DOCUMENT RESUME

ED 287 175	CS 210 837

AUTHOR	Atkinson, Dianne; Murray, Mary
TITLE	Improving Interrater Reliability.
PUB DATE	20 Mar 87
NOTE	17p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (38th, Atlanta, GA, March 19-21, 1987).
PUB TYPE	Viewpoints (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Evaluation Methods; *Holistic Evaluation; *Interrater Reliability; *Measurement Techniques; Research Methodology; *Statistical Analysis; Writing (Composition); *Writing Evaluation

ABSTRACT

Noting that improvement in rater reliability means eliminating differences among raters, this paper discusses ways to assess writing evaluator reliability and methods for achieving higher levels of interrater reliability. After showing that reliability can be improved two ways--by increasing the number of raters or measurements made, and by increasing the systematic variance among essays relative to error variance--the paper cites common problems in reporting and assessing reliability. The paper then recommends that researchers (1) use an "analysis of variance" approach in assessing reliability; (2) indicate the number of independent observations; (3) use a two-way analysis of variance if more than one dimension is rated; (4) use "repeated measures" analysis of variance if rating more than one sample per student; and (5) use an "intraclass correlation coefficient" such as coefficient alpha in reports of research, or the "Pearson r" when two raters rate one dimension of the sample. Finally, the paper describes methods to increase interrater reliability such as controlling the range and quality of sample papers, specifying the scoring task through clearly defined objective categories, choosing raters familiar with the constructs to be identified, and training the raters in systematic practice sessions. (Formulas for calculating reliability and training procedures for raters are included.) (JG)

Dianne Atkinson
Educational Psychology Section/SCC-G, Purdue University

Mary Murray
Rhetoric and Composition Studies/Heavilon, Purdue University

Conference on College Composition and Communication
Westin Peachtree Plaza
Atlanta, Georgia
March 20, 1987

IMPROVING INTERRATER RELIABILITY

In our presentations today on improving interrater reliability, we will be dividing our concerns into two major categories. First, I will discuss appropriate and meaningful ways to assess reliability. Secondly, Mary will present important considerations for achieving higher levels of interrater reliability.

An informal survey we conducted on the reporting of reliability in the journal *Research in the Teaching of English* over the last five years has convinced us that we need to move toward more uniform and interpretable reporting of reliability.

Our specific recommendations today about reporting reliability grow out of a "theory of measurement" perspective in which reliability is conceptualized as variance. Specifically, reliability is defined as a ratio of true variance to total variance, as shown in the transparency (A). That is, reliability is the ratio of variance due to real differences divided by variance due to both real differences and error. Measurement with zero error would then have a reliability of one.

As indicated on the transparency (B), reliability is calculated as as a ratio of estimated true variance (between-product variance minus error variance) to total variance (between-product variance plus error variance). A correlation coefficient calculated in this way is known as an "intraclass correlation coefficient" and is familiar to us in the context of reliability assessment as a "coefficient alpha," also presented on the transparency (B). Coefficient alpha or intraclass correlation coefficients can be calculated directly from an analysis of variance summary table as indicated on the transparency (C).

In other words, reliability is subtracting out differences among raters, in order to find out how much of the total variance is due to "true" differences among essays. The reliability of a measure is the ratio of true or systematic variance to total variance in ratings associated with an "average" or composite single rater. If the ratings of all raters can be used to estimate true scores, then the reliability estimated by the intraclass correlation coefficient is increased to reflect the greater stability and accuracy of rater means. The Spearman-Brown prophecy formula estimates this increased reliability which is due to the utilization of multiple ratings. The calculations on the transparency (C) show how the reliability coefficient shifts from .71 in this case to over .90 when four raters are used rather than just one.

Now we are in a position to anticipate the kinds of suggestions that Mary will make for increasing the level of interrater reliability achieved. According to the Spearman-Brown prophecy formula--shown on transparency (D), we have two options: one--increasing the number of raters or measurements that we make--that is "K" in the formula, and two--increasing essay or product variance relative to error variance--that is the term rii. Besides recruiting more raters then, we can take steps that increase systematic variance and decrease random or error variance in order to achieve more reliability. Here in fact we have a recipe for increasing reliability.

However, we want to emphasize that it is the analysis of variance approach that helps us see the effects of taking such steps as increasing the number of raters, and decreasing error variance by facilitating the perception of systematic differences.

Instead of a composite index which reflects the impact of all factors affecting reliability, an analysis of variance approach allows us --to some extent-- to examine the relative importance of various factors contributing to achieved reliability.

