Comparing Reliability Indices Obtained by Different Approaches for Performance Assessments

CSE Technical Report 401

Jamal Abedi, Eva L. Baker, and Howard Herl
CRESST/University of California, Los Angeles

August 1995

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

COMPARING RELIABILITY INDICES OBTAINED BY DIFFERENT APPROACHES FOR PERFORMANCE ASSESSMENTS *

Jamal Abedi, Eva L. Baker, and Howard Herl
CRESST/University of California, Los Angeles

Abstract

The literature has reported many different statistical procedures for assessing interrater, test, and scorer reliability. These procedures are based on different statistical concepts and could provide different estimates of reliability coefficients. In this study, we have examined the efficiency of several statistical techniques under violation of assumptions using two different data sets: (a) a Monte Carlo data set, and (b) a data set obtained from field research. The results of the analyses on both data sets indicate large discrepancies between different indices under violation of certain assumptions. Percent of agreement has been shown to be less stable and more affected by the number of items (raters) and the shape of the distributions. Cronbach's alpha was affected by the number of items, shape of the distribution and the dimensionality of items (raters). The Generalizability approach (G-coefficient) was influenced by the violation of the ANOVA assumption and the number of raters. Also, the Product Moment (PM) correlation coefficient was not consistent under certain violations of assumptions. However, the PM correlation and the G-coefficient were more robust to the violation of assumption when compared with percent of agreement and alpha.

Objective

The purpose of this study was to compare methods used for estimating reliability of a scoring rubric for performance assessment under different conditions (e.g., when some of the assumptions of Product Moment (PM) correlation or ANOVA are violated) and to determine which method is more appropriate for a particular condition.

Background

Different techniques have been used for estimating the reliability of the essay scoring rubric developed at the UCLA National Center for Research on

* Paper presented at the annual meeting of the American Educational Research Association, Atlanta GA, April 1993.

Evaluation, Standards, and Student Testing (CRESST) (see Baker, Freeman, & Clayton, 1991). These techniques are based on different statistical theories and approaches and thus provide different estimates of reliability of the rubric. To get a better picture of the reliability of a scoring rubric, researchers may need to obtain a variety of estimates; however, some of these estimates may not be quite appropriate for certain cases. Among the various techniques introduced in the literature, index of internal consistency (Cronbach's alpha), percent of exact agreement, percent of agreement within explicit point range, generalizability approach, and factor analysis have been used by CRESST's researchers to obtain estimates of reliability of the history and science performance assessment essay scoring rubric.

In the percent of agreement technique, as its name suggests, the number of agreements among raters is counted, then converted to a percentage. This statistic is a distribution-free statistic. It should perform similarly in normal and non-normal cases; therefore, problems like restriction of range and outliers may not have much impact on this technique. On the other hand, Cronbach's alpha may be sensitive to the violation of normality assumption and to number and dimensions of raters (Cortina, 1993). PM correlation and factor analysis techniques may be sensitive to linearity, normality assumptions, and so forth.

As an example of comparing percent of agreement with Cronbach's alpha or PM correlation, assume that all students in a group wrote excellent essays in history, and two raters gave all of the students perfect scores. In this case, there may be 100% agreement but PM correlation may be zero because there is no variability within scores. As another example, suppose that one of the two raters is stringent in rating and usually scores everybody 1 point lower than the other rater. A percent of exact agreement for this case would be zero, but a PM correlation of near 1.0 could result because the ratings changed equally and in the same direction.

Extreme restriction of range could seriously reduce the size of alpha or PM correlation but may not have any effect on percent of agreement. This is also true of distributions with many outliers. The percent of agreement technique, on the other hand, is extremely sensitive to the range of the scores. Percent of agreement computed based on scores with wider range could, in general, be lower than the percent of agreement computed based on scores with tighter range.

