Open access • Journal Article • DOI:10.1177/014662168500900103

The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability : A Monte Carlo Investigation: — Source link

Domenic V. Cicchetti, Donald Shoinralter, Peter Tyrer

Institutions: Veterans Health Administration

Published on: 01 Mar 1985 - Applied Psychological Measurement (SAGE Publications)

Topics: Scale (ratio), Reliability (statistics), Rating scale and Inter-rater reliability

The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability: A Monte Carlo Investigation

Domenic V. Cicchetti
West Haven VA Medical Center and Yale University

Donald Showalter
West Haven VA Medical Center

Peter J. Tyrer
Mapperley Hospital, England

A computer simulation study was designed to investigate the extent to which the interrater reliability of a clinical scale is affected by the number of categories or scale points (2, 3, 4, . . . ,100). Results indicate that reliability increases steadily up to 7 scale points, beyond which no substantial increases occur, even when the number of scale points is increased to as many as 100. These findings hold under the following conditions: (1) The research investigator has insufficient a priori knowledge to use as a reliable guideline for deciding on an appropriate number of scale points to employ, and (2) the dichotomous and ordinal categories being considered all have an underlying metric or continuous scale format.

The rationale for this computer simulation study derives from the fact that after more than six decades of theorizing and empirical research (i.e., dating back to the work of Symonds, 1924), the clinical investigator still has no satisfactory solution to the fundamental a priori problem of deciding how many rating categories should be used to define a given clinical scale. In certain areas (such as determining stages of human cataract development), the possible number of categories that can be defined is well agreed on by investigators in the field (e.g., see Cicchetti, Sharma, & Cotlier, 1982; Cotlier, Fagadau, & Cicchetti, 1982; Pirie, 1968). However, in many other areas of biomedical or behavioral science, no specific guidelines exist for the clinical or research investigator to decide on how many categories to employ.

This unfortunate state of affairs reflects itself in the wide variety of response formats that clinical scales typically use. These run the gamut from the simplest "presence" or "absence" of clinical signs and symptoms (Koran, 1975a, 1975b); to the dichotomous-ordinal ("absent," "mild to moderate," "severe," Cicchetti, 1976) category format of the Present State Examination (PSE; Wing, Nixon, Mann, & Leff, 1977); to the 7-category format of the Overall and Gorham (1962) Brief Psychiatric Rating Scale; to the 5- to 9-category formats of various agoraphobia scales (Gelder & Marks, 1966; Watson, Gaind, & Marks, 1971); and, finally, to the "continuous" scales (0 to 100 points) of Aitken (1969) and Remington, Tyrer, Newson-Smith, and Cicchetti (1979).

It should be noted that the scales discussed here all have an underlying metric (continuous or dimensional scale format). It follows that this body of research, as well as the current investigation, has no application to the reliability of categorical scales that have no underlying metric. An example of a clinical scale of this type is the classification schema reported by Hakama, Franssila, and Saxen (1973) for differentiating three basic types of lym-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 1, March 1985, pp. 31–36

0146-6216/85/010031-06$1.55

phatic cancers (Lymphosarcoma, Reticulosarcoma, Hodgkin's Disease), each of which is qualitatively distinct from one another.

Despite a number of experimental, quasi-experimental, and computer simulation studies designed to determine the optimal number of categories to employ for a given clinical rating scale, the conclusions that authors have reached vary considerably. These range from seven as the optimal number (Ramsay, 1973; Symonds, 1924); to seven plus or minus two as the "magic number" (Miller, 1956); to more than seven as the "best" number of categories to employ (Champney & Marshall, 1939). A recent study, Remington et al. (1979), showed that clinical assessments using the PSE (Wing et al., 1977) were as reliable for the aforementioned dichotomous-ordinal category format ("absence," "mild to moderate," "severe" symptomatology), as for either a dichotomous (absence-presence) or continuous format (scale ranging between 0 and 100 points). Finally, the results of two recent computer simulation studies on the problem indicated that reliability increases up to 5 categories, beyond which no further substantial gains are made (Jenkins & Taber, 1977; Lissitz & Green, 1975).

These and other investigations have not resolved the problem of the optimal number of categories for several reasons: (1) In some studies, sample sizes have been relatively small (e.g., ≤ 40 in Ramsay, 1973); (2) other studies have produced findings that are most probably sample specific (e.g., Komorita & Graham, 1965; Remington et al., 1979); (3) the two aforementioned computer simulation studies employed only 100 computer simulations per condition, which may be too few to produce valid results; and, finally, (4) the methodologies have varied so greatly from one study to another that it is not possible to directly compare results across investigations. It should also be stressed that the reliability statistics employed have most often been either coefficient alpha (Cronbach, 1951) or the Pearson product-moment correlation coefficient. Coefficient alpha is the well-known measure of the internal consistency of items within a given test or subtest for which the concept of interrater reliability is not applicable. The Pearson correlation coefficient, though very widely used as a statistic for measuring the level of interrater reliability, has been severely criticized by statisticians since it merely measures similarity in judges' rankings of subjects *rather* than levels of interrater agreement per se (e.g., see Bartko, 1976; Bartko & Carpenter, 1976; Kazdin, 1980; Robinson, 1957).

