DOCUMENT RESUME

ED 324 336 TM 015 552

AUTHOR Micceri, Theodore; And Others
TITLE Interrater Agreement: Same Data, Different Definitions, Different Outcomes.
PUB DATE Nov 87
NOTE 23p.; Paper presented at the Annual Meeting of the Florida Educational Research Association (Jacksonville, FL, November 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Data Analysis; Educational Research; *Estimation (Mathematics); *Evaluation Methods; Evaluators; *Interrater Reliability; *Performance; Scaling; Scores; *Test Items
IDENTIFIERS Intraclass Correlation

ABSTRACT

Several issues relating to agreement estimates for different types of data from performance evaluations are considered. New indices of agreement are presented for ordinal level items and for summative scores produced by nominal or ordinal level items. Two sets of empirical data illustrate the performance of the two formulas derived to estimate agreement. Three different agreement estimates, one at the item level and two at the score level, were computed for 2- and 5-point scalings and were compared using data from two human sciences studies: (1) ratings of speech production quality of an English-language consonant pair by 29 Japanese students; and (2) ratings of the quality of published research by 100 judges. In Study 1, each of the subjects was rated on a 5-point, 30-item scale by 4 raters. In Study 2, the judges rated the quality of 50 different research articles on a 5-point, 33-item scale. Intraclass Correlation Coefficient estimates of reliability were computed for each score. Results show that the first formula appeared best for total scores based on a 2-point scale, while for scales with a greater number of scale points, the second formula produced a more conservative estimate of agreement. It is concluded that different situations call for different estimates of consistency. Researchers must determine what score level is of interest prior to conducting agreement/reliability estimates. For multi-point scales, researchers must make sure that the technique used to compute agreement is appropriate for the logic inherent to the rating technique used. Six tables present data from the two studies. (SLD)

Interrater Agreement: Same Data, Different Definitions, Different Outcomes

Theodore Micceri,
Bruce W. Hall
Underbakke, Melva E.

Paper Presented at the
Florida Educational Research Association Conference,
Jacksonville, Fl, November 1987

Abstract

Agreement among raters has become increasingly more important as observational performance evaluations proliferate. Although instrument reliability is essential to theory building, field practitioners are often more interested in consistency among raters. For them, estimates of agreement are more important than estimates of reliability and the effort required to conduct G-type studies is unproductive. Unfortunately, different definitions of agreement may operationally produce substantially different results. Compounding this, agreement indices are almost strictly limited to item level estimates for nominal category data. Such indices are not very accurate at either the item or total score level for many commonly used rating scales. The current paper considers several issues relating to agreement estimates for different types of data and presents new

indices of agreement for ordinal level items and for summative scores produced by either nominal or ordinal level items. Two sets of empirical data are used to illustrate the indices' performance.

Rationale

In the broadest sense, a reliable instrument exhibits stability, consistency, dependability and small errors of measurement on the characteristic being measured. Although all measures include some measurement error, judgements made by humans are especially plagued by this problem. The recent growth of process studies in Psychology, Education, Medicine and Business has generated considerable interest in performance evaluation. Consequently, as observation instruments proliferate, estimates of the consistency of agreement among raters has become increasingly important. Reliability may be defined as the consistency with which an instrument discriminates among (ranks) a group of subjects. A major source of error for observation instruments is disagreement among raters. Interrater agreement may be defined as the extent to which two or more observers, working independently, agree on which phenomena occur to what degree in the target of interest.

In her seminal article, Mitchell (1979) claims that the most common index of rater consistency in observational studies is the interobserver agreement percentage. This technique, although an excellent measure of the absolute

magnitude of one type of error (observer disagreement), provides no information about an instrument's ability to discriminate among subjects. For this reason, it is possible for extremely high interrater agreement percentages to associate with very low reliabilities and vice versa.

