DOCUMENT RESUME

ED 414 319 TM 027 822

AUTHOR Moore, Alan D.; Young, Suzanne
TITLE Clarifying the Blurred Image: Estimating the Inter-Rater Reliability of Performance Assessments.
PUB DATE 1997-10-00
NOTE 19p.; Paper presented at the Annual Meeting of the Northern Rocky Mountain Educational Research Association (Jackson, WY, October 1997).
PUB TYPE Information Analyses (070) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accountability; *Correlation; Elementary Secondary Education; Generalizability Theory; *Interrater Reliability; Literature Reviews; *Performance Based Assessment; *Research Methodology; Test Use; *Test Validity

ABSTRACT

As schools move toward performance assessment, there is increasing discussion of using these assessments for accountability purposes. When used for making decisions, performance assessments must meet high standards of validity and reliability. One major source of unreliability in performance assessments is interrater disagreement. In this paper, the literature on interrater reliability is reviewed, and a useful and understandable summary of methods is presented for estimating interrater reliability that can be used in performance assessments. Methods of quantifying the degree of interrater reliability are classified into three categories: (1) methods based on bivariate correlation, such as the Pearson product-moment correlation and Spearman's rank correlation coefficient; (2) methods based on the percent of interrater agreement; and (3) methods based on intraclass correlation or by treating raters as a facet in a generalizability study. Examples illustrate use of these methods and issues related to their use. (Contains 5 tables and 17 references.) (Author/SLD)

Clarifying the blurred image: Estimating the inter-rater reliability of Performance assessments

Alan D. Moore, Suzanne Young

University of Wyoming

This paper is prepared for the:
Annual Meeting of the Northern Rocky Mt. Educational Research Association
October 1997, Jackson, Wyoming

Abstract

As schools move toward performance assessment, there is increasing discussion of using these assessments for accountability purposes. When used for making decisions, performance assessments must meet high standards of validity and reliability. One major source of unreliability in performance assessments is inter-rater disagreement. In this paper we review the literature on inter-rater reliability and provide a useful, understandable summary of methods for estimating inter-rater reliability which can be used in performance assessments together with examples which illustrate their use and issues related to their use.

Clarifying the blurred image:

Estimating the inter-rater reliability of performance assessments

The past two decades have seen increased interest in the use of more authentic assessments and particularly performance assessments in education. These assessments are seen to better measure what we want students to know and be able to do than more traditional, often pencil-and-paper multiple choice tests, with which they are most often contrasted. When used informally by teachers to make instructional decisions, there is no more concern for the validity of these assessments than there has been for assessments seen to be less performance-based in the past. Certainly no teachers are encouraged to compute reliability coefficients for performance assessments any more than they have been asked to do so for other forms of assessment. Classroom assessments have never been held to high standards of reliability. But, when performance assessments are used to make decisions about individual students, or about their teachers, or their school, district or state, all assessments whether performance-based or not, must be held to a higher standard of validity (Baker, 1992). The higher the stakes for assessments, the more rigorous must be the evidence for the validity of the assessments.

Reliability of scores is a major necessary condition for the validity of inferences and decisions based on performance assessments no less than for those based on assessments of other kinds. Because performance assessments often have relatively few tasks, consist of complex tasks, and employ more subjective judgments of raters, traditional approaches to estimation of reliability fall short. Since performance assessments almost always use one or more raters to assign scores, or categories to those being assessed, one major potential source of unreliability is inter-rater disagreement. As Frick and Semmel (1978) pointed out, observer disagreement is important because it limits the reliabilities of observational measures. To avoid this limitation, observers should be trained, and criterion-related and intraobserver agreement measures should be used both before and during a study. In these situations, where portfolios, performance, or student products are judged by individuals or groups of individuals, it is still important for the judgments to be consistent.