Method

The present investigation employed monte carlo methodology, appropriate reliability statistics, adequate sample sizes, and large enough numbers of computer simulations to further investigate the optimal number of categories to use when no specific guidelines are available.

The basic question addressed in this research was: How does interrater reliability, under a variety of different conditions, compare for dichotomous, ordinal, and continuous scales of measurement (i.e., $2 \leq k \leq 100$ categories or scale points).

Input Parameters

In order to answer the above question the following input parameters were systematically varied:

1. The scale of measurement: (a) categorical-dichotomous (2 categories); (b) ordinal ($3 \leq k \leq 10$) categories of classification; and (c) continuous (dimensional) scale of measurement (i.e., $15 \leq k \leq 100$ scale points).
2. The average level of simulated absolute interrater agreement: 30%, 50%, 60%, or 70% (on the average), across the main diagonal of a Rater 1 × Rater 2 contingency table. These levels were chosen in order to be consonant with clinical applications; this strategy stands in distinct contrast to allowing levels of interrater agreement to simply vary about chance expectancies.
3. The average proportion of cases in which one simulated rater gave higher scores than the other when the two raters were not in complete agreement: 50/50 split on the off-diagonals, 60/40 split, 70/30 split, or 90/10 split.
4. The sample size or N for each computer simulation was 200, based on the results of monte

carlo research undertaken by Cicchetti (1981) and Cicchetti and Fleiss (1977).

5. Given the very large number of possible rater pairings as k approached 100 (here 10,000), it was considered appropriate to utilize 10,000 as the number of computer runs per simulated condition. In previous research k was also taken into account in deciding on the number of runs to employ (e.g., Cicchetti, 1981; Cicchetti & Fleiss, 1977; Fleiss & Cicchetti, 1978).

This procedure resulted in 240 conditions (10 $\times$ 3 $\times$ 4 $\times$ 2), each based on an N of 200 and 10,000 computer runs per simulation, as well as one condition of 30% absolute rater agreement with a 90/10 off-diagonal split.

Reliability Statistics

The selection criteria for choosing an appropriate interrater reliability statistic were that:

1. It would measure levels of interrater agreement rather than similarity in the ordering of rater rankings,
2. It would correct for the amount of agreement expected on the basis of chance alone, and
3. It could be validly applied to all three types of scales that were investigated (categorical-dichotomous, ordinal, and continuous/dimensional).

Given these stipulations, the statistic of choice was the intraclass correlation coefficient (R/intra), Model II, as described in Bartko (1966, 1974). As shown by Fleiss (1975), R/intra, Model II, with an underlying metric, for the categorical-dichotomous case, with $N - 1/N$ close to 1, is identical to Cohen's (1960) kappa. Also, R/intra, Model II, when applied using a quadratic weighting system, is the same as weighted kappa (e.g., see Cohen, 1968; Fleiss & Cohen, 1973; Fleiss, Cohen, & Everitt, 1969; Krippendorff, 1970).

Output

In comparing the extent to which interrater reliability levels might be affected by the number of scale points or categories, there was interest in noting possible differences in the sizes of R/intra (weighted kappa with quadratic weights), and in levels of both statistical and substantive significance of R/intra values. For the former, the standard Z of R/intra values (a two-tailed test in which values of ± 1.96 are significant at the .05 level, values of ± 2.57 are significant at the .01 level, and values of ± 3 are significant at the .003 level) was utilized. However, since it is well-known that even the most trivial of effects can produce statistical significance at the .05 level (e.g., an R/intra value of only .10 with sufficient size N), at least some rough guide to gauge levels of substantive or clinical significance of resulting R/intra values seemed relevant. A rather simple set of guidelines (due to Cicchetti & Sparrow, 1981, and Fleiss, 1981) was applied, in which the clinical significance of R/intra values was judged as follows: $<$ .40 = poor; .40 to .59 = fair; .60 to .74 = good; and .75 to 1.00 = excellent.

Results

Assessment of Randomness

Before proceeding to the major results of this computer simulation study, it is essential to answer whether the random number generators employed were really producing the simulated conditions they were intended to produce, namely, the four conditions of complete rater agreement (30%, 50%, 60%, 70%, on the average) and off-diagonal splits (representing disagreement cells) of 50/50, 60/40, 70/30, and 90/10, on the average. The latter strategy was in keeping with the recommendations of previous investigations of reliability, as it is affected by number of scale points (e.g., Jenkins & Taber, 1977).

