THE YALE JOURNAL OF BIOLOGY AND MEDICINE 49, 373-383 (1976)

A Statistical Analysis of Reviewer Agreement and Bias in Evaluating Medical Abstracts[1]

DOMENIC V. CICCHETTI[2] AND HAROLD O. CONN[3]

Veterans Administration Hospital, West Haven, Connecticut 06516

Received January 5, 1976

Observer variability affects virtually all aspects of clinical medicine and investigation. One important aspect, not previously examined, is the selection of abstracts for presentation at national medical meetings. In the present study, 109 abstracts, submitted to the American Association for the Study of Liver Disease, were evaluated by three "blind" reviewers for originality, design-execution, importance, and overall scientific merit. Of the 77 abstracts rated for all parameters by all observers, interobserver agreement ranged between 81 and 88%. However, corresponding intraclass correlations varied between 0.16 (approaching statistical significance) and 0.37 ($p < 0.01$). Specific tests of systematic differences in scoring revealed statistically significant levels of observer bias on most of the abstract components. Moreover, the mean differences in interobserver ratings were quite small compared to the standard deviations of these differences. These results emphasize the importance of evaluating the simple percentage of rater agreement within the broader context of observer variability and systematic bias.

INTRODUCTION

As noted recently by Conn (1), the blind evaluation of abstracts has been tried as a selection procedure by a number of professional organizations, e.g., the American Federation for Clinical Research (2), the American Association for the Study of Liver Disease, the European Association for the Study of the Liver, and the Society for Investigative Dermatology. Nonetheless, the specific procedures involved have never been described in detail, nor have the results been presented in formal, systematic statistical fashion.

This investigation was undertaken to assess the extent of observer agreement and systematic bias in the blind review of abstracts submitted for presentation at an annual meeting of the American Association for the Study of Liver Disease (AASLD). In a recent editorial published by Conn in 1974 (1) the clinical and subjective aspects of the study were described in some detail, e.g., the procedures used to select a program of 40 abstracts for presentation; the criteria for balancing the program; the regional composition of the abstract submitters; etc. The present report, in contrast, will focus upon a critical statistical evaluation of the more objective numerical aspects of the review process.

The purpose of this report is to give a statistical answer to two questions which

[1]Liver Seminar Series. This article is the ninth in a series entitled "Seminars on Liver Disease" that has been presented as part of the Training Program in Liver Disease at the Veterans Administration Hospital, West Haven, Conn. Dr. Harold O. Conn, Professor of Medicines Yale University School of Medicine, Director of the Training Program in Liver Disease, is Guest Editor. A portion of this work was presented at the American Statistical Association Meeting in St. Louis, Mo., August 1974.

[2]Dr. Cicchetti is a research psychologist and biostatistician at the West Haven VAH and Associate Yale University School of Medicine.
requests should be sent to the V.A. address.

[3]Dr. Conn is Chief of the Liver Disease Program at the West Haven VAH and Professor of Medicine, Yale University School of Medicine.

characterize clinical studies of observer variability, irrespective of the specific area under investigation. The first question posed by the clinical investigator is: To what extent do pairs of observers agree about the phenomenon under study, e.g., how often will two endoscopists agree in making the diagnosis of esophageal varices? The second question is of much clinical interest: Are the disagreements between independent observers random or systematic, e.g., does one observer consistently overdiagnose esophageal varices? The statistical procedures introduced have relevance for answering a wide range of other questions, as well as in the assessment of observer variability. These procedures require only that the data be measurable on an *ordinal* or *interval* scale, in contrast to *nominal* or categorical scales. These scales are defined here in the manner of Stevens (3).