In the generalizability technique, the choice between a relative or an absolute decision could make a difference in the size of the G-coefficient. Furthermore, since the G-coefficient is based on analysis of variance, violation of the assumptions of repeated measures ANOVA could seriously affect the results of a G-study.

The literature reports numerous studies using different statistical techniques for estimating reliability of educational measures. Very few, however, have focused on the issue of interrater reliability in performance assessment. Safrit (1976) recommended the use of intraclass correlation as an appropriate measure of interjudge reliability because of robustness of ANOVA its violations of some of the ANOVA assumptions. Based on this recommendation, Stewart and Blair (1982) obtained interjudge reliability by applying intraclass correlation technique to their data.

Data Sources

Data for the Monte Carlo Study

A Monte Carlo approach was used in this study to examine the effects of violations of assumptions of the parametric statistics used in this study.

Data for the Monte Carlo phase of this study were obtained by generating uniformly distributed pseudo-random numbers in the range of 0.0 to 1.0 with a mean of 0.5 using the RAN function in the FORTRAN compiler of an IBM 3090 mainframe computer. These numbers were then transformed to a scale of 1 to 5 to represent ratings. To generate different sets of random numbers, every time the program was run, another set of random numbers was generated and was used as "SEED" of the RAN function. These numbers were then transformed to normally distributed random numbers using Box and Muller's procedure (Box & Anderson, 1955) as follows:

$$X_1 = (-2 \log_e U_1)^{1/2} \cos (2p\ U_2),$$

$$X_2 = (-2 \log_e U_1)^{1/2} \sin (2p\ U_2),$$

where U_1 and U_2 are two uniformly distributed random numbers, and X_1 and X_2 are two normally distributed random numbers.

Since these random numbers were the basis for all subsequent analysis, their randomness was tested using the following statistics:

1. The rectangularly distributed random numbers were transformed to normally distributed random numbers. Numbers of cases falling outside of 10%, 5%, and 1% points were counted. The obtained numbers were 10.20%, 4.98% and 1.10% for the 10%, 5%, and 1% nominal values, respectively, which showed close agreement with the theoretical distribution.

2. The means of several groups of 1,000 random numbers were obtained, and the hypothesis of H: $\mu = 0.5$ was tested. There was not enough evidence to reject the hypothesis of $\mu = .5$ at the .01 level of significance for any group of numbers.

3. The Kolmogorov-Smironov test was applied to determine the goodness of fit of normally distributed random numbers to the theoretical normal distribution. The result was a maximum discrepancy between cumulative proportions of 0.08 with a probability of 0.71, showing very close agreement between obtained and theoretical distributions. The results of these tests showed that the generated data were randomly distributed.

Violation of normality assumptions. Three different distributions were used to test the effect of violation of normality assumptions. These distributions have more commonly been used in educational research.

1. normal distribution (using Box and Muller's procedure);
2. rectangular distribution;
3. J-shaped distribution (chi-square with 2 degrees of freedom).

Data From the Field Study

A group of 250 students completed performance tasks relating to their understanding of Civil War and General Immigration texts. Trained raters scored essays for each of the two topics on six different dimensions (see Baker et al., 1991) plus some other sets of essays. The scores for the first dimension (General

Content Quality) were used for this study because this is a more holistic index of students' understanding of the text.

Procedure

For the Monte Carlo part of this study, scores of four raters on six dimensions of essays for three different topics were obtained. For this purpose, different sets of data were generated with three different types of distributions (normal, J-shaped, and rectangular) with three different levels of interrater agreement (less than 50% agreement, between 50% and 75% agreement, and above 90% agreement). All these conditions were used with two different sample sizes ($N = 50$ and $N = 500$). The reason for including different sample sizes was to see if the size of the sample has any impact on the power of statistical tests used in this study. For each condition (different types of distributions, different levels of agreement, and different sample sizes), 20 replications were generated. The Multi-Approach Reliability System (MARS) software (Abedi, 1993) was then used to compute the five different indices of interrater reliability mentioned earlier (that is, percent of agreement, PM correlations, intraclass correlation, alpha, and G-coefficient). Similarly, the Multi-Approach Reliability System was used to analyze the field study data and provided different estimates of reliability.

