.

Cason, G.J., & Cason, C.L. (1985). A regression solution to Cason and Cason's model of clinical performance rating: Easier, cheaper, faster. Presented at AERA, Chicago.

Cason, G.J., & Cason, C.L. (1984). A deterministic theory of clinical performance rating: Promising early results. Evaluation & the Health Professions, 7(2), 221-247.

Cason, G.J., Cason, C.L., & Littlefield, J.H. (1983). Controlling rater stringency error in clinical performance rating: Further validation of a performance rating theory. Presented at AERA, Montreal. Resources in Education, 18(8), 176. (ERIC

ED-228-314)

Ebel, R.E. (1951). Estimation of reliability of ratings. Psychometrika, 16: 407-424.

Gottfredson, S.D. (1978). Evaluating psychological research reports: Dimensions, reliability and correlates of quality judgements. American Psychologist, 33, 915-929.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Hays, W.L. (1963). Statistics. New York: Holt, Rinehardt & Winston.

Littlefield, J.H., Ellis, R., Cohen, P., & Herbert, R. (1984). Leniency and score distribution differences among clinical raters. In Research in Medical Education 1984: Proceedings of the 23rd Annual Conference. Washington, D.C.: Association of American Medical Colleges, 199-204.

Littlefield, J.H., & Troendle, G.R. (1986). Rating format effects on rater agreement and reliability. Presented at AERA: San Francisco.

Marsh , W.H., & Ball, S. (1981). Interjudgemental reliability of reviewers for the Journal of Educational Psychology. Journal of Educational Psychology, 73(6), 872-880.

McNemar, Q. (1969). Psychological Statistics (4th ed.). New York: Wiley.

Stanley, J.C. (1961). Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. Psychometrika, 26(2), 203-219.

Ward, J., & Jennings, E. (1973). Introduction to linear models. Englewood Cliffs, NJ: Prentice-Hall.

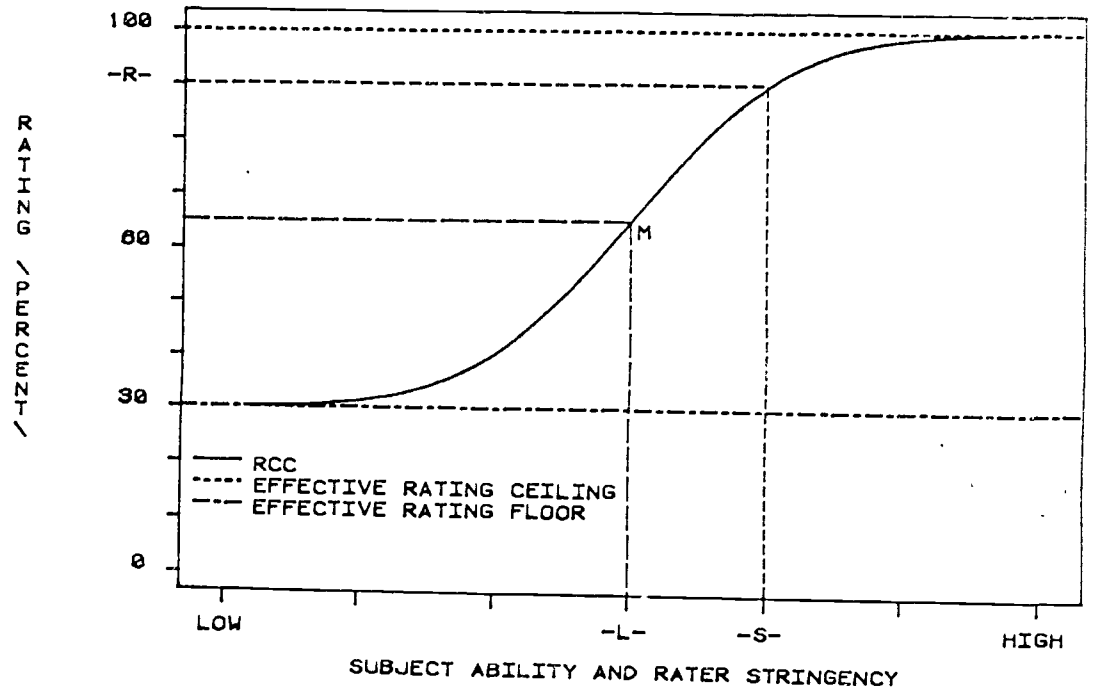

Figure 1 Example rater characteristic curve (RCC) Stringency L gives rating R to ability S