MATERIALS AND METHODS

One-hundred nine abstracts submitted for consideration for presentation at the 1974 national meeting of the AASLD were evaluated blindly by three independent observers (A, B, and C). Each abstract was rated on a 0 to 100 scale for originality, design–execution, and importance. In addition, observers were asked to use the same scale to judge *overall scientific merit*. The ratings of three measures of quality were presented to the reviewers as follows:

Originality. A completely new innovative use of current knowledge should rate a very high score, e.g., 80 to 100. The most original concept imaginable should be rated 100. Projects devoid of originality should receive very low grades. Studies of intermediate originality should receive appropriate intermediate grades.

Design–Execution. A thoughtful, innovative, well-controlled experimental design executed using sound, carefully calibrated techniques and assessed using appropriate statistical tests should rate very highly. Poorly designed, poorly performed, and poorly evaluated design and execution should rate very low. Intermediate design and execution should be rated on a continuum between these extremes.

Importance. Experiments productive of very useful clinical or laboratory knowledge should receive very high grades. Trivial experiments providing no useful or potentiallly useful information should be scored very low. Studies of intermediate importance should receive intermediate ratings.

In addition, observers were asked to use the same scale to judge *overall scientific merit*. They were told that the overall merit score should not be determined by simply averaging the other three ratings. The same abstracts were also rated independently by three nonblind reviewers (D, E, and F). This report will focus only upon the blind reviews. The reviewers did not rate abstracts originating from their own institutions, in order to control for this possible source of bias.

Two methods were applied to assess the extent of reviewer agreement: the simple percentage of agreement, and the intraclass correlation coefficient.

l. Simple Percentage of Agreement

Two basic types of assessment were made: (a) the extent of agreement among each of the three pairings of observers (A vs B; A vs C; B vs C) by each of the four abstract components (this yielded 12 evaluations); and (b) the average or composite agreement[4] among the three observers (A, B, and C) for each abstract quality (this

[4]The rationale for using a composite index is that it provides a single, overall score reflecting the extent of agreement among three or more observers. This contrasts with the more specific pair by pair agreement indexes. Both indexes are rather commonly reported in observer variability studies in medicine.

produced four additional evaluations). These forms of assessment will now be discussed in more detail.

(a) Paired agreement. The formula utilized here was the standard

$$\text{Paired agreement (\%)} = 1 - (\Sigma \mid D \mid / \Sigma \mid D_{\max} \mid) \times 100,$$

in which $\Sigma \mid D \mid$ refers to the sum of the absolute differences in the paired ratings of two observers, over the number of abstracts reviewed and $\Sigma \mid D_{\max} \mid$ refers to the maximum value $\Sigma \mid D \mid$ possible for a given set of data.

As an example of how the formula is applied, suppose two independent reviewers rate 100 abstracts for overall scientific merit using our 0 to 100 scale. The maximum disagreement possible is $100 \times 100 = 10{,}000$. If, now, the sum of absolute differences for these observers across the 100 abstracts were 2000, then the simple percentage of agreement would become 80% as shown below:

$$\begin{aligned} \text{Observer agreement} &= 1 - (\Sigma \mid D \mid / \Sigma \mid D_{\max} \mid) \times 100 \\ &= 1 - (2000 / 10{,}000) \times 100 \\ &= 80\% \end{aligned}$$

This simple formula has been used since 1957 when it was presented by Robinson (4) and, in fact, is mathematically equivalent to the formula used for determining observer agreement, not only with interval data (as in our research) but also with nominal (5–8) or ordinal (9–16) data.

Of the 109 abstracts submitted, we analyzed only those 77 abstracts which were rated on all four attributes by all three blind and all three nonblind reviewers.[5]

(b) Agreement among all three reviewers. This composite assessment was determined by simply averaging the percentages of agreement among each of the three possible rater pairings for each abstract component. Although the percentages of both composite agreement and agreement among specific pairs of raters are very simple to calculate and interpret, we recognized from the onset that considered alone, these crude percentages could be quite misleading. For example, suppose two independent reviewers evaluate the same abstracts and obtain a seemingly high percentage of agreement, e.g., 85%. It is nonetheless quite possible that individual variations on given abstracts might be enormous. Moreover, consistent overreading or underreading by a given rater would also not be detected by this statistical method.