The purpose of this paper is to review literature and history of the estimation of interrater reliability and to provide a useful, understandable summary of methods for estimating interrater reliability that can be used in performance assessments. We present a clear "user-friendly" presentation of how to carry out the procedures, and examples which illustrate the various methods and issues relating to the methods.

General history of estimation of interrater reliability

The oldest method of indexing the degree of interrater agreement was the use of bivariate correlation coefficients. The use of the Pearson product-moment correlation coefficient was widespread by the earliest 20th century following its introduction by Karl Pearson (1857-1936) (Glass & Hopkins, 1984). When ratings could not be considered to be measured at an interval level, rank correlation coefficient such as that attributed to Charles Spearman (1963-1945) have been used. (See Spearman, 1904). Both these coefficients are used when only two raters are being compared.

In response to the need for an overall measure of rater agreement when more than two raters are used, various intraclass correlation coefficients were developed. The intraclass correlation is the ratio of 'true' score variance to observed score variance. Cronbach and colleagues (1963). According to Cronbach, Pearson originally developed the intraclass correlation. (Cronbach, et al., 1972). The Kuder Richardson formulas, two of which, KR-20 and KR-21 are well-known, for estimating internal consistency were derived by Kuder and Richardson (1937). Later, Cronbach showed that these and other intraclass correlation coefficients were subsumed in a single formula known as Cronbach's alpha. (Cronbach, 1951).

Another line of thought is based on the use of percentage agreement between raters. Here, 100% agreement would be taken to be a high rate of interrater agreement, and 0% would be seen as low. Frick & Semmel (1978) reviewed several of these. Cohen's kappa is often recommended to aid in the interpretation of percentage agreement. But Crocker and Algina (1986) caution that these indices based on percentage agreement are conceptually different from reliability estimates and should not be substituted for reliability estimates.

The newest, most general method of estimating interrater reliability is in the application of generalizability theory to ratings. The method was first described by Cronbach and colleagues (1963). Complete discussions of the theory and examples of its application are in Cronbach, et al., 1972, Brennan (1983), and Crocker and Algina (1986). This method is flexible enough to handle any combination of sources of measurement error, including raters, tasks, occasions, forms of assessment, for example. In this method, what we have referred to as interrater reliability is called *generalizability across raters*. Raters may be considered one of several *facets* in a G*eneralizability study* (G-study) which is used to estimate variance components for each source of measurement error. Once these estimates are made, they can be used to determine how many raters, tasks, or occasions would be necessary to reach desired levels of score generalizability using *Decision studies* (D-studies). *Generalizability coefficients* may be calculated and interpreted in a manner similar to reliability coefficients.

Methods of Measuring Interrater Agreement

The methods of quantifying the degree of interrater agreement can be classified into three categories. The oldest of these are methods based on bivariate correlation such as Pearson product-moment correlations and Spearman's rank correlation coefficient. Second are methods based on percentage of interrater agreement. These can be interpreted either by taking into account the interrater agreement that would be expected due to chance or by interpreting some function of the percentage agreement which does take into consideration the expected agreement due to chance alone. Third, are the methods based on the intraclass correlation or by treating raters as a facet in a generalizability study. The generalizability coefficient for this facet can be considered an index of interrater reliability.

To illustrate the use of the various methods of estimating interrater agreement, we have selected a hypothetical data set from Crocker and Algina (1986). The original data set consists of ratings of three raters on 10 individuals (Table 1). To illustrate how the different reliability statistics change when fewer rating categories are employed, we have collapsed the original data into 4 rating categories (Table 1b) and 2 categories (Table 1c) while maintaining the original rating distribution shapes as much as possible. In addition, to illustrate the effects of a large percentage of classification in one category, we have dichotomized the ratings in Table 1d using a cut-score between 6 and 7.