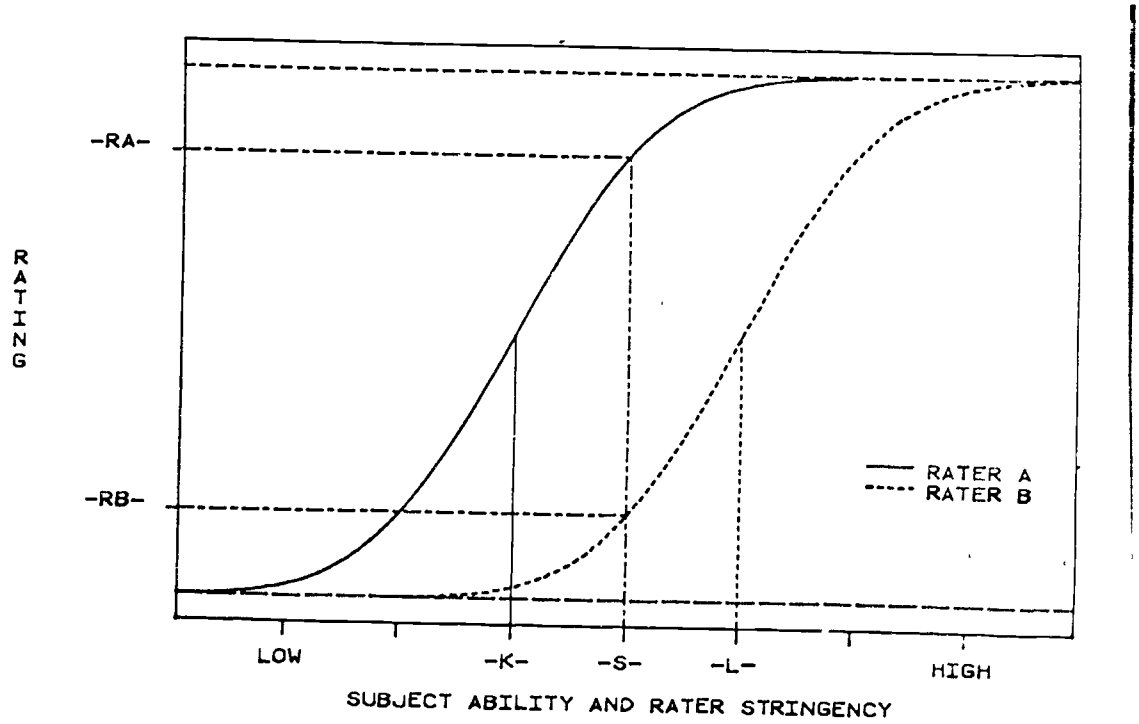

Figure 2. Raters A and B of stringencies K and L give subject of ability S ratings RA and RB respectively