Although frequently utilized, interrater agreement is unappreciated, misunderstood and often belittled. Those wishing to apply inferential statistical procedures must use measures exhibiting variability among subjects/cases, since the statistical procedures require this. Thus reliability is a necessity. However, for many field applications, such variability is neither useful nor used. For those seeking purely descriptive information about the presence, absence or quantity of particular attributes, variability among targets may not be required. Professionals such as clinical psychologists, reading specialists personnel managers and educational evaluators, attainment of a criterion and agreement among raters regarding said attainment may be the only issues of interest. For these, the much maligned interrater agreement percentage is both useful and appropriate.

Interrater agreement may be computed in several ways. A simple, traditional technique computes mean item level agreement by assigning a value of 100% for each agreement and 0.00% for each disagreement. Although reasonable for dichotomously scored items (e.g. yes/no), it does not work well for multi-point items (e.g. 1 for very bad and 5 for

very good). Obviously, different raters scoring respectively a 4 and a 5 agree to a greater degree than do raters scoring a 1 and a 5. To obtain an accurate estimate of "TRUE" item level agreement among raters, such information must be considered. Moreover, the level at which data will be aggregated for use in decisions or evaluations is pertinent. Although some uses require item level information, many applications are based on total or sub scores. If item level information is used for decisions or analyses, then it makes sense to estimate reliability/agreement at that level. However, if the unit in analysis is the total or subscore, then it appears reasonable to estimate reliability/agreement at that level rather than the item level. A tangential benefit of score level agreement is the reduced computations/programming required to produce estimates. This, of course, reduces the sources of possible computational error.

If item level agreement is high, then scale level agreement must be high. However, it is possible for item level agreement to be low while scale level agreement is high. For instance, imagine a scale consisting of 20 dichotomous items rating a subject as begin tactful or tactless in specific predetermined situations. Given this scale, it would be possible for each of two raters to disagree on every item, yet both assign the subject a score of 10 (average tact). Their mean item level agreement would

be zero while their scale level agreement would be 100%, since both would consider the subject "average" on tact.

Although perhaps a bit farfetched, the preceding example points to an inherent problem of observation instruments: specifying unique behaviors. Since most, if not all, human behaviors are interdependent; item definitions precise enough to generate perfect agreement must, almost by definition, detail behaviors so specific they provide little useful information to an evaluator or researcher. On the other hand, item definitions involving behaviors capable of producing useful relationships with outcome variables, again, almost by definition, must overlap with other similar items and thereby create item level disagreements among raters.

Thus, if one wishes to know how well raters agree at the score level, agreement must be computed at the score level. This may be accomplished by determining the proportion of possible variability in the measure shared among raters. One may define the proportion of shared variability (agreement) as:

$$A_o = 1 - \frac{X_h - X_l}{\text{RANGE}} \qquad 2.0$$

where:

X_h = higher score among 2 raters

X_l = lower score among 2 raters

RANGE = Maximum minus minimum possible scores

For an instrument having a possible range of scores from 10 to 50 (range = 40), if two different raters assign the same subject respectively scores of 35 and 30, formula 1.0 produces an agreement percentage of .875.

$$A_o = 1 - \frac{(35 - 30)}{40} = .875.$$

Although rational, this technique fails to assess limits in ranges that may occur under conditions of application. For example, although the possible range is 40 points, the range in application may be only 30 points (e.g. 20 - 50). In addition, such limitations are probably indeterminable until an instrument has been widely applied, at which point estimates of its consistency may be superfluous. A more conservative estimate for our example may be computed by determining the proportion of instrument variability actually credited to a subject by raters. Computation involves subtracting a scale's minimum possible score from the obtained score for each rater and dividing the smaller by the larger.