Furthermore, the very activity of setting up an analysis of variance table forces us to carefully identify the various components of our measurement design. An investigator must specify the number of independent observations, the number of dimensions assessed, and the number of products per subject included. THESE SAME ITEMS OF INFORMATION SHOULD ACCOMPANY ANY RELIABILITY REPORT IN THE LITERATURE. Otherwise, interpretation of the numerical value provided may be difficult.

To illustrate some common difficulties that frequently arise in assessing reliability or "measuring measurement," I've pulled out some typical descriptions from recent issues of Research in the Teaching of English. I've omitted citations not just in the interest of collegial harmony but also because these instances are not atypical--additional examples could easily be added. On the transparency (E) you see four statements about reliability assessment.

Beginning our discussion at the top of the list, let's examine the statement that "Each essay was read by at least two raters." Although the implication that "extra" raters were used as needed seems initially comforting, a serious problem is generated for calculating a useful measure of reliability in such a case. We need to know the exact number of raters.

The same problem is compounded in the next example: "In cases where the scores differed by more than two points, a a third rater was used and the extreme score dropped." Such an approach to rater-disagreement has the effect of leveling final scores and results in an information loss, as well as rendering problematical the meaningful assessment of reliability. If reliability is calculated anyway, an inflated value will result. Any rater-exchange must be related to factors extraneous to the rating situation--such as a rater dropping out because his four-year-old has the chicken-pox. This kind of rater-switch may well lower reliability, but does not constitute systematic biasing.

Our third selection--"Reliability was .97"--is one of our favorites although it probably just represents an oversight. It is of course essential to identify the way in which reliability was calculated. We do not know--although we hope the researcher does--whether the .97 represents a proportion of agreement, or a coefficient of some kind.

Our final quote reports that "Correlations among the three pairs of raters were .75, .84, and .79." First, these ARE reasonable levels--remember Diederich's forceful assertion that it's very hard to get more that a Pearson r of .70 for two raters holistically rating essays. However, the situation is one in which three raters were used, while reliability was only assessed two-raters-at-a-time, so reliability is probably underestimated.

To address these problems and to help both researchers and consumers of research meaningfully interpret reliability measures, we make the following recommendations shown on the transparency (F):

RECOMMENDATIONS FOR CALCULATING AND REPORTING RELIABILITY ESTIMATES

A. Use an "analysis of variance" approach in assessing reliability.

a. Indicate number of independent observations. If pairs of raters confer before giving a rating, N = number of pairs. If raters work alone while rating, even though they train with other raters, or receive periodic feedback, N = number of raters.

b. Number of dimensions assessed. If more than one dimension is rated, such as both "quality of ideas" and "correctness," use a two-way analysis of variance.

c. Number of essays per student. If more than one sample of writing is used to estimate achievement, use "repeated measures" analysis of variance.

B. Use an "intraclass correlation coefficient" such as coefficient alpha in reports of research. In the special case of two raters rating one dimension of the product for one product per student, the familiar Pearson r is an equivalent measure of reliability. Both coefficient alpha and the Pearson coefficient of correlation can be readily generated by such widely available software packages as SPSSX--the Statistical Package for the Social Sciences, recently EXpanded.

MARY WILL NOW DISCUSS PROCEDURAL STRATEGIES FOR INCREASING INTER-RATER RELIABILITY.

CALCULATING RELIABILITY

$$r_{ii} = \frac{\text{TRUE VARIANCE}}{\text{TOTAL VARIANCE}}$$

transparency (A)

$$r_{ii} = \frac{\text{VARIANCE BETWEEN} - \text{VARIANCE WITHIN}}{\text{TOTAL VARIANCE}}$$

+ transparency

INTRACLASS CORRELATION COEFFICIENT

$$\text{ICCC} = \frac{\text{VARIANCE BETWEEN} - \text{VARIANCE WITHIN}}{\text{VARIANCE BETWEEN} + (\text{ave.\#cases per class} - 1)\ \text{VARIANCE WITHIN}}$$

(B)

COEFFICIENT ALPHA (CRONBACH'S ALPHA)

$$\text{ALPHA} = \frac{(\text{number of raters})\ (\text{average interrater correlation})}{1 + (\text{average interrater correlation})\ (\text{number of raters})}$$

(B)

THE SPEARMAN-BROWN FORMULA

$$r_{kk} = \frac{k\ r_{ii}}{1 + (k - 1)\ r_{ii}}$$

transparency

RELIABILITY ASSESSMENT WHEN MORE THAN TWO RATERS ARE USED

Numerical Example:
(See Winer, pp. 288-289)

ANALYSIS OF VARIANCE

Source of variation		SS		df	MS
Between essays		122.50		5	24.50
Within essays		36.00		18	2.00
Between judges	17.50		3		5.83
Residual	18.50		15		1.23
TOTAL		158.50		23	

INTRACLASS CORRELATION COEFFICIENT:

$$r_i = \frac{\text{Variance between} - \text{Variance within}}{\text{Variance between} + (\text{ave.\# cases per class} - 1)\text{Variance within}}$$

$$= \frac{24.50 - 2.00}{24.50 + (4-1)\ 2.00}$$

$= .7377$ reliability coefficient of single judgment

If mean of all four judges is used, reliability is higher.
Using Spearman-Brown formula:

$$r_4 = \frac{4\ (.7377)}{1 + (4-1)\ (.7377)}$$

$= .9184$ reliability coefficient for mean of four judgments.