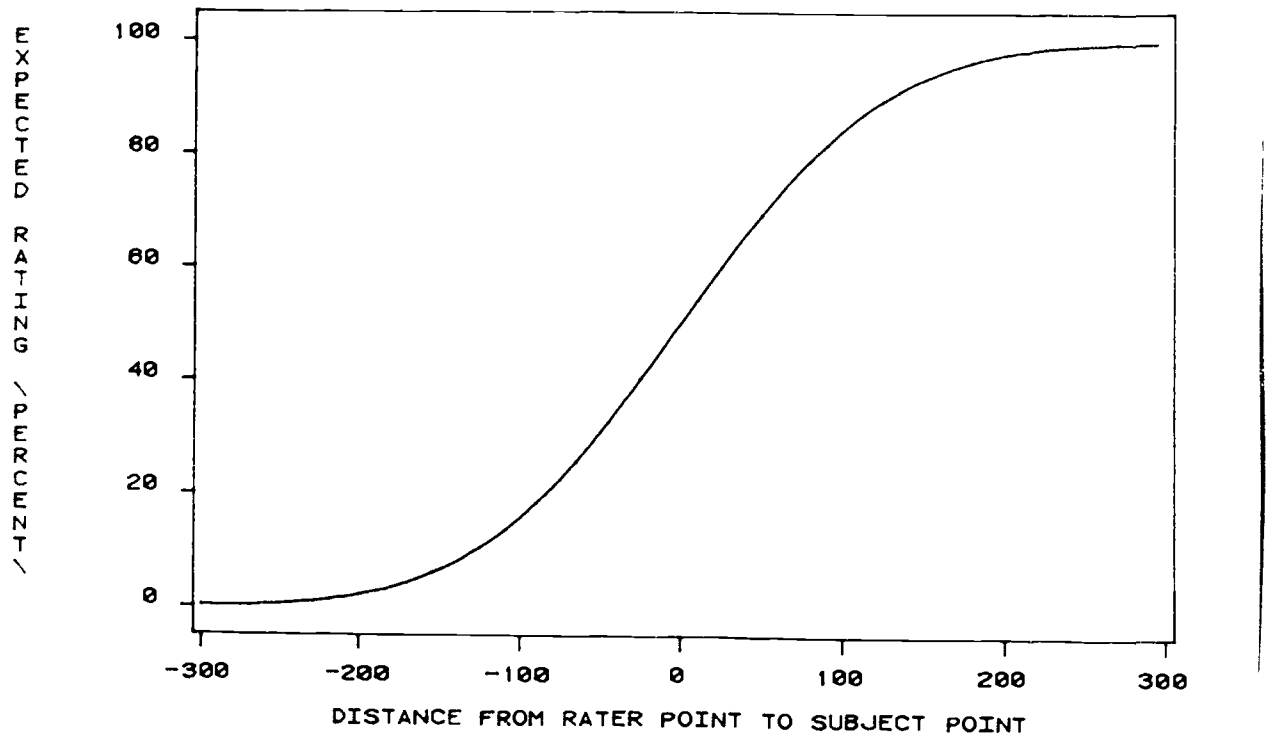

Figure 3 Expected Rating as a Function of Distance Between Rater Reference Point and Subject Ability Point

FIGURE 4. Reviewer Inventory

DIVISION I -- AERA
1985 PAPER PROPOSAL EVALUATION FORM

21

Paper ID#: P-____ Reviewer: _____

Paper title: _____

Recommendation:

1. Definitely accept.
2. Acceptable; suggest minor changes.
3. Accept only if space permits. Weaknesses noted in comments.
4. Reject.

If rated 1 or 2, list suggestions for discussants below.

Name: _____ Phone: ___ _____

Affiliation: _____

Name: _____ Phone: ___ _____

Affiliation: _____

Below this line will be sent to authors.

ID#: P-_____ Paper title: _____

	Not Applicable		
Clarity of Summary	☐	+ + + + +	Obscure (incomplete) ... Clear (all elements treated)
Relevance of Problem (to education, to society)	☐	+ + + + +	Insignificant ... Important to field
Theoretical Framework	☐	+ + + + +	Non-existant ... Well grounded
Methodology or Mode of Inquiry	☐	+ + + + +	Insuffiently developed ... Highly Appropriate
Execution of Study (coherence, clarity)	☐	+ + + + +	Unsystematic ... Carefully done
Findings & Conclusions	☐	+ + + + +	Lacks credibility or overstated ... Well founded
Appropriateness to Division I	☐	+ + + + +	Inappropriate (Where should it be _____) ... Highly Appropriate

22

EXPLAIN REASON FOR RANKS (use reverse side of paper ~~if necessary~~)

A2303

NCS Trans-Optic® B03-9542-65

SUBJECT NAME (PERSON RATED): P ABE, Z.

SUBJECT TYPE (OR RANK):

ACTIVITY OR TOPIC RATED: 1986 PROPOSALS

SUBJECT IDENTIFICATION

INITIALS (MARK EMPTY CIRCLE IF NO INITIAL)

FIRST Z
MIDDLE
LAST P

SUBJECT I.D. NUMBER: 0 0 8 8 0 6 8 2 9

PLEASE:

SEE INSTRUCTIONS ON BACK BEFORE STARTING. UNLESS OTHERWISE DIRECTED, ALL I.D. GRIDS TO RIGHT MUST BE COMPLETE BEFORE SUBMITTING FORM FOR MACHINE READING & SCORING.