Results

Since it is impossible to present all the results obtained from the analyses of Monte Carlo and field study data, we will present samples of the results of the possible combinations of raters. We will discuss the results of this study in two different sections. First, we discuss the results obtained from a performance assessment in history and second, the results of a Monte Carlo study.

Results of the Field Study

Interrater reliabilities were estimated for the field study data by computing percent of exact agreement, percent of agreement within one point, PM correlation coefficients between raters (minimum, maximum, and average correlations), and Cronbach's alpha for each of the possible combinations of 10 raters ($2^{10} = 1{,}024$) and for each of the two topics. A two-facet analysis of variance repeated measures design (with rater and topic facets) was also applied to each of the combinations. A summary ANOVA table was obtained for each combination and

provided the data including sums of squares, mean squares, degrees of freedom, *F*-ratios, and expected values of mean squares. Based on these results, a G-coefficient was computed for each combination. All the computations were done by the Multi-Approach Reliability System (MARS) (Abedi, 1993).

Since it is not possible to present the results for all 1,024 combinations, we have selected a few combinations for our discussion. These combinations were selected based on the raters' background information such as rater's field of study and teaching experience. Combination number 1 was history teachers, number 2 was all teachers and so on. Table 1 presents different estimates of reliability for these selected combinations. As Table 1 indicates, the first set consists of history teachers, the second set is all teachers, the third set consists of two history teachers and one math teacher, the fourth set is all graduate students in history, and the fifth set is all graduate students. The major problems in interpreting the data in Table 1 are unequal number of cases and unequal number of raters for the different combinations selected. Having a very small number of cases in some combinations makes the interpretation difficult; however, one clearly can see that the history teachers have much higher percent of agreement than other groups (see Table 1). This higher level of agreement may be due to the smaller number of raters in the first category, but when the results of this combination are compared with those of the third combination in Table 1, which has the same number of raters, then one can see the higher interrater agreement for the first combination in which all raters are history teachers (percent of agreement of 35.5, average PM of .90, and alpha of .91 as compared with 27.8, average PM of .86, and alpha of .89). The average PM correlation and the alpha coefficient of the first group of raters are about the same as those of other groups, however.

Results of the Monte Carlo Study

As mentioned earlier, reliability coefficients of the different sets of computer generated data were estimated by different statistical techniques. Since it is impossible to present all or even a part of the results in this paper, a few cases were selected for discussions. Tables 2 through 7 present the results for the selected sets.

Careful examinations of these results suggest the following points.

Table 1

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies

Rater group	Number of raters	% of Exact agreement		% of Agreement within 1 point		Average PM		Alpha		G	N	
		T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2		T_1	T_2
History teachers	3	35.48	22.58	79.84	83.06	.90	.83	.91	.85	.83	124	124
All teachers	4	14.81	10.00	67.59	73.00	.91	.89	.92	.89	.85	108	100
2 history teachers and 1 math teacher	3	27.78	17.00	77.78	80.00	.86	.83	.89	.82	.83	108	100
All graduate students in history	5	4.88	1.04	45.45	46.15	.95	.93	.95	.85	—	20	20
All graduate students	6	4.88	1.04	33.33	42.31	.95	.95	.95	.90	—	20	20

Note. T_1 = Topic 1; T_2 = Topic 2; PM = Product Moment correlation coefficient; G = Generalizability coefficient.

With 500 cases and level of agreement above 90%: (Table 2):

- The percent of exact agreement and the percent of agreement within one point are slightly higher for the data sets with fewer raters. Compare, for example, a percent of agreement of 91.60 for two raters to 90.4 for four raters. Alpha is slightly higher for the data sets with more raters. The G-coefficient is considerably higher for the data sets with more raters. Compare, for example, a G of .65 for two raters with a G of .99 for four raters

- The average PM correlations are very similar across different types of distributions.