Correlations

When independent raters are considered two at a time, and if the range of rating categories possible is not too restricted, the Pearson product-moment and Spearman correlations are indices of interrater agreement. But, when the number of rating categories is quite small, say, 2 to 5 categories, which is often the case in performance assessment, the use of correlation coefficients becomes problematic. In performance assessment, it is quite common to ask raters to classify performance responses into those that are on a 4 or 5-point scale. Restriction of the range of variables (in this case the ratings for each rater) lowers the correlation between variables (Glass and Hopkins, 1984) In our example (Table 2) this effect can be seen by comparing the correlations between ratings as the number of rating categories is collapsed. The restriction of range, decreases limits the size of the correlation coefficient in spite of the apparent increase in agreement among the raters.

Another problem with the use of correlation to estimate interrater reliability is that this estimate can be grossly exaggerated because the person by task interaction is included in the numerator of the coefficient in cases where multiple tasks (i.e. performances) are rated by two or more raters (Brennan & Johnson, 1995). It is particularly important to use the generalizability theoretic approach, discussed below, when measuring interrater agreement when more than one task is being rated.

Percentage Agreement

A seemingly straightforward attempt to answer the question, "to what degree do the raters agree," is to calculate the percentage of agreement between two raters. One simply calculates the number of times a pair of judges agrees in their ratings compared to the total number of performance rated. Here, 100% agreement would be taken to be a high rate of interrater agreement, and 0% would be seen as low. However, as Koretz and others (1994) pointed out, " simple agreement rates can be seriously misleading in the case of scales with only a few score points (p. 7). As the number of rating categories decreases, the likelihood of raters agreeing by chance alone increases and the percentage agreement is inflated accordingly. The problem becomes even more severe the more the scoring distributions depart from a uniform distribution. In our example (Table 3) we see the effect of fewer rating categories on the percentage agreement. The effect of radical departure from a fairly unifrom distribution is particularly apparent in the comparison of Tables 3c and 3d.

In order to take into account agreement due to chance, Swaminathan, Hambleton, and Algina (1974) suggested using Cohen's kappa. This index is considered to be a measure of reliability of mastery classification by Crocker and Algina (1986).

Cohen's kappa is

kappa = (P – Pc)/(1-Pc)

where P is the proportion of agreement for two raters and Pc is the chance probability of agreement. Pc is calculated by summing the joint probabilities using each rater's scoring probabilities as marginal probabilities.

For example, in Table 3c, raters 1 and 2 agreed on 9 out of 10 ratings, so P = 9/10 = .90. Rater 1 assigned 6 ratings of 1 and 4 ratings of 2, therefore her marginal probabilities are .6 and .4, respectively. Rater 2 assigned 7 ratings of 1 and 3 ratings of 2. His marginal probabilities are .7 and .3, respectively. The probability, Pc, of their ratings agreeing by chance is Pc = (.6 x ..7) + (.4 x ..3) = .42 + .12 = .54.

Cohen's kappa, for these two raters is kappa = (.90 - .54)/(1-.54) = .36/.46 = 0.78.

The proportion of agreement and chance probability of agreement for our example are displayed in Table 3 together with for these ratings are displayed in Table 3. It is apparent from these tables that kappa adjusts the proportion of agreement downwardly as the chance probability of agreement approaches the actual proportion of agreement

Intraclass Correlation

As discussed above, the intraclass correlation has been used to estimate internal consistency of multi-item tests. In the context of interrater reliability, raters take the place of items, so Coefficient alpha can be interpreted as a measure of interrater reliability. Hoyt (1941) developed a method based on analysis of variance for estimating this same intraclass correlation. For the 2-factor, random effects analysis of variance, our ratings may be represented as a persons-by-raters matrix. (Table 1a). The Persons effect is one factor, and the Rater effect is the other. The analysis of variance for our example is displayed in Table 5.. Hoyt's formula is

Alpha = (MSp – MSr)/MSp

where MSp is the mean square for the person effect and MSr is the mean square residual.