USE SOFT, BLACK PENCIL. ONE MARK PER ITEM. USE POINT SCALE BELOW.

COURSE CLERKSHIP OR UNIT

NAME: AERA DIVISION I PROGRAM: 1986

CHIEF INSTRUCTOR, COORDINATOR, OR DIRECTOR: CAROLE J. BLAND, PHD

NUMBER: 5 0 2 1 1

RATER NUMBER

NAME:

TYPE OR RANK:

THIS IS SHEET ___ OF ___

LINE PRINTER ALIGNMENT CHECK

MARKING EXAMPLE

OK
WRONG
WRONG
WRONG
WRONG
WRONG
OK

ALIGNMENT CHECK

5 = OUTSTANDING
4 = VERY GOOD
3 = GOOD
2 = POOR
1 = VERY POOR

USE TO RATE PAPER PROPOSAL (SUMMARY).

UNIVERSITY OF ARKANSAS MEDICAL SCIENCES CAMPUS

GENERAL PURPOSE RATING FORM GPR-1(a)

FIGURE 5

Rating	NOT APPLICABLE	NO OPINION	Item		Rating	NOT APPLICABLE	NO OPINION
1 2 3 4 5			CLARITY		1 2 3 4 5		
1 2 3 4 5			PROBLEM RELEVANCE/IMPORTANCE		1 2 3 4 5		
1 2 3 4 5			THEORETICAL FRAMEWORK		1 2 3 4 5		
1 2 3 4 5			METHODOLOGY/MODE OF INQUIRY		1 2 3 4 5		
1 2 3 4 5			EXECUTION OF STUDY		1 2 3 4 5		
1 2 3 4 5			CREDIBILITY OF FINDINGS/CONCLUSIONS		1 2 3 4 5		
1 2 3 4 5			APPROPRIATENESS TO DIVISION I		1 2 3 4 5		

MAKE NO FOLDS WHICH CROSS EDGE BELOW.

TOTALS (LINE PRINTER ENTRY ONLY) P.L.G. RAW= %= RANK= Z=

BEST COPY AVAILABLE

RATING ANALYSIS SUMMARY (Current Rating)

Prepared 10-Jun-86 00:06 by the UAMS OIS/PR System (version A1) as implemented at UAMS

Test: 1 - 1986 PROPOSAL EVALUATIONS
Instructor: CAROLE J. BLAND, PHD
Course: AERA DIVISION I PROGRAM: 1986

Dept: Educational Development
Slot: 595 Phone:661-5720
Subjects rated:115 Absent: 2 Withdrawn: 1

Category	# Rated Items & Total Points (1)		Average Score Raw	Average Score 5 Pt	1 Rater Reliab.	Mean # Raters & Reliability (2)		Standard Dev Raw	Standard Dev 5 Pt	Std Error of Measure Raw	Std Error of Measure 5 Pt	Measure Z (3)
1 CLARITY	1	5	3.5	3.52	0.24	3.8	0.55	0.8	.755	0.51	.51	67
2 PROBLEM RELEVANCE/IMPORTANCE	1	5	3.8	3.80	0.18	3.8	0.46	0.6	.645	0.47	.47	74
3 THEORETICAL FRAMEWORK	1	5	3.2	3.20	0.23	3.7	0.52	0.7	.743	0.51	.51	69
4 METHODOLOGY/MODE OF INQUIRY	1	5	3.2	3.15	0.24	3.8	0.55	0.8	.845	0.57	.57	67
5 EXECUTION OF STUDY	1	5	3.3	3.32	0.23	3.4	0.50	0.8	.822	0.58	.58	71
6 CREDIBILITY OF FINDINGS/CONCLUSIONS	1	5	3.0	3.01	0.19	3.3	0.43	0.9	.889	0.67	.67	75
7 APPROPRIATENESS TO DIVISION I	1	5	4.0	3.97	0.17	3.8	0.43	0.7	.681	0.51	.51	75
Overall	7	35	24.0	3.43	0.25	3.9	0.56	4.0	.575	2.65	.38	66

(1) The Number of Rated Items does not include category header items, i.e. unrated items which head categories with more than one item.