One method which obviates these serious problems is the *intraclass correlation coefficient* (R_1) described by a number of investigators (4, 7, 16–23). The intraclass r (R_1) like the Pearsonian r, can vary between -1 and $+1$. However, the higher the R_1 scores, the more the observer agreement. It should be stressed here that the Pearsonian r is often used to assess rater agreement with ordinal or continuous data, in spite of the fact that it is inadequate as a measure of agreement (4, 9). The basic problem with this type of analysis is that it takes into account only the extent to which two independent observers put their measurements of the same subjects in the same order but tends to ignore the magnitude of the discrepancy on individual pairs

[5]It should be noted that the results of all the statistical analyses were virtually the same, whether or not incomplete abstract reviews were included in the data. However, since the research design is one of repeated measures (the three raters) on the same variables (the four abstract qualities), it was decided to include only those abstracts for which there were complete reviews. This would then facilitate a comparison of blind and nonblind reviews in a future investigation.

of ratings. As a result, raters can be far apart on individual measurements, but, as long as the trend of their rankings is similar, the extent of correlation will be high, giving the impression of better agreement than actually exists. R_I is a valid measure of agreement, rather than one of mere association as is true of the Pearsonian r.

The formula for the interrater R_I, based upon the simple two-way, repeated measures analysis of variance design (ANOVA), and due to Bartko (17, 18), can be expressed as:

$$R_I = [ms\ S - \text{ms}\ E] / [ms\ S + \text{ms}\ E\,(O - 1) + O\,(ms\ O - ms\ E) / S],$$

in which *ms S* refers to the mean square (or variance) due to differences among the subjects (in our case abstracts) being rated; *ms O* denotes the mean square (or variance) due to observers; *ms E* refers to the mean square (or variance) due to the interaction between *S* and *O*; *O* alone refers to the number of observers; *S* by itself denotes the number of subjects.

It should be noted that *ms* in the above formula is nothing more than the variance which, in other contexts, is symbolized by the much more familiar s^2. What the ANOVA model allows us to do is to identify and quantify the several sources of s^2. One is due to the inherent variability among the abstracts being reviewed (*ms S*). A second is due to the variability caused by differences between the reviewers (*ms O*). The third (*ms E*) is that source of s^2 that remains after both *ms S* and *ms O* are taken into account. It is the variability (or s^2) in evaluating abstracts which is caused by unknown factors, e.g., specific idiosyncratic preferences or dislikes of different reviewers for different abstracts. The formula for R_I, above, allows one to evaluate the relative importance of each of these three sources of variability in a given situation. In the *ideal* situation, *ms S* would be very high relative to both *ms O* and *ms E*. This, in turn, would result in very high interrater agreement (reflected in an R_I value approaching +1) and no appreciable rater bias. Such an ideal result would, in our application, tell us that mostly all the variance in abstract reviewing is due to differences in the quality of the abstracts themselves and is neither a function of differences in reviewers nor of other as of yet unknown causes.[6]

R_I as defined above, has the distinct advantage over the Pearsonian r, in that it allows the investigator to assess how much of the variability among independent observers is due to the raters themselves and how much is a function of differences among the subjects (in our case, abstracts) being rated. Also, it is simple to calculate and simple to interpret. The significance of R_I is assessed by referring the value of the F ratio for subjects (S) to a standard ANOVA table with degrees of freedom (df) in the numerator associated with S, and the df in the denominator associated with E. It should be stressed that, unlike the simple percentage of agreement, the value of R_I is indeed affected by the extent of observer bias in the ratings under investigation. Thus, with the level of agreement held constant the more the tendency of one observer to evaluate abstracts consistently more stringently than a second observer, the lower the value of R_I. The value of R_I is also adversely affected by the magnitude of the variability of the paired differences among any set of observers rating the same clinical phenomena. Thus, the same level of simple percentage of observer agreement will be associated with varying values of R_I depending upon the extent of this source of variance. The logic here is that in the ideal case we would expect the following: (a) R_I to be high (see (19) for the recommendation of ≥ 0.75); and (b) the F