- Shape of distribution does not have much impact on any of the statistics used to compute reliability when level of agreement is high.

With 500 cases and level of agreement between 50% and 75% (Table 3):

- As the number of raters increases, the percent of exact agreement and the percent of agreement within one point decrease but alpha and G-coefficients increase. Compare, for example, a percent of agreement of 71.2 for two raters with a percent of agreement of 61.8 for four raters in Table 3.

- The percent of exact agreement and the percent of agreement within one point are highest for the J-shaped distribution, lower for normal distribution, and lowest for the rectangular distribution. Compare, for example, a percent of agreement of 77.0 for a J-shaped (two raters) with a percent of 71.2 for a normal distribution.

- Alpha and G-coefficients stay relatively unchanged across the different types of distributions.

Table 2

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Above 90% (*N*=500)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	91.60	97.00	.88	.88	.88	.94	
		2	92.00	96.60	.89	.89	.89	.94	.65
		3	92.80	96.20	.87	.87	.87	.93	
1,2,3,4	Normal	1	90.40	93.40	.86	.92	.89	.97	
		2	90.40	94.80	.88	.92	.90	.97	.99
		3	90.60	93.00	.84	.89	.87	.96	
3 and 4	JSHP	1	94.00	98.00	.84	.84	.84	.91	
		2	94.60	98.80	.90	.90	.90	.95	.66
		3	96.80	99.20	.93	.93	.93	.96	
1,2,3,4	JSHP	1	91.40	97.00	.82	.95	.88	.97	
		2	92.20	98.20	.90	.92	.91	.98	.99
		3	93.40	98.00	.91	.95	.93	.98	
3 and 4	RECT	1	92.00	94.20	.89	.89	.89	.94	
		2	92.80	96.00	.93	.93	.93	.96	.66
		3	92.20	96.20	.91	.91	.91	.95	
1,2,3,4	RECT	1	90.40	90.80	.89	.93	.91	.98	
		2	90.40	92.40	.90	.93	.91	.98	.99
		3	90.60	91.40	.87	.91	.89	.97	

Note. Percent of agreement among raters is 90. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

With 500 cases and level of agreement below 50% (Table 4):

- As the number of raters increases, the percent of exact agreement and the percent of agreement within one point decrease but alpha and G-coefficients increase considerably, much more than in the cases of higher level of agreements. Compare, for example, a percent of agreement of 48.2 for two raters with a percent of 31.6 for four raters; also, compare a G of .47 for two raters with a G of .83 for four raters.

- The percent of exact agreement and the percent of agreement within one point are highest for J-shaped distributions, lower for normal distributions, and lowest for rectangular distributions. Compare, for example, a PA of 65.8 for the J-shaped distribution with a PA of 48.2 for a normal distribution.

- Average correlations are slightly higher for the normal distributions. These results are similar for different numbers of raters.

With 50 cases and level of agreement above 90% (Table 5):

- The percent of exact agreement and the percent of agreement within one point are about the same for the conditions with different numbers of raters.

- Average correlations and alpha coefficients are very similar for different distributions and different numbers of raters.

- The G-coefficients are considerably lower for cases with smaller numbers of raters. Compare, for example, a G of .64 for two raters with a G of .99 for four raters.