For the original data of our example,

Alpha = (10.310 – 1.043)/10.310 = 0.90

Intraclass correlations are displayed in Table 5 for all the data sets in our example An important characteristic of this statistic in the context of interrater reliability is that equal variances and intercorrelations among the ratings are assumed (Cronbach, Gleser, and Rajaratnam, 1963). The use of a generalizability theoretic approached, discussed next, avoids these assumptions.

Generalizability Theory

Though a complete discussion of generalizability theory is beyond the scope of this paper, its use will be illustrated for the estimation of interrater reliability for our example. In our persons-by-raters matrix (Table 1a), the row and column means for each person and rater, respectively are included. An *effect* is the difference between the grand mean (here, 4.13) and a row or column mean. We can model each rating as the sum of a person effect, a rater effect, and a residual. The residual is simply the discrepancy of the actual rating from what would be expected based on the two means. For example, the decomposition of the rating for Person 1 by Rater 1 is 2 = (4.13 – 2.33) + (4.13 – 4.80) + .87 = 1.80 + (-0.67) + .87. In our generalizability study, we estimate the *variance component* for each effect, each interaction, and the residual. These variance components are estimated using a random effects analysis of variance. In this simple design, a formula for the generalizability coefficient for raters is

Rho-hat-squared = (MSp – MSr)/ [MSp + (ni – 1)MSr]

where MSp is the mean square for persons, MSr is the mean square residual, and ni is the number of raters.

For our example,

Rho-hat-square = (10.310 – 1.043)/[10.31 + (3 – 1)1.043 = 0.75

This generalizability coefficient can be interpreted as an index of interrater agreement. With similar raters, trained in the same way, rating under similar conditions, we could expect the reliability of ratings averaged across the three raters to be. 0.75. Generalizability coefficients for each of the four sets of ratings in our example are displayed in the second column of Table 5. It is clear, by comparing of Cronbach's alpha in the first column with the generalizability coefficients, these two statistics are not the same.

For the purposes of comparison, we have displayed the generalizability coefficients for a G-study in which only Raters 1 and 2 were included. In Table 6 are displayed the correlation, percentage agreement, kappa, Cronbach's alpha, and generalizability coefficient for the ratings of Raters 1 and 2.

The use of generalizability theory in performance assessment

There are many examples of the use of generalizability to estimate interrater reliability for performance assessments. In one of the first uses of portfolios for a state-level assessment, the evaluators of the Vermont Portfolio Assessment Program reported that there was overall satisfaction with the portfolio assessment among teachers and principals. However, the authors (Koretz, Stecher, Klein, and McCaffrey, 1994) continue, "[t]he positive news about the reported effects of the assessment program contrasted sharply with the empirical findings about the quality of the performance data it yielded. Rater reliability was very low in both subjects in the first year of statewide implementation. It improved appreciably in 1993 in mathematics but not in writing. The unreliability of scoring alone was sufficient to preclude most of the intended uses of scores.

Not all performance assessments have been found to have low interrater reliability, however. Linn and Burton (1994) reviewed several performance assessments which have demonstrated high levels of generalizability across raters when well-defined scoring rubrics with intensive training and ongoing monitoring during rating sessions is used. However across-task generalizability is relatively limited

In a recent article, Brennan and Johnson (1995) demonstrated the use of generalizability to assess the relative sizes of the many sources of errors in performance assessment. They used data from a study of performance assessment used in math and science (Shavelson, Baxter, and Gao, 1993.