A more conservative estimate for our example may be computed by determining the proportion of instrument variability actually credited to a subject by raters. Computation involves subtracting a scale's minimum possible score from the highest obtained score and comparing rater differences against this difference.

$$A_o = 1. - \frac{X_h - X_l}{X_{Max} - MIN} \qquad 2.1$$

where:

X_h = higher score among 2 raters

X_l = lower score among 2 raters

X_{Max} = maximum score from any rater

For tne preceding example, formula 1.1 produces an agreement estimate of .80.

$$A_o = 1 - \frac{(35 - 30)}{(35 - 10)} = 1 - \frac{5}{25} = .80.$$

Although the minimum (zero) and maximum (100) possible agreements for any pair of raters are the same for both formulae 1.0 and 1.1, formula 1.1 will necessarily assign lower agreement percentages to lower scoring subjects than to higher ones given the same absolute difference between scores assigned by raters. Therefore, as always when estimating reliability/agreement, it would prove useful to select a sample including representatives from high, low and intermediate scoring strata of the population to be evaluated.

A problem arises in the computation of mean agreement if one uses percentages other than 100% and 0%. Unfortunately, the arithmetic simplicity of ratios and proportions masks a subtle pitfall in that the transformation of a scale into a ratio alters the relations among scale points. For any proportion, a value of .50 indicates the denominator is twice as large as the numerator , a value of .25 four times as large

and a value of .125, eight times as large. The complementary ratios greater than 1.00 are 2 to 1, 4 to 1 and 8 to 1. It is obvious that a problem arises here in the measurement assumption of equal intervals and therefore in the interpretation one can apply to a given proportion. If one is interested only in the relationship among ratios with respect to zero, no problem arises. However, if one is interested in the relationship among ratios with respect to the denominator and/or the value 1.00, then the equal interval measurement assumption is false. This confounds data interpretation. How therefore, can one retain symmetry about 1.00 and proportional intervals among values? Just as in the solution to $\int \frac{1}{x}$ the log to the base e comes galloping to the rescue. The $\log_e$ of proportions retains both symmetry about 1.00 and proper relationships among values as shown in Table 1.

Table 1

Proportions, Ratios and the $\log_e$

Ratio	Percent	$\log_e$
1/20	.05	-2.995
1/10	.10	-2.302
1/ 5	.20	-1.609
1/ 4	.25	-1.386
1/ 2	.50	-0.693
1/ 1	1.00	0.000
2/ 1	2.00	0.693
4/ 1	4.00	1.386
5/ 1	5.00	1.609
10/ 1	10.00	2.302
20/ 1	20.00	2.995
Mean	3.92	0.000

A little appreciated aspect of agreement percentages is their inherently conservative nature. For example, agreement among three of four raters produces a percentage of .50 (not .75). This results because six possible agreement pairs occur for four raters (1 & 2, 1 & 3, 1 & 4, 2 & 3, 2 & 4 and 3 & 4). This is particularly true for dichotomously coded items. Thus, for dichotomous items, if one rater codes YES while three code NO, only three of the six pairs agree (50%). This, despite the fact that 75% of the raters agree that the behavior is absent. Given the same situation for a 5 point item with agreement computed as suggested in the methods section of this paper, the conservatism of the estimate depends on the degree of disagreement by the one differing rater. In the situation (maximum), where three assign a code of 5 and one a code of 4, agreement would be .87. However, in the situation (minimum), where three assign a code of 5 and one a code of 2, agreement would be .50. In the first situation, three raters agree that the characteristic is present in an optimum/maximum form, and one that it is present in an almost optimum/maximum form, thus, the agreement percentage of .87 appears appropriate. However, in the second situation, the problem of three coding presence in optimum/maximum form, with one coding presence, but in sub-optimum/minimum form obtains an agreement of only 50%, again, despite the fact that three of four code optimal (75%). Table 2 shows how many raters must agree to produce various percentages of agreement for a dichotomous scaling. This

conservatism issue may be addressed by establishing a set of criteria (TRUTH) for a measure. Generally accomplished using preplanned, demonstration tapes for which a true score has been established by "experts", this technique involves both agreement and validity. If three of four raters agree with the truth, the percentage of agreement

TABLE 2

Agreement Percentages Among Raters for Dichotomous Scaling

Number of Raters Showing Agreement	Percentage of Agreement
3 of 4	50%
4 of 5	60%
5 of 6	67%
6 of 7	71%
7 of 8	75%

is .75 rather than the .50 of all rater pairs. In addition, one may be "sure" the agreement is with truth rather than a result of consistent errors in perceived item definitions among raters.