Transparency (E)

A NON-RANDOM SAMPLE OF STATEMENTS ABOUT RELIABILITY ESTIMATION TAKEN FROM THE LAST FIVE YEARS OF RESEARCH IN THE TEACHING OF ENGLISH

1. Each essay was read by at least two raters.

2. When scores differed by more than two points, a third rater was used and the extreme score dropped.

3. Reliability was .97.

3. Correlations among the three pairs of raters were .75, .84, and .79.

A SELECTED LIST OF RESOURCES FOR RELIABILITY ISSUES:

Measurement and Reliability in Education & Composition Studies:

Asher, W.J. (1967). *Measurement in educational research.* Educational research and evaluation methods. Little, Brown.

Cooper, C.R. & Odell, L. (1977). *Evaluating writing: Describing, measuring, judging.* Urbana, Il: NCTE.

Diederich, P. (1974). *Measuring growth in English.* Champaign, Il., NCTE.

Lauer, J.M. & Asher,W.J. (1987). *Composition research: Empirical designs.* Oxford University Press.

Comprehensive Discussions of Reliability Issues in Measurement Theory:

Cronbach, L.J. (1971). Test validation. In R.L.Thorndike (Ed.). *Educational measurement.* Washington: American Council in Education (pp. 443-507).

Kerlinger, F. N. (1986). Reliability. *Foundations of Behavioral* Research. 3rd ed. New York: Holt, Rinehard and Winston. (p. 404-416).

Nunnally, J. (1979). *Psychometric theory.* New York: McGraw Hill (Chapters 6 and 7).

Stanley, J. (1971). Reliability. In R.L. Throndike (Ed.) *Educational measurement.* Washington: American Council in Educational. (pp. 356-442).

Statistical Procedures for Assessing Reliability:

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.

Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16, 407-424.

SPSSX. (1986). *User's Guide*, 2nd. Chicago: McGraw-Hill. (The Statistical Package for the Social Sciences provides both software and extended discussion of the available options.)

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw Hill (pp. 283-289).

PROCEDURES FOR SECURING HIGH RELIABILITY

In the previous section the Intraclass Correlation Coefficient was presented as the most visibly clear means of calculating reliability; also mentioned were problems of interpreting results when researchers do not specify how reliability was obtained. This section will address methods of improving interrater reliability. These methods can be applied during or after training sessions, but they are best used as preparation for rating.

WRITTEN PRODUCTS

Content analysis experts like Krippendorf (1980) and Holsti (1969) advise analyzing only well-defined writing tasks. Here the composition researcher is in trouble since the essay has multiple ways of being developed and organized. (See DeShields, Hsieh, and Frost (1984) for more on essay grading and reliability.) Nonetheless, the researcher can still take the precaution of removing any essays that clearly do not respond to the task. For example, if a persuasive essay was assigned and a student produced an expressive essay, the researcher should remove that essay and not force raters to identify a construct in it that probably doesn't exist. The confusion that would result from trying to score this essay introduces unsystematic error, thereby lowering the reliability. In the previous section, it was stated that reliability is the ratio of true variance to total variance. Increasing the denominator with error variance yields a smaller reliability figure.

Another precaution to take before giving essays to raters would be to check for restriction of range. Although you may wish to generalize to a population of all freshmen writers, your particular student body may not represent them fully. For example, if your admission standards are very high, your freshmen may not reveal national trends even though you have a variety of types of writers in your classes. Scores tend to cluster around a few categories, and raters may have a hard time distinguishing among papers. Raters dutifully try to stretch the papers over the scale, yet they end up quibbling over small details they never would have seen in a more representative sample. In other words, restriction of range means insufficient product variance. Without enough variation in the products, findings are restricted and reliability can be lowered because of rater confusion.[*]

MEASUREMENT INSTRUMENT

The measurement instrument is really the rater, a person sorting written products according to the categories assigned by the researcher. Therefore, the issue of reliability is bound up in many factors. Before discussing the human factors, let us consider the scoring task. Categories should be clearly defined. Raters should be told the basic unit of text to be classified--be it a word, paragraph, theme, essay, or otherwise (Weber, 1985: 22-23). The more objective the scoring task, the higher the reliability because an easy task promotes greater systematic variance (Nunnally, 1971).