Summary and Conclusion

Interrater agreement is an important subject of study for performance assessment. As Koretz, et al. (1994) points out, "[a]lthough rater reliability limits the value of the scores derived form an assessment, it is, of course, only one aspect of the broader question of consistency of scores across theoretically comparable instances of measurement, or 'score reliability.' High rater reliability need not imply that score reliability is satisfactory." (p. 7) So interrater reliability is a necessary but not sufficient condition for score reliability in performance assessment. Though correlational statistics, and statistics based on percentage agreement are easy to compute, their use in performance assessment is fraught with problems. Even the intraclass correlation, in the form of KR-20 or Cronbach's alpha is not ideal. Instead, interrater reliability should be studied using generalizability theory. The machinery exists, and is well understood. The good news from the measurement literature related to performance assessment is that high rater reliability is quite possible and feasible with as few as two, and even one rater, if there are specific scoring guidelines and sufficient training for the raters. The bad news is that rater reliability may be the least of our worries. The biggest validity challenge faced by performance assessment is increasingly seen to be score variability due to inadequate task sampling. (e.g. Mehrens, 1992; Shavelson, Baxter and Gao,1993).

References

Baker (1992). The role of domain specifications in improving the technical quality of performance assessment. (Technical Report). Los Angeles, CA: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City: The American College Testing Program.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. Educational Measurement: Issues and Practice, 14, 9-12, 27.

Crocker, L. and Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart, and Winston.

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

Cronbach, L. J, Gleser, G. C., and Rajaratnam, N. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: John Wiley.

Glass, G. V. and Hopkins, K. D. (1984). Statistical Methods in Education and Psychology, 2nd Ed. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Frick, T. and Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48, 157-184.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. Educational Measurement: Issues and Practice, 13(3). 5-16.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Linn, R. L. and Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13, 1, 5-8.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11(1), 3-9, 20.

Shavelson, R. J., Baxter, G. P. and Gao, X. (1993). Sampling variability in performance assessments. Journal of Educational Measurement, 30, 215-232.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Swaminathan, H., Hambleton, R. K., and Algina, J. (1974). Reliability of criterion referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 11, 263-268.

Table 1. Ratings of 10 performances by 3 raters

a. Original Data - 8 categories

Person	Rater 1	Rater 2	Rater 3	Average
1	2	3	2	2.33
2	8	5	7	6.67
3	4	2	2	2.67
4	4	3	6	4.33
5	8	5	5	6.00
6	8	5	7	6.67
7	6	4	5	5.00
8	4	3	3	3.33
9	3	2	2	2.33
10	1	2	3	2.00
Averag	4.80	3.40	4.20	4.13

b. Collapsing to 4 categories

Person	Rater 1	Rater 2	Rater 3	Average
1	1	1	1	1.00
2	4	3	3	3.33
3	2	1	1	1.33
4	2	1	3	2.00
5	4	3	3	3.33
6	4	2	3	3.00
7	3	2	3	2.67
8	2	2	1	1.67
9	1	1	1	1.00
10	1	1	2	1.33
Averag	2.40	1.70	2.10	2.07

c. Collapsing to 2 categories, cut between 4 and 5

Person	Rater 1	Rater 2	Rater 3	Average
1	1	1	1	1.00
2	2	2	2	2.00
3	1	1	1	1.00
4	1	1	2	1.33
5	2	2	2	2.00
6	2	2	2	2.00
7	2	1	2	1.67
8	1	1	1	1.00
9	1	1	1	1.00
10	1	1	1	1.00
Averag	1.40	1.30	1.50	1.40

d. Collapsing to 2 categories, cut between 6 and 7

Person	Rater 1	Rater 2	Rater 3	Average
1	1	1	1	1.00
2	2	1	2	1.67
3	1	1	1	1.00
4	1	1	1	1.00
5	2	1	1	1.33
6	2	1	2	1.67
7	1	1	1	1.00
8	1	1	1	1.00
9	1	1	1	1.00
10	1	1	1	1.00
Averag	1.30	1.00	1.20	1.17