Assuming one has chosen an appropriate agreement estimate to assess the relevant question, if raters exhibit substantial disagreement, it is not likely that one has a reliable instrument. In addition to mere agreement among raters, reliability requires variability among targets (subjects). Most human sciences researchers are inescapably wedded to reliability estimates, since, in order to evaluate

the theories of their disciplines, one must distinguish among individuals on variables measuring traits such as knowledge and attitude. For the average evaluator, teacher or clinical psychologist, however, it may well be more important that every subject attain an objective, and that raters agree when this occurs. If this is so, then reliability is unimportant, while (1) interrater agreement, and (2) the validity of an instrument's definitions are of primary concern. After all, an agreement of 100% with 100% attainment produces a reliability of zero but suggests successful treatment.

Methods

This inquiry compares several different agreement estimates using data from two "real life" human sciences studies.

1. Ratings of speech production quality of the American English /r/-/l/ contrast by Japanese students of English as a second language, and
2. ratings of the methodological quality of published research by experienced researchers.

In study 1, each of 29 subjects was rated on a 5 point, 30 item scale by 4 raters. A good example of the correct phoneme (r or l) was given a rating of 5, a good example of the incorrect phoneme (r or l) a rating of 1, and an indeterminate phoneme a value of 3. The values 2 and 4 associated with "fairly" good examples of the incorrect and correct phoneme respectively. It is logical to collapse this into a 2 point scale where a 5 (correct and proper phoneme) receives the value of 1 and any other code (incorrect

pronunciation) the value of 0. In this form, percentage of agreement based on 100% and 0% is appropriate.

In Study 2, 100 different judges rated the quality of 50 different research articles on a 5 point, 33 item scale. Each article was rated by a different pair of raters. Although the dichotomizing of items is less logical in this instance, any item rated 5 was designated as 1 while all other scores received a value of zero for purposes of developing a 2 point scaling. In the production of total scores, mean item level scores were substituted for missing values, where raters failed to provide ratings for specific items.

For study 1, reliability was the major issue, since differentiation among subjects was desired. For study 2, the issue of reliability was relatively unimportant, since attainment of a maximum score by all targets would in no way hinder the study's objectives, therefore, interrater agreement was the issue of greatest interest.

Three estimates of agreement, one at the item level and two at the score level were computed for both 2 and 5 point scalings. Scalings at the item level were (1) dichotomous - 100% or 0.00% agreement and (2) multipoint - where the proportion of agreement is represented by the smaller item rating divided by the greater rating with proportions based on the number of scale points minus 1.00. For example, given a 5 point scale, the proportion of agreement for two raters coding and item respectively a 4 and a 5 is .75 as computed below:

$$A_i = \frac{4 - 1}{5 - 1} = \frac{3}{4} = .75$$

The mean item score across all items, all rater pairs and all subjects was used as an estimate of item level agreement for total scores. Additionally, two score level agreements were computed, separately for 2 and 5 point scalings using formulae 1.0 and 1.1.

Intraclass Correlation Coefficient (ICC) estimates of reliability were computed for each score placing the between rater variance in the error term with the between subject variance (BMS) treated as the "TRUE" variability and within cell (WMS - based on differences among raters) treated as error. Two forms of this estimate were computed: 2.0 estimating the reliability of an average judge and 2.1 estimating the reliability of an average of judges:

$$R = \frac{BMS - WMS}{BMS + (k-1)WMS} \qquad 2.0$$

$$R = \frac{BMS - WMS}{WMS} \qquad 2.1$$

where:
BMS = between subjects mean square
WMS = error or within subjects mean square
k = number of raters/judges

Formula 2.0 is appropriate when scores produced by a single rater will be used and formula 2.1 when the mean score

produced by a group of k raters will represent a target's performance (Shrout and Fleiss, 1978).