[6]For a more thorough description of various R_I models, depending upon the specific type of research question, see, most recently, Ref. (7) and (16).

ratio for abstracts to be much greater than that due to observers. In this type of comprehensive analysis, these results would tell us that the major source of variance (or variability) in reviewing medical abstracts is due to differences among the scores given to abstracts themselves rather than to differences in scoring between the two raters. It should be stressed that there is a direct mathematical relationship between the value of F and that of t. This holds in the simplest case of comparing two observers in the scoring of a given number of abstracts. If we did a paired t test, the value of t we would obtain would exactly equal the value of the square root of the F ratio for raters, if we had instead performed the appropriate analysis of variance on the same data. The problem with the paired t test is that it does not allow one to compare the variability due to raters with that due to abstracts, and it is this relative concept that we are most interested in investigating.[7]

In order to compare reviewer agreement based upon R_I with that based upon the percentage of agreement, the same two types of assessment will again be made (i.e., paired agreement and composite agreement).

2. Reviewer Bias

The overall level of reviewer bias was tested by the same, simple repeated measures analysis of variance used for computing R_I. In this model, if the variance due to observers is statistically significant, it tells us that there is an overall tendency for raters to be biased with respect to each other. With respect to bias vis-à-vis Rater A vs B, A vs C, and B vs C, the most widely used test here is the simple paired t test, comparing each pair of raters on each abstract component. However, recent work (25) indicates clearly that the t test (paired or unpaired) is only valid for sets of data involving only two groups, e.g., abstracts rated by Reviewers A and B only. When more than two groups are being studied, the probability of a Type I error (falsely claiming statistical significance) increases as a direct function of the number of groups involved. In order to control for this source of error, a number of tests have been developed. Petrinovich and Hardyck (25) show, empirically, that the one developed by Tukey (26–28) most adequately controls for this source of error. This standard test, in effect, makes the results comparable to what one would have obtained had the t test been modified to fit the situation of multiple group comparisons.[8]

A final assessment concerned the extent to which the evaluations a specific reviewer made on one abstract component were related to the evaluations he made on the remaining abstract components. For example, are high ratings on overall merit associated also with high scores on originality? The standard Pearsonian product moment correlation was utilized here.

RESULTS

Reviewer Agreement

1. Simple percentage of agreement. The range of paired, interobserver agreement for any given abstract component is between 81 and 88% (Table 1). Moreover, the variability between any pair of observers, on any given abstract component, never differed by more than 4%. Composite rater agreement, or the

[7]For further information concerning these relationships and their bearing upon the interpretation of medical data, the interested reader should consult Amenta (24, pp. 145–170).

[8]Snedecor (28, p. 251) gives the mathematical relationship between the Tukey multiple-range test and the t test (paired or unpaired) when the number of groups (e.g., raters) being compared is restricted to two only. The point is that only under conditions in which the t test is appropriate will it produce results identical to that obtained by applying the Tukey test to the same data.

TABLE 1
Simple, Unadjusted Mean Percentages of Reviewer Agreement[a]

Abstract component	Reviewer A vs B (%)	Reviewer A vs C (%)	Reviewer B vs C (%)	Average, composite agreement among reviewers A, B, & C (%)
Originality	88	85	84	86
Execution	86	84	86	85
Importance	85	82	81	83
Overall merit	84	84	82	83

[a]Each comparison is based upon a complete evaluation of 77 abstracts. This is also true of Tables 2 through 6.

average percentage of agreement over the three possible rater pairings, varied between 83 and 86% across the abstract components.