(2) The Reliability is reported for a single rater and for the average rating across the average (harmonic mean) number of raters that rated each subject. This statistic is calculated as recommended by Ebel (PSYCHOMETRICA, 16, 1951) as an intra-class inter-rater reliability for a single rater and is expanded by Spearman-Brown's formula to estimate the reliability of the average rating across multiple raters. Variation in rater standards (stringency/leniency) will reduce reliability. Increases in inter-rater agreement (single rater reliability) and the number of raters rating each subject increase reliability. When only one rater rates each subject reliability cannot be estimated. Reliabilities may be interpreted as:

Below 0.60	UNACCEPTABLE	as sole basis for evaluation
0.60 - 0.69	POOR	as sole basis for evaluation
0.70 - 0.79	FAIR	as sole basis for evaluation
0.80 - 0.90	GOOD	as sole basis for evaluation
0.90 - 1.00	EXCELLENT	as sole basis for evaluation

(3) The Standard Error of Measurement (SEM) allows one to estimate the probability that a subject's score on a subsequent rating will fall within a given range of his original rating score. This estimate assumes that the number of raters is the same for both ratings and that no real change has occured in the level of knowledge or skill of the subject. The following chart gives the probabilities that re-rating scores will fall within certain ranges of original scores:

Level of Confidence (Probability)	Predicted Score Range Original Score + and -
75%	1.15 times SEM
95%	1.96 times SEM
99%	2.58 times SEM

The value of the SEM decreases as the value of the Reliability increases. Thus, performance is more precisely measured for larger values of the Reliability.

Figure 6. Rating analysis summary

INDIVIDUAL PERFORMANCE REPORT (Current Rating)

Prepared 10-Jun-8f 00:10 by the UAMS OTS/PR System
(version A1) as implemented at UAMS

To Student: 585000441
From: CAROLE J. BLAND, PHD
Re: Rating/test 1-1986 PROPOSAL EVALUATIONS

Dept: Educational Development
Course: AERA DIVISION I PROGRAM: 1986

Class Overall Mean Rating = 3.43-->
Your Overall Mean Rating = 4.54-->

Item	5 Point Scale 1.....2.....3.....4.....5	5 Pt Score — Your Mean=x	SEM	Class Mean=c	StdDev	Raw Score — Perfect p	Yours xp/5	# of Raters
1 CLARITY	.c—x—	4.50	.50	3.52	.755	5	4.50	4
2 PROBLEM RELEVANCE/IMPORTANCE	.—c—.—x—	4.75	.47	3.80	.645	5	4.75	4
3 THEORETICAL FRAMEWORK	c—.—x—.—	4.25	.50	3.20	.743	5	4.25	4
4 METHODOLOGY/MODE OF INQUIRY	c—.—.—x	5.00	.56	3.15	.845	5	5.00	4
5 EXECUTION OF STUDY	c.—.—x—	4.50	.56	3.32	.822	5	4.50	4
6 CREDIBILITY OF FINDINGS/CONCLUSIONS	c—.—x—.—	4.00	.64	3.01	.889	5	4.00	4
7 APPROPRIATENESS TO DIVISION I	.—c—.—x—	4.75	.51	3.97	.681	5	4.75	4

Rating Scale

5 = OUTSTANDING
4 = VERY GOOD
3 = GOOD
2 = POOR
1 = VERY POOR

Definition of Symbols

C = class overall 5 pt score: 3.43 StdDev: .575
X = your overall 5 pt score: 4.54 SEM : .38
c = class mean 5 pt score on item (or category)
x = your mean 5 pt score on item (or category)
SEM= Standard Error of Measurement

Your overall raw score 31.75 (out of perfect 35) yields: 90.7% Z= 693 Rank= 1 (out of 115). Class ave raw score 23.99

Figure 7A. Individual performance report for best proposal

INDIVIDUAL PERFORMANCE REPORT (Current Rating)