Table 2. Correlations among rating of 3 raters

a. Original Data - 8 categories

	Rater 1	Rater 2
Rater 2	0.91	
Rater 3	0.79	0.83

b. Collapsing to 4 categories

	Rater 1	Rater 2
Rater 2	0.87	
Rater3	0.76	0.58

c. Collapsing to 2 categories, cut between 4 and 5

	Rater 1	Rater 2
Rater 2	0.80	1.00
Rater 3	0.82	0.65

d. Collapsing to 2 categories, cut between 6 and 7

	Rater 1	Rater 2
Rater 2	0.00	1.00
Rater3	0.76	0.00

Table 3. Percentage Agreement and Expected Percentage Agreement

a. Original Data - 8 categories

Percent Agreement

	Rater 1	Rater 2
Rater 2	0	
Rater 3	10	40

Expected Percent Agreement

	Rater 1	Rater 2
Rater 2	9	
Rater 3	6	21

kappa

	Rater 1	Rater 2
Rater 2	-0.10	
Rater 3	0.04	0.24

b. Collapsing to 4 categories

Percent Agreement

	Rater 1	Rater 2
Rater 2	40	
Rater 3	30	50

Expected Percent Agreement

	Rater 1	Rater 2
Rater 2	26	
Rater 3	20	33

kappa

	Rater 1	Rater 2
Rater 2	0.19	
Rater 3	0.12	0.25

c. Collapsing to 2 categories, cut between 4 and 5

Percent Agreement

	Rater 1	Rater 2
Rater 2	90	
Rater 3	90	80

Expected Percent Agreement

	Rater 1	Rater 2
Rater 2	54	
Rater 3	50	50

kappa

	Rater 1	Rater 2
Rater 2	0.78	
Rater 3	0.80	0.60

d. Collapsing to 2 categories, cut between 6 and 7

Percent Agreement

	Rater 1	Rater 2
Rater 2	70	
Rater 3	90	80

Expected Percent Agreement

	Rater 1	Rater 2
Rater 2	70	
Rater 3	62	80

kappa

	Rater 1	Rater 2
Rater 2	0.00	
Rater 3	0.74	0.00

Table 4. Intraclass correlations and generalizability coefficients

	Intraclass correlation	Generalizability coefficient
a. Original Data - 8 categories	0.90	0.75
b. Collapsing to 4 categories	0.88	0.72
c. Collapsing to 2 categories, cut between 4 and 5	0.90	0.76
d. Collapsing to 2 categories, cut between 6 and 7	0.65	0.38

Table 5. Correlations, percentage agreement, kappa, intraclass correlations and generalizability coefficients

	Correlation	Percentage Agreement	Cohen's kappa	Intraclass correlation	Generalizability coefficient
a. Original Data - 8 categories	0.91	0	-0.10	0.90	0.75
b. Collapsing to 4 categories	0.87	40	0.19	0.88	0.72
c. Collapsing to 2 categories, cut between 4 and 5	0.80	90	0.78	0.90	0.76
d. Collapsing to 2 categories, cut between 6 and 7	0.00	70	0.00	0.65	0.38

TM027822

ERIC

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Rocky Mt Res Assoc

Title: Clarifying the blurred image: Estimating the interrater reliability of performance assessments	
Author(s): Alan D. Moore and Suzanne Young	
Corporate Source: University of Wyoming	Publication Date: October 3, 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ Sample sticker to be affixed to document

Sample sticker to be affixed to document ☐

Check here

Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Alan D. Moore	Position: Associate Professor
Printed Name: Alan D. Moore	Organization: University of Wyoming
Address: P.O. Box 3374 Laramie, WY 82071	Telephone Number: (307-766-2071
	Date: 10/3/97

Over

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of this document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents which cannot be made available through EDRS).

Publisher/Distributor:	
Address:	
Price Per Copy:	Quantity Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant a reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name and address of current copyright/reproduction rights holder:
Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

You can send this form and your document to the ERIC Clearinghouse on Assessment and Evaluation. They will forward your materials to the appropriate ERIC Clearinghouse.

ERIC Acquisitions/ RMRA
ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

(800) 464-3742
e-mail: eric_ae@cua.edu