2. *Intraclass correlation coefficients.* The 12 paired values of R_1 are given in Table 2. These coefficients range between 0.16 ($p < 0.10$) and 0.37 ($p < 0.01$) which fall far short of the ≥ 0.75 recommended by Burdock *et al.* (19). The overall, composite R_1 coefficients were more uniform than those between individual pairs of raters. These coefficients, also given in Table 2, ranged between 0.24 and 0.30. Thus, both paired and composite agreement, as measured by R_1, was statistically adequate but clinically inadequate.

Reviewer Bias

The great inconsistency between the high percentages of agreement and the low intraclass r's strongly suggested a considerable amount of interobserver variation in specific abstract ratings. This suspicion was tested by comparing the means of interobserver differences to the respective standard deviations of these differences. As shown in Table 3, the 12 standard deviations of the differences were, without exception, much higher than the means of these differences. Specifically, the ratio of the standard deviation to the mean (the so-called coefficient of variation, when multiplied by 100) ranged between 1.55 and 68.13, indicative of extensive interobserver variability in agreement on individual abstracts.

In addition, an examination of mean abstract ratings for each component (Table 4) revealed results suggestive of rater bias. Specifically: (i) Reviewer B assigns higher scores than either reviewer A or C on each of the four abstract components; and (ii)

TABLE 2
Intraclass Correlation as a Measure of Reviewer Agreement

Abstract component	Reviewer A vs B	Reviewer A vs C	Reviewer B vs C	Composite rater agreement
Originality	0.37***	0.21**	0.32***	0.30***
Execution	0.28***	0.21**	0.34***	0.29***
Importance	0.22**	0.15*	0.31***	0.24***
Overall merit	0.16*	0.18*	0.33***	0.24***

*$p < 0.10$.
**$p < 0.05$.
***$p < 0.01$.

TABLE 3
Mean Differences in Reviewer Agreement Compared to Standard Deviations of These Differences

Abstract component	Mean difference	SD of difference	Ratio of SD to mean difference
Reviewer A vs Reviewer B			
Originality	2.47	15.51	6.28
Execution	9.54	14.80	1.55
Importance	6.43	18.86	2.93
Overall merit	7.08	19.40	2.74
Reviewer B vs Reviewer C			
Originality	3.38	20.46	6.05
Execution	4.80	19.46	4.05
Importance	9.51	23.16	2.44
Overall merit	6.78	21.38	3.15
Reviewer A vs Reviewer C			
Originality	0.91	19.17	21.07
Execution	4.74	18.64	3.93
Importance	3.08	22.76	7.39
Overall merit	0.30	20.44	68.13

Reviewers A and C seem to differ appreciably on only one abstract component, design–execution. The data were analyzed statistically to answer two basic questions: (i) Is there an overall tendency for one rater to be consistently more lenient than the other reviewers? (ii) Is there evidence of rater bias when pairs of reviewers are compared to each other? Each of these questions will be answered in turn.

The overall level of reviewer bias was tested by the same, simple, standard, two-way, repeated measures analysis of variance used for computing each of the composite R_1 values given in Table 2. The results (Table 5) indicate bias among the three raters, separately, for each of three abstract components at the following levels of statistical significance: design–execution and importance at $p < 0.001$ and overall scientific merit at $p = 0.004$. The Tukey modification of the paired t test (26–29) was applied to compare the three reviewers with each other. For three abstract qualities, Reviewer B assigned more favorable ratings ($p < 0.05$) than did Reviewer A or C; Reviewer A give significantly harsher ratings on the abstract quality design–execution than did Reviewer C; and Reviewers A and C showed no bias on evaluating abstracts for originality, importance, or overall scientific merit. The specific ratings given by each reviewer for each abstract component are presented in Table 4.