Prepared 10-Jun-86 00:10 by the UAMS OIS/PR System (version A1) as implemented at UAMS

To Student: 585000300
From: CAROLE J. BLAND, PHD
Re: Rating/test 1-1986 PROPOSAL EVALUATIONS

Dept: Educational Development
Course: AERA DIVISION I PROGRAM: 1986

Class Overall Mean Rating = 3.43—>
Your Overall Mean Rating = 3.43—>

Item	5 Point Scale (1.....2.....3.....4.....5)	5 Pt Score — Your Mean=x	SEM	Class Mean=c	StdDev	Raw Score — Perfect p	Yours xp/5	# of Raters
1 CLARITY	.C—x—	3.75	.50	3.52	.755	5	3.75	4
2 PROBLEM RELEVANCE/IMPORTANCE	.—x—	3.75	.47	3.80	.645	5	3.75	4
3 THEORETICAL FRAMEWORK	x—C—.	2.75	.50	3.20	.743	5	2.75	4
4 METHODOLOGY/MODE OF INQUIRY	x—C—.	2.75	.56	3.15	.845	5	2.75	4
5 EXECUTION OF STUDY	C.x—	3.50	.56	3.32	.822	5	3.50	4
6 CREDIBILITY OF FINDINGS/CONCLUSIONS	C—.x—	3.50	.64	3.01	.889	5	3.50	4
7 APPROPRIATENESS TO DIVISION I	.—C—x—	4.67	.54	3.97	.681	5	4.67	3

Rating Scale

5 = OUTSTANDING
4 = VERY GOOD
3 = GOOD
2 = POOR
1 = VERY POOR

Definition of Symbols

C = class overall 5 pt score: 3.43 StdDev: .575
X = your overall 5 pt score: 3.43 SEM : .38
c = class mean 5 pt score on item (or category)
x = your mean 5 pt score on item (or category)
SEM= Standard Error of Measurement

Your overall raw score 24.04 (out of perfect 35) yields: 68.7% Z= 501 Rank= 59 (out of 115). Class ave raw score 23.99

Figure 7B. Individual performance report for proposal of average quality

INDIVIDUAL PERFORMANCE REPORT (Current Rating)

Prepared 10-Jun-86 00:10 by the UAMS OTS/PR System (version A1) as implemented at UAMS

To Student: 585001027
From: CAROLE J. BLAND, PHD
Re: Rating/test 1-1986 PROPOSAL EVALUATIONS

Dept: Educational Development
Course: AERA DIVISION I PROGRAM: 1986

Class Overall Mean Rating = 3.43—>
Your Overall Mean Rating = 1.87—>

Item	5 Point Scale (1.....2.....3.....4.....5; C = class mean, X = your mean)	— Your Mean=x	5 Pt Score — SEM	— Class Mean=c	— StdDev	Raw Score — Perfect p	Yours xp/5	# of Raters
1 CLARITY		2.00	.54	3.52	.755	5	2.00	3
2 PROBLEM RELEVANCE/IMPORTANCE		2.00	.47	3.80	.645	5	2.00	4
3 THEORETICAL FRAMEWORK		1.75	.50	3.20	.743	5	1.75	4
4 METHODOLOGY/MODE OF INQUIRY		1.75	.56	3.15	.845	5	1.75	4
5 EXECUTION OF STUDY		2.00	.65	3.32	.822	5	2.00	2
6 CREDIBILITY OF FINDINGS/CONCLUSIONS		1.00	.74	3.01	.889	5	1.00	2
7 APPROPRIATENESS TO DIVISION I		2.00	.51	3.97	.681	5	2.00	4

Rating Scale
5 = OUTSTANDING
4 = VERY GOOD
3 = GOOD
2 = POOR
1 = VERY POOR

Definition of Symbols
C = class overall 5 pt score: 3.43 StdDev: .575
X = your overall 5 pt score: 1.87 SEM : .38
c = class mean 5 pt score on item (or category)
x = your mean 5 pt score on item (or category)
SEM= Standard Error of Measurement

Your overall raw score 13.09 (out of perfect 35) yields: 37.4% Z= 229 Rank= 115 (out of 115). Class ave raw score 23.99

Figure 7C. Individual performance report for weakest proposal