Table 2

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Above 90% (*N*=500)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	91.60	97.00	.88	.88	.88	.94	
		2	92.00	96.60	.89	.89	.89	.94	.65
		3	92.80	96.20	.87	.87	.87	.93	
1,2,3,4	Normal	1	90.40	93.40	.86	.92	.89	.97	
		2	90.40	94.80	.88	.92	.90	.97	.99
		3	90.60	93.00	.84	.89	.87	.96	
3 and 4	JSHP	1	94.00	98.00	.84	.84	.84	.91	
		2	94.60	98.80	.90	.90	.90	.95	.66
		3	96.80	99.20	.93	.93	.93	.96	
1,2,3,4	JSHP	1	91.40	97.00	.82	.95	.88	.97	
		2	92.20	98.20	.90	.92	.91	.98	.99
		3	93.40	98.00	.91	.95	.93	.98	
3 and 4	RECT	1	92.00	94.20	.89	.89	.89	.94	
		2	92.80	96.00	.93	.93	.93	.96	.66
		3	92.20	96.20	.91	.91	.91	.95	
1,2,3,4	RECT	1	90.40	90.80	.89	.93	.91	.98	
		2	90.40	92.40	.90	.93	.91	.98	.99
		3	90.60	91.40	.87	.91	.89	.97	

Note. Percent of agreement among raters is 90. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

Table 3

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Between 50% and 75% (N=500)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	71.20	86.20	.57	.57	.57	.73	
		2	70.00	87.40	.58	.58	.58	.73	.59
		3	72.60	89.20	.63	.63	.63	.77	
1,2,3,4	Normal	1	61.80	77.40	.53	.63	.60	.85	
		2	61.80	75.80	.55	.61	.58	.85	.94
		3	61.60	78.00	.57	.63	.60	.86	
3 and 4	JSHP	1	77.00	94.60	.58	.58	.58	.73	
		2	80.20	95.00	.60	.60	.60	.75	.61
		3	78.40	94.80	.59	.59	.59	.74	
1,2,3,4	JSHP	1	66.40	89.80	.56	.67	.60	.85	
		2	68.40	90.80	.57	.65	.59	.85	.95
		3	68.20	90.00	.55	.63	.60	.86	
3 and 4	RECT	1	66.80	80.00	.58	.58	.58	.74	
		2	66.20	79.00	.59	.59	.59	.74	.59
		3	68.80	80.60	.64	.64	.64	.78	
1,2,3,4	RECT	1	61.00	67.00	.58	.67	.63	.87	
		2	60.60	67.40	.59	.66	.62	.87	.94
		3	60.80	66.00	.60	.64	.62	.87	

Note. Percent of agreement among raters is 60. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

With 500 cases and level of agreement below 50% (Table 4):

- As the number of raters increases, the percent of exact agreement and the percent of agreement within one point decrease but alpha and G-coefficients increase considerably, much more than in the cases of higher level of agreements. Compare, for example, a percent of agreement of 48.2 for two raters with a percent of 31.6 for four raters; also, compare a G of .47 for two raters with a G of .83 for four raters.

- The percent of exact agreement and the percent of agreement within one point are highest for J-shaped distributions, lower for normal distributions, and lowest for rectangular distributions. Compare, for example, a PA of 65.8 for the J-shaped distribution with a PA of 48.2 for a normal distribution.

- Average correlations are slightly higher for the normal distributions. These results are similar for different numbers of raters.

With 50 cases and level of agreement above 90% (Table 5):

- The percent of exact agreement and the percent of agreement within one point are about the same for the conditions with different numbers of raters.

- Average correlations and alpha coefficients are very similar for different distributions and different numbers of raters.

- The G-coefficients are considerably lower for cases with smaller numbers of raters. Compare, for example, a G of .64 for two raters with a G of .99 for four raters.

Table 4

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Below 50% (N=500)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	48.20	79.20	.29	.29	.29	.45	
		2	47.00	81.00	.34	.34	.34	.51	.47
		3	51.20	78.60	.29	.29	.29	.45	
1,2,3,4	Normal	1	31.60	56.20	.27	.34	.30	.63	
		2	31.20	60.60	.28	.35	.31	.64	.83
		3	33.20	57.60	.24	.31	.28	.61	
3 and 4	JSHP	1	65.80	92.20	.24	.24	.24	.39	
		2	66.00	91.60	.25	.25	.25	.39	.48
		3	64.60	91.20	.24	.24	.24	.39	
1,2,3,4	JSHP	1	42.60	84.00	.22	.35	.26	.58	
		2	44.40	84.60	.23	.33	.28	.61	.82
		3	44.40	82.20	.22	.32	.27	.59	
3 and 4	RECT	1	42.60	64.00	.20	.20	.20	.34	
		2	43.20	65.00	.25	.25	.25	.40	.50
		3	45.00	67.80	.32	.32	.32	.48	
1,2,3,4	RECT	1	30.60	38.80	.20	.29	.26	.58	
		2	30.80	41.00	.23	.31	.27	.60	.81
		3	31.80	41.40	.20	.32	.26	.59	