Results and Discussion

Table 3 shows results of these analyses. It is clear that 2 point scalings generally produce lower estimates of reliability than 5 point scalings no matter what the estimate used. This result supports the work of Green and Rao (1970) who contend that reliability increases as the number of scale points increases up to seven or more.

It is also clear that agreement estimates produced at the score level (1.0, 1.1) tend to be higher and tend to agree more with corresponding ICC reliability estimates than do those at the item level. As suggested, formula 1.1 is more conservative than formula 1.0. Interestingly, for the 2 point scalings, estimates produced using formula 1.1 were substantially lower than those produced using 1.0 because of the presence of very low or zero level scores. If one rater assigns a zero and another a 1 on a scale of 30, the agreement percentage is zero for formula 1.0, despite the proximity of the ratings. This suggests that for 2 point scalings, or any other for which a total score of zero is a strong possibility, formula 1.0 is more appropriate than formula 1.1. For 5 point scalings, however, the estimates produced by 1.0 and 1.1 were close. It should be noted that both of these studies included numerous subjects and/or raters, and that the targets spanned the possible score

ranges of both instruments. The tables in Appendix A show total scores for each target across raters for both studies. A quick glance at these shows the general agreement among raters as well as the considerable variation between targets. This explains the high ICC estimates in Table 3. These ICC estimates both for mean raters and mean of raters more closely correspond to the agreement estimates produced using formula 1.0 and 1.1 than the mean of item level agreements. Appendix B contains the ANOVA tables used to produce these ICC estimates.

TABLE 3

Interrater Agreement and ICC Reliability Estimates

		4 Raters 45 Subjects		2 Raters 50 Subjects	
		Scale Points		Scale Points	
		2	5	2	5
INTERRATER AGREEMENT					
Mean Item Level		.75	.83	.27	.73
Score Level	1.0*	.86	.95	.91	.93
	1.1	.61	.92	.53	.89
ICC Estimates					
Mean Rater	2.0	.83	.95	.85	.88
Mean of Raters	2.1	.90	.98	.92	.93

* formula noted in body of paper

Implications

To reiterate:

1. Different situations call for different estimates of consistency:

   a) INTERRATER AGREEMENT

      i) When purely descriptive information is desired, or

      ii) when attainment of a criterion is of interest.

   b) RELIABILITY

      i) When one wishes an instrument to differentiate among subjects on the trait of interest. This is essential for purposes of comparisons and statistical analysis. In general, when this is of interest, interrater agreement is not, and vice versa.

These results suggest two important issues in the estimation of observer/rater instrument reliability.

1. One must determine what score level is of interest prior to the conduct of agreement/reliability estimates.

2. For multi-point scales, one must be sure that the technique used to compute agreement is appropriate for the logic inherent to the rating technique used.

For these reasons, it appears most appropriate to use score level estimates of agreement for those situations where

total or subscores are applied, and item level estimates only where they are appropriate. Some measures may require both; for instance in evaluation situations, when item level information is used for remediation or feedback while score level data determines criterion attainment or placement.

For total scores based on a 2 point scale or applied frequently to low scoring individuals, it appears best to apply formula 1.0 in the computation of rater agreement, while for scales having a greater number of scale points, formula 1.1 produces a more conservative estimate of agreement.

For those situations in which one wishes to apply what have been termed inferential statistics, the ICC analyses will indicate whether an instrument differentiates consistently enough to produce reasonable information from such statistical applications. These ICC estimates may be computed at either the score or item level, or both if appropriate.

REFERENCES

Green, P. E. and Rao, V. R. (1970). Rating scales and information recovery - How many scales and response categories to use? Journal of Marketing, 84, 33-39.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and the generalizability of data collected in observation studies. Psychological Bulletin, 86, 376-390.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Tinsley, H. E. A. and Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology, 22, 358-376.

Appendix A

Raw Total Scores For Studies 1 and 2

Tables 4 and 5 show total scores across raters for both studies using both 2 and 5 point scalings. It is clear that percentages of agreement for the 2 point scale will frequently have a numerator of zero.