TABLE 4
Extent of Reviewer Bias[a]

Abstract component	Reviewer A		Reviewer B		Reviewer C		Grand Mean
	Mean	SD	Mean	SD	Mean	SD	
Originality	65.26	10.69	67.73	16.50	64.35	18.78	65.78
Execution	65.97	10.03	75.52	15.53	70.71	18.55	70.74
Importance	64.80	11.58	71.23	18.32	61.73	21.88	65.92
Overall merit	63.70	11.41	70.78	18.15	64.00	19.50	66.16

For all four abstract qualities, Reviewer B assigned more favorable ratings than did Reviewer A or C. This reached statistical significance ($p < 0.05$) for execution, importance, and overall merit. Reviewers A and C differed significantly on only one abstract component, execution ($p < 0.05$). The term bias is used here to refer to the predominantly systematic variation in abstract reviewing.

TABLE 5
Overall Tests of Reviewer Bias

Source	*df*	*ss*	Variance or *ms*	*F*	*p*
Originality					
Raters (R)	2	470.128	235.064	1.36	0.26
Abstracts (A)	76	30591.000	402.513		
R × A	152	26346.293	173.331		
Total	230	57407.421			
Execution					
Raters (R)	2	3508.009	1754.004	10.99	<0.001
Abstracts (A)	76	28556.438	375.743		
R × A	152	24258.414	159.595		
Total	230	56322.861			
Importance					
Raters (R)	2	3623.458	1811.729	7.61	<0.001
Abstracts (A)	76	36827.813	484.576		
R × A	152	36200.316	238.160		
Total	230	76651.587			
Overall scientific merit					
Raters (R)	2	2467.697	1233.848	5.84	0.004
Abstracts (A)	76	32536.688	428.114		
R × A	152	32113.387	211.272		
Total	230	67117.772			

With respect to the final question, are high ratings on one abstract component also associated with high ratings on other components, the data indicate that the correlations among scores for each possible pairing of abstract components were all statistically significant ($p < 0.001$). Moreover, this was true for each of the three raters considered separately (Table 6). The correlation of each abstract component with the rating of overall scientific merit was higher than any other correlation. Of particular interest is the fact that the ordering of the correlation coefficients was the same for each rater, with the highest correlation being between overall score and importance (0.89 for Rater A, 0.90 for C, and 0.96 for B) and the lowest being between overall score and originality (0.55 for A, 0.70 for B, and 0.75 for C). Correlations between overall score and design–execution were intermediate in magnitude (0.72 for A, 0.78 for C, and 0.84 for B). These data indicate that the various abstract components were *not* considered as independent qualities by the three reviewers.

DISCUSSION

The results of this investigation underscore the importance of evaluating the simple percentage of agreement within the broader context of observer variability

TABLE 6
Intrarater Correlations between Overall Scientific Merit and Scores on Originality, Execution, and Importance in the Blind Review of AASLD Abstracts

Reviewer	Overall vs originality	Overall vs execution	Overall vs importance	Originality vs importance	Execution vs importance	Originality vs execution
A	0.55	0.72	0.89	0.40	0.53	0.66
B	0.70	0.84	0.96	0.63	0.80	0.70
C	0.75	0.78	0.90	0.56	0.65	0.68

TABLE 7
Effects of Reduction in Variability upon Observer Agreement and Bias in the Blind Review of Abstracts Submitted to the AASLD: Overall Merit

Source of comparison	*N*	Agreement (%)	Intraclass *r*	*p* of bias
Reviewer A vs Reviewer B				
All abstracts	77	84	0.16	0.003
Abstracts yielding ≥80% agreement	57	90	0.40	0.004
Reviewer B vs Reviewer C				
All abstracts	77	82	0.33	0.008
Abstracts yielding ≥80% agreement	52	90	0.79	0.11
Reviewer A vs Reviewer C				
All abstracts	77	84	0.18	n.s.[a]
Abstracts yielding ≥80% agreement	50	92	0.73	n.s.