Note. Percent of agreement among raters is 30. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

Table 5

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Above 90% (N=50)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	94.00	100.00	.97	.97	.97	.97	
		2	94.00	96.00	.88	.88	.88	.88	.64
		3	96.00	98.00	.96	.96	.96	.96	
1,2,3,4	Normal	1	92.00	96.00	.88	.97	.93	.98	
		2	92.00	94.00	.88	.96	.90	.97	.99
		3	92.00	96.00	.88	.96	.93	.98	
3 and 4	JSHP	1	96.00	100.00	.96	.96	.96	.98	
		2	96.00	98.00	.84	.84	.84	.91	.65
		3	94.00	98.00	.82	.82	.82	.90	
1,2,3,4	JSHP	1	94.00	98.00	.75	.96	.87	.96	
		2	96.00	98.00	.84	1.00	.92	.98	.99
		3	92.00	96.00	.79	.98	.88	.96	
3 and 4	RECT	1	92.00	100.00	.98	.98	.98	.99	
		2	94.00	94.00	.85	.85	.85	.92	.66
		3	94.00	98.00	.97	.97	.97	.98	
1,2,3,4	RECT	1	92.00	92.00	.88	.98	.92	.98	
		2	92.00	96.00	.85	.98	.92	.98	.98
		3	92.00	94.00	.81	.97	.89	.97	

Note. Percent of agreement among raters is 90. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

With 50 cases and level of agreement between 50% and 75% (Table 6):

- The percent of exact agreement and the percent of agreement within one point are higher for the cases with smaller numbers of raters. Compare for example a percent of agreement of 68 for two raters with a percent of agreement of 64 for four raters (Table 6).

- Alpha and G-coefficients are higher for the cases with more raters. The shape of distribution of scores has some impact on the average correlation and alpha. The average correlations and alpha coefficients are highest for the J-shaped distributions and lowest for the rectangular distributions. Compare, for example, an alpha of .98 for the J-shaped distribution (four raters) with an alpha of .96 for the normal distribution (Table 7).

With 50 cases and level of agreement below 50% (Table 7):

With smaller number of subjects and lower agreement rates:

- The percent of exact agreement and the percent of agreement within one point are lower for the cases with more raters.

- The percent of exact agreement and the percent of agreement within one point are highest for the J-shaped distributions and lowest for the rectangular distributions.

- Average correlations are highest for the normal distributions and lowest for the J-shaped distributions.

- The alpha coefficients are very high for the rectangular distributions with 2 raters and very low for the J-shaped distributions with 2 raters.

- G-coefficients are highest for the normal distributions with more raters and lowest for the J-shaped distributions with 2 raters. Compare, for example, a G of .99 for normal distribution (four raters) with a G of .54 for the J-shaped distributions with two raters (Table 7).