TABLE 4

Study 1 Raw Scores Across Raters for 2 and 5 Point Scales

Rater #				Rater #			
1	2	3	4	1	2	3	4
29	30	27	28	147	150	145	146
30	30	27	29	150	150	147	148
11	2	0	0	81	76	72	71
28	24	21	17	142	137	135	128
15	11	10	10	90	87	85	93
17	11	9	3	102	109	98	99
24	14	2	16	128	111	99	120
29	19	12	22	146	134	127	133
15	8	3	3	90	93	81	83
20	1	7	5	121	76	93	102
3	0	0	2	68	72	82	79
27	17	14	24	144	134	116	136
14	6	4	9	91	89	89	88
14	9	4	9	102	104	95	92
14	13	11	14	86	83	83	88
24	20	18	17	126	119	116	119
29	23	18	23	146	140	135	136
25	26	20	20	142	144	136	131

Table 4 continued

2 Point Scale				5 Point Scale			
Rater #				Rater #			
1	2	3	4	1	2	3	4
30	18	21	25	150	138	139	143
25	26	25	26	133	134	133	142
12	11	6	10	116	120	105	110
29	12	12	22	146	116	126	134
25	26	20	17	133	137	130	118
16	8	6	3	112	109	96	96
21	20	11	27	138	135	119	144
18	14	12	9	102	111	98	109
14	12	10	5	89	98	96	105
14	11	8	3	87	86	83	90
25	14	6	19	134	116	108	126
26	25	15	15	137	138	127	130
28	27	26	21	142	138	141	132
30	30	30	24	150	150	150	144
27	26	21	27	142	142	137	142
27	25	21	27	141	136	132	145
30	28	28	29	150	148	148	148
28	30	28	30	144	150	148	150
15	15	15	9	90	92	90	100
15	13	14	3	90	88	89	90
15	8	9	2	106	94	91	94
21	19	14	6	131	116	118	102
26	19	22	15	134	118	131	120
22	16	12	14	129	120	108	118
27	26	25	20	147	143	145	130

TABLE 5

Study 2 Raw Scores Across Raters for 2 and 5 Point Scales

Two Point Scale				Five Point Scale			
Rater #		Rater #		Rater #		Rater #	
1	2	1	2	1	2	1	2
19	13	0	0	150	143	104	105
2	1	4	3	94	71	132	122
0	0	11	15	80	97	134	137
3	5	0	0	103	115	99	105
7	13	7	14	132	143	119	137
9	3	3	3	126	123	109	113
0	0	6	4	87	85	106	84
1	3	9	19	122	124	128	144
2	1	11	11	91	87	122	116
5	8	5	2	115	121	130	123
9	13	19	10	134	140	141	122
0	0	0	0	117	115	91	85
2	1	3	3	107	113	109	119
3	3	16	13	79	93	145	139
10	13	6	7	131	138	115	121
19	10	7	8	145	138	124	134
23	15	9	9	153	134	125	116
2	3	14	20	116	122	121	137
0	2	8	4	103	114	118	95
2	1	1	0	121	108	98	96
20	15	13	7	152	145	140	130
17	11	13	11	143	133	136	131
12	7	0	1	140	129	90	87
0	0	0	0	80	90	85	8

Appendix B

Source Data for ICC Estimates

Table 5 contains Sums of Squares computed for the two scalings (2 and 5 point) for studies 1 and 2.

TABLE 6

Analysis of Variance for Speech Production Ratings

Source		Df	Sum of Squares	Mean Square
STUDY 1				
2 pt. Scale	BMS	44	10242.300	232.779
	WMS	135	2989.500	22.144
5 pt. Scale	BMS	44	91186.300	2072.416
	WMS	135	6784.500	50.255
STUDY 2				
2 pt. Scale	BMS	49	4752.810	96.996
	WMS	50	382.500	7.650
5 pt. Scale	BMS	49	41971.890	856.569
	WMS	50	2831.500	56.630