[a] n.s. means does not reach statistical significance at or beyond the 0.05 level.

and systematic bias. Specifically, the data show that the percentage of observer agreement, considered in isolation, can be quite misleading. Thus, one might conclude that rater agreement varying between 81 and 88% is impressively accurate. One cannot accept such a conclusion, however, until individual rater differences have been evaluated further. An ideal index of observer agreement is dependent on a high intraclass correlation coefficient, e.g., $R_I \geq 0.75$ as recommended by Burdock *et al.* (19). In this study, the R_I values were far below ideal levels.

In an attempt to explore this issue further, each abstract for which there was less than 80% interrater agreement on overall scientific merit was removed from the data. The resulting data were then reanalyzed in order to note what effect this reduction in variability would have upon the size of the intraclass *r*, the percentage of agreement, and the extent of reviewer bias. This analysis revealed that the increases in percentage of agreement were quite similar for each pair of raters. They ranged between 6 and 8% (Table 7). The intraclass *r*, as expected, also increased for each rater comparison. It is interesting to note that the increase was least for the case in which rater bias was greatest, despite the removal from the data of abstracts for which there was < 80% agreement. Both the B–C comparison (in which bias dropped from the $p < 0.01$ to the $p > 0.05$ level) and the A–C reviewer pairing (for which there was no significant bias either for the full set of rated abstracts or for those yielding at least 80% agreement) showed much more dramatic increases in the size of the intraclass *r*. In fact, in the former case the increase in R_I meets the criterion of ≥ 0.75 set by Burdock *et al.* (19).

Do these observations have practical significance? We believe that these types of analysis are important in a number of ways. First, they permit identification of those abstracts which are responsible for the major source of variability in the evaluation process. Thus, for example, Table 7 shows that reviewers B and C manifest little agreement ($R_I = 0.33$) and appreciable bias ($p < 0.01$) in the rating of 77 abstracts, indicating that the 82% agreement level is misleading. By deleting the 25 abstracts which yielded <80% agreement, R_I increases dramatically to 0.79, reviewer bias disappears, and the high percentage of agreement (90%) thereby becomes clinically significant, as well as statistically significant. Identification of the controversial abstracts permits abstracts in question to be critically scrutinized (Table 7). Second,

the analyses of rater biases show that some observers are intrinsically easy graders. Rater B, for example, was consistently more generous than the other two; Rater A was harsher than Rater C (Table 4). Furthermore, Rater B was a more variable rater than A and C, who were more consistently in accord with each other (Table 4). Identification of intrinsically biased or inconsistent raters permits their exclusion from the selection process. Finally, the recognition that the overall score was the dominant one (Table 6) is extremely useful information. It could simplify the rating process by eliminating the individual component evaluations and requiring only a single overall score. On the other hand, it is possible that despite directions to the contrary, the overall score represented a consensus or "loose" mean score of the individual components. Indeed, two of the three raters felt that their overall score was to a large extent determined by their component scores. One might thus require the rating of multiple components but would base abstract selection primarily on the overall score.

The ability to identify controversial abstracts, biased or variable observers, and other factors responsible for rater disagreement permits re-evaluation of difficult abstracts, exclusion of variable observers, and elimination of other discordant factors. The recognition of the pathophysiology of the selection process permits the design of more simple and precise procedures for abstract evaluation. Such factors are of more than passing interest, since the selection process determines to a large extent what information is to be presented, which in turn may affect subsequent trends in investigation.

ACKNOWLEDGMENTS

The authors acknowledge the careful blind and open-eyed abstract assessment of Drs. Burton Combes, Robert L. Scheig, Steven Schenker, and Hyman J. Zimmerman.