The number of categories also influences the reliability. The decision of how many to include depends on how many the raters can perceive. Using the maximum number of categories that

[*] Since reliability can be expressed as $1 - \frac{\text{error variance}}{\text{total variance}}$, we see that restriction of range ("falsely low" total variance) results in a larger quantity subtracted from unity and consequently, an attenuated estimate of reliability.

raters can perceive gives you the most information about your essays, maximizing systematic variance. For example, if raters can perceive 5 categorios of audience-adaptation, and those categories are well defined and easily scored, the researcher will then have more information about the essays; that is, between-essay variance is increased. Raters will tend to achieve higher reliability with 5 categories than if only 3 categories were used.

RATER SELECTION

Since your measurement instrument is the rater, and you need the most precise judgment possible, you want your raters to be experts by selection and by training. Therefore choose raters familiar with the construct you wish to identify. I suggest environmental training methods much like Hillocks (1984) mode of composition instruction; namely a session where you elicit raters' preconceptions about that construct and then build a fresh notion together using their preconceptions and your definitions. I. you see a rater who cannot or will not adjust his or her preconceptions to match your scoring task, do not use that rater. They simply increase unsystematic variance by their inability to internalize that scoring system. SPSSx Reliability softwar helps you detect such raters. It deletes each rater and figures a subsequent reliability (alpha). When a person is deleted and the alpha <u>increases</u>, you know who to remove. You may also wish to figure the reliability of various groups of raters.

Adding raters can dramatically increase the reliability if the scoring task is fairly objective. The Spearman-Brown

Prophecy Formula indicates this fact and reveals that the number of raters is perhaps the most easily adjusted factor a researcher has control of before and after rating. The reason why additional raters can help so much is that as raters are added, systematic variance accumulates faster than non-systematic variance, or error. If the scoring task is too vague however, increasing the number of raters will probably not affect the reliability.

TRAINING PROCEDURES

The goal of training is to build a firm knowledge base. Therefore, the researcher should begin training with sets of anchor papers for each of the categories on the scale. Once raters have a firm grasp of each category, then they can begin practice rating. It would be a mistake to hand raters a mixed pile of essays before they achieve this grasp of their task. Systematic variance, or rater agreement, is enhanced greatly by this firm notion of each category.

Early in training also, the following types of rater errors should be discussed (Corsini, 1984: 205-206):

--halo (one trait influences the scoring)
(balloon handwriting for example)
--carryover (knowledge of student's ability)
(knowing your students, esp. at small school)
--central tendency (hesitancy to score on extremes)

--sequential (order of papers affects scoring)

--recency (emotional influences on raters).
(death of the Challenger crew, personal things)

When raters make these kinds of errors, their scoring patterns become erratic thus lowering the reliability.

Finally, you are ready for practice sessions after the knowledge base is firm and errors have been discussed. Be firm about resolving differences as this builds knowledge. Raters should confer and adjust scores during training sessions only or if their original scores will be retained. Hillocks (1983) provides discussion of this procedure. In the previous section, it was noted how important rater independence is; if raters confer there is no interrater reliability since only one conglomerate score exists.

As a final word, we would like to note how important conditions are for rating. Training and rating sessions should be short, about 2 hours, to avoid fatigue. Refreshments should be provided; raters should be paid a fair wage. Copies of papers should be dark enough and legible. We would also suggest using pairs of raters to add to the interest and reliability of the scoring task.

In closing, we suggest using the above methods for securing higher reliability before the rating sessions and using the Intraclass Correlation Coefficient for assessing reliability.

Works Cited

Corsini, R. J., ed. (1984) "Rater errors," in Encyclopedia of Psychology Vol. 3: 205-206, NY: Wiley.

DeShields, S. M., H. K. Hsieh, and D. Frost (1984) "The measurement of writing skills: some problems and probable solutions." Educational and Psychological Measurement 44: 101-112.

Hillocks, G., Jr. (1983) "Teaching defining strategies as a mode of inquiry." Research in the Teaching of English 17: 275-284.

---. "What works in teaching composition: a meta-analysis of experimental treatment studies." American Journal of Education 93: 133-70.

Holsti, O. R. (1969) Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley.

Krippendorf, K. (1980) Content Analysis: An Introduction to its Methodology. Beverly Hills, CA: Sage.

Weber, R. P. (1985) Basic Content Analysis. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-049. Beverly Hills, CA: Sage.