Table 6

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Between 50% and 75% (N=50)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	68.00	86.00	.61	.61	.61	.76	
		2	74.00	86.00	.64	.64	.64	.78	.63
		3	74.00	88.00	.71	.71	.71	.83	
1,2,3,4	Normal	1	64.00	80.00	.61	.79	.68	.89	
		2	64.00	80.00	.64	.74	.70	.90	.96
		3	66.00	78.00	.48	.79	.64	.97	
3 and 4	JSHP	1	86.00	98.00	.83	.83	.83	.91	
		2	82.00	92.00	.60	.60	.60	.75	.65
		3	86.00	94.00	.77	.77	.77	.87	
1,2,3,4	JSHP	1	72.00	94.00	.78	.93	.82	.95	
		2	70.00	86.00	.51	.89	.69	.89	.98
		3	74.00	94.00	.77	.90	.84	.95	
3 and 4	RECT	1	70.00	80.00	.64	.64	.64	.78	
		2	66.00	78.00	.60	.60	.60	.75	.60
		3	68.00	78.00	.48	.48	.48	.65	
1,2,3,4	RECT	1	64.00	70.00	.63	.83	.69	.90	
		2	64.00	74.00	.60	.74	.66	.89	.96
		3	64.00	70.00	.48	.83	.67	.89	

Note. Percent of agreement among raters is 60. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

Table 7

Comparison of Different Approaches for Computing Reliability for Content Assessment Studies on the Monte Carlo Data Set, Level of Agreement Below 50% (N=50)

Raters	Distribution	Topic	% of Exact agreement	% of Agreement within 1 point	PM correlations			Alpha	G
					Min.	Max.	Average		
3 and 4	Normal	1	46.00	78.00	.50	.50	.50	.66	
		2	44.00	74.00	.26	.26	.26	.41	.56
		3	58.00	84.00	.54	.54	.54	.70	
1,2,3,4	Normal	1	34.00	54.00	.25	.50	.40	.73	
		2	32.00	60.00	.26	.49	.40	.73	.99
		3	34.00	68.00	.41	.54	.46	.77	
3 and 4	JSHP	1	58.00	88.00	.24	.24	.24	.38	
		2	62.00	84.00	.21	.21	.21	.34	.54
		3	72.00	84.00	.29	.29	.29	.45	
1,2,3,4	JSHP	1	40.00	80.00	.24	.48	.31	.64	
		2	44.00	74.00	.12	.48	.27	.57	.83
		3	42.00	76.00	.10	.42	.29	.61	
3 and 4	RECT	1	40.00	66.00	.45	.45	.45	.94	
		2	48.00	70.00	.44	.44	.44	.96	.66
		3	52.00	62.20	.27	.27	.27	.95	
1,2,3,4	RECT	1	32.00	42.00	.06	.45	.29	.62	
		2	34.00	50.00	.25	.44	.36	.69	.78
		3	32.00	38.00	.15	.55	.30	.63	

Note. Percent of agreement among raters is 30. PM = Product Moment; G = Generalizability coefficient; JSHP = J-shaped distribution; RECT = Rectangular distribution.

Discussion

Different indices of reliability were computed on data sets obtained from two different sources: (a) data sets from a performance assessment study and (b) computer generated data sets. The Multi-Approach Reliability System (MARS) (Abedi, 1993) was used to compute different indices of reliability for all possible combinations of raters for the different data sets. The results showed that the different statistical techniques for computing reliability indices provided estimates which in some cases were very different. Differences were more evident for cases with different numbers of raters, different shapes of distributions, and different numbers of subjects.

In general, when interrater agreement was extremely high, say 90% or higher, and there was a fair amount of variation and sample sizes were large, all different approaches yielded similar results. The shape of distribution and the number of raters did not have much impact on the estimates of reliability (see Table 2).

With relatively large numbers of subjects, when a moderate interrater agreement exists (between 50% to 75%), the percent of exact agreement and the percent of agreement within one point are influenced by the number of raters (see Table 3). For a small number of raters, say two, the percent of exact agreement and the percent of agreement within one point are relatively high. As the number of raters increases, the percent of exact agreement and the percent of agreement within one point decrease. On the other hand, when the number of raters increases, alpha increases, because the coefficient alpha is sensitive to the number of items (raters) (Cortina, 1993). The G-coefficient is also affected by the number of raters because the larger the number of raters, the smaller is the effect of "rater-by-subject" error in computation of a G-coefficient. In Table 3, the G-coefficients are considerably higher for the cases with larger numbers of raters. G-coefficients were not much affected by the shape of the distributions. Among all statistics reported in Table 3, PM correlation coefficients seemed to be less sensitive to the number of raters. All the average correlations stayed at about medium-high range for all combinations under different distributions. These data suggest that an average PM correlation could be a better choice with a relatively large number of subjects when interrater agreement is moderate and number of raters varies across different combinations.