REFERENCES

1. Conn, H. O., An experiment in blind program selection. *Clin. Res.* **22,** 128–134 (1974).
2. Ball, M. F., Selection of programs. *Clin. Res.***20,** 318 (1972).
3. Stevens, S. S., Mathematics, measurement, and psychophysics. *In* "Handbook of Experimental Psychology." (S. S. Stevens, Ed.), pp. 1–49. Wiley, New York, 1951.
4. Robinson, W. S., The statistical measurement of agreement. *Amer. Sociol. Rev.* **22,** 17–25 (1957).
5. Cohen, J., A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20,** 37–46 (1960).
6. Fleiss, J. L., "Statistical Methods for Rates and Proportions." Wiley, New York, 1973.
7. Fleiss, J. L., Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* **31,** 651–659 (1975); also presented at the Joint Meetings of the ASA, New York, 1973.
8. Fleiss, J. L., Cohen, J., and Everitt, B. S., Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **72,** 323–327 (1969).
9. Cicchetti, D. V., A new nonparametric statistic for measuring agreement between two ordinal variables. *Biometrics* **27,** 478 (1971).
10. Cicchetti, D. V., Determining observer variability when the data are ordinal: An alternative to Kendall's Tau. *Biometrics* **28,** 260 (1972).
11. Cicchetti, D. V., and Allison, T., A new procedure for assessing reliability of scoring EEG sleep records. *Amer. J. EEG Technol.* **11,** 101–109 (1971); also, *Proc. Electro-Physiol. Technol. Ass.* **18,** 203–213 (1971).
12. Cicchetti, D. V., and Allison, T., Assessing the reliability of scoring EEG sleep records: An improved method. *Proc. Electro-Physiol. Technol. Ass.* **20,** 92–102 (1973).
13. Cicchetti, D. V., Keitges, P. W., and Barnett, R. N., Effects of and justification for reducing the number of unknown syphilis serology specimens from 65 to 20. Presented at Joint Meetings of the ASA, New York, 1973.
14. Cicchetti, D. V., Keitges, P. W., and Barnett, R. N., How many is enough? A statistical study of proficiency testing of syphilis serology specimens. *Health Lab. Sci.* **11,** 299–305 (1974).
15. Cicchetti, D. V., and Fleiss, J. L., A comparison of the null distributions of weighted kappa and the C

ordinal statistic. Presented at the Joint Central Regional Meetings of the ASA, St. Paul, Minnesota, 1975; also *Appl. Psychol. Meas.*, in press (1976).
16. Fleiss, J. L., and Cohen, J., The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **33,** 613–619 (1973).
17. Bartko, J. J., The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* **19,** 3–11 (1966).
18. Bartko, J. J., Corrective note to: The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* **34,** 418 (1974).
19. Burdock, E. I., Fleiss, J. L., and Hardesty, A. S., A new view of interobserver agreement. *Pers. Psychol.* **16,** 373–384 (1963).
20. Ebel, R. L., Estimation of the reliability of ratings. *Psychometrika* **16,** 407–424 (1951).
21. Guilford, J. P., "Fundamental Statistics in Psychology and Education." McGraw-Hill, New York, 1950.
22. Haggard, E. A., "Intraclass Correlation and the Analysis of Variance." Dryden Press, New York, 1958.
23. Horst, P., A generalized expression for the reliability of measures. *Psychometrika* **14,** 21–31 (1949).
24. Amenta, J. A., Analysis of variance for the clinical laboratory. *In* "Statistical Methods in the Clinical Laboratory" (R. N. Barnett, Ed.), pp. 145–170. American Society of Clinical Pathologists, Chicago, 1968.
25. Petrinovich, L. F., and Hardyck, C. D., Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychol. Bull.* **71,** 43–54 (1969).
26. Tukey, J. W., "The Problem of Multiple Comparisons." Privately circulated monograph, pp. 1–396, 1953.
27. Snedecor, G. W., "Statistical Methods," 5th Ed. Iowa State College Press, Ames, Iowa, 1956.
28. Winer, B. J., "Statistical Principles in Experimental Design." McGraw-Hill, New York, 1962.