When the number of subjects is small and interrater agreement is high, the discrepancies between different estimates of reliability are a bit higher than in the similar case with a larger number of subjects. With a small number of subjects and a high level of interrater agreement (Table 5), the estimates of reliability are very close to each other, and the impact of number of raters and shape of distribution is minimal for all indices except for the G-coefficient, which is affected by the number of raters. For the data sets with smaller numbers of raters, the G-coefficient is considerably lower. In such cases the percent of exact agreement and the percent of agreement within one point or average correlations could be used.

With small numbers of subjects and moderate levels of agreement (Table 6), which is not unusual in educational research settings, the reliability indices computed in this study seem to be affected by the number of raters and the shape of distributions. The alpha and G-coefficients are more affected by the number of raters than by the shape of the distribution. In such cases among different statistics, the average PM correlation, which is less affected by the shape of the distribution and number of raters, may be used.

The cases with small numbers of subjects combined with low-level interrater agreement created the most discrepancies between different indices of reliability computed in this study (Table 7). Unlike most other cases discussed earlier in this paper, the average PM correlations are higher for the data with normal distribution. Alpha is affected by the number of raters. For the normal and J-shaped distributions the value of alpha increases as the number of raters increases, but for the rectangular distribution, this effect is in the opposite direction.

In general, Cronbach's alpha is sensitive to number of items (number of raters) and item (rater) dimensionality. PM correlation is less affected by the number of raters and provides lower bound estimates of reliability for cases with different numbers of raters and different sample sizes. Percent of exact agreement and the percent of agreement within one point should be used when no systematic differences exist between the ratings provided by different raters; however, it should be kept in mind that percent of agreement is affected by chance agreement and may not be a valid statistic when the level of agreement is moderate to low. Here, Cohen's kappa would be recommended over percent of agreement (Cohen, 1960, 1968).

For the field data, because we did not have data sets with different distributions, the efficiency of the statistical techniques for estimating reliability used in this study could not be compared under different distributions. However, some statements can be made with respect to the number of raters. As Table 1 indicates, as the number of raters increases, the percent of exact agreement and the percent of agreement within one point decrease. The average PM correlation and the alpha and G-coefficients are similar and are not much affected by the number of raters, however.

In further investigations, we plan to use several other field study data sets and compute reliability indices to see how the number of raters and number of subjects affect the estimates of reliability.

In general, the percent of agreement is negatively related to the number of raters: the higher the number of raters, the lower the percent of exact agreement, other things being equal. In contrast, alpha and G coefficient show considerable increase as the number of raters increases.

These results suggest that the statistics proposed in the literature for estimating interrater reliability are all affected by some conditions in the study (e.g., number of raters, sample size, etc.). It would be advisable for researchers to compute different indices and, based on specific conditions of the research, decide on which index to use.

References

Abedi, J. (1993). *Multi-Approach Reliability System* (MARS) [Computer program]. Submitted for publication.

Baker, E. L., Freeman. M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.

Box, G. E., & Anderson, S. L. (1955). *Statistics Social Series B*.

Cohen, J. (1960). A coefficient of agreement for normal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement and partial credit. *Psychological Bulletin, 70*, 213-220.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 1,* 98-104.

Safrit, J. M. (1976). *Reliability theory*. Washington DC: American Alliance for Health, Physical Education and Recreation Publication.

Stewart, M. J., & Blair, W. O. (1982). Agreement among high school diving judges. *Perceptual and Motor Skills, 55*, 3-7.