The data examining the four conditions of simulated complete interrater agreement showed that, in fact, the intended and observed, or actual, proportions of interrater agreement were virtually interchangeable, and were in no way affected by either the absolute level of agreement or k, the number of categories of classification:

For 30% agreement—range = 29.95 to 30.07, mean = 29.99;

For 50% agreement—range = 49.93 to 50.05, mean = 50.00;
For 60% agreement—range = 59.93 to 60.01, mean = 60.00; and
For 70% agreement—range = 69.94 to 70.05, mean = 70.00.

The data comparing the intended and observed off-diagonal splits were also quite convincing, with obtained values very closely approximating the expected levels of 50/50, 60/40, 70/30, and 90/10:

For the 50/50 split—range = 49.92/50.08 to 50.11/49.89, mean = 50.01/49.99;
For 60/40 split—range = 59.93/40.07 to 60.07/39.93, mean = 60.00/40.00;
For 70/30 split—range = 69.90/30.10 to 70.14/29.86, mean = 70.02/29.98; and
For 90/10 split—range = 89.94/10.06 to 90.04/9.96, mean = 90.00/10.00.

Major Findings

The pattern of results that occurred remained very consistent over each level of absolute interrater agreement (30%, 50%, 60%, 70%), as well as over the four conditions of off-diagonal splits (50/50 or balanced off-diagonal split, or skews of 60/40, 70/30, or 90/10 on the off-diagonals). Specific findings are given in Table 1 and indicate the following:

1. The level of interrater agreement was always lowest for 2 categories of classification and highest for 100 categories. In some instances the level of agreement based on 2 categories did not even reach statistical significance, whereas this never happened for 3 or more categories of classification.

2. There was always an increase in reliability levels as the number of categories increased with

Table 1
Intraclass r Values and Corresponding Z Values for 50% and 60% Exact Interrater Agreement Levels and Off-Diagonal Splits of 50/50 and 60/40

	50% Agreement				60% Agreement			
	50/50		60/40		50/50		60/40	
K	R/intra[a]	Z	R/intra[a]	Z	R/intra[a]	Z	R/intra[a]	Z
2	-.0004	-0.02(NS)	.01	0.14(NS)	.20	2.91	.21	3.01
3	.25	3.51	.25	3.64	.40	5.65	.40	5.75
4	.33	4.71	.33	4.81	.46	6.59	.47	6.67
5	.37	5.30	.38	5.39	.50	7.05	.50	7.13
6	.40	5.65	.40	5.73	.52	7.33	.52	7.42
7	.41	5.89	.42	5.97	.53	7.54	.53	7.61
8	.43	6.05	.43	6.15	.54	7.66	.54	7.73
9	.44	6.20	.44	6.27	.55	7.76	.55	7.83
10	.44	6.28	.45	6.38	.55	7.85	.55	7.89
15	.46	6.55	.46	6.64	.57	8.08	.57	8.13
30	.48	6.82	.48	6.90	.58	8.28	.58	8.32
50	.49	6.94	.49	7.00	.59	8.36	.59	8.40
100	.49	7.00	.50	7.06	.59	8.42	.59	8.44

[a]As rough guidelines concerning the clinical significance of these R/intra values the criteria of Cicchetti and Sparrow (1981) and Fleiss (1981) were applied, in which: < .40 = poor; .40 - .59 = fair; .60 - .74 = good; and .75 - 1.00 = excellent.

the most dramatic increase being between 2 and 3 categories of classification.

3. However, beyond 7 categories (or scale points), the increases in interrater reliability levels were, relatively speaking, almost trivial compared to the rather dramatic increases in reliability between 2 and 7 scale points. In fact, the strength of interrater agreement was only slightly increased between 7 and 100 scale points (e.g., from .54 [Z of 7.66] to .59 [Z of 8.44] for 60% exact agreement [50/50 off-diagonal split]).

Discussion and Conclusions

The implications of this computer simulation research into the question of the a priori optimal number of categories to employ in a given research situation are quite straightforward. An investigator will sacrifice the most if a dichotomous format is used, will suffer intermediately for using 3 to 6 categories of classification, and will pay the smallest penalty (optimize the odds of producing a more reliable scale) if he/she employs about 7 categories of response. In fact, the differences in scale reliability between a 7-, 8-, 9-, or 10-category ordinal scale, on the one hand, and a 100-point or continuous scale on the other, is trivial not only from a statistical point of view but clinically, as well. These findings are quite consistent with the recent caveats expressed by Cohen (1983), who quantified the substantial losses in information (accuracy) that can occur when a continuous scale of measurement is dichotomized. The results of the present study further indicate that in the context of interrater reliability estimates, 7 ordinal categories of response appear at least functionally interchangeable with as many as 100 such ordered categories.

It is important to stress that the findings deriving from this monte carlo investigation hold under the following two conditions: (1) The research investigator has insufficient information available for deciding what number of scale points is optimal for studying a given clinical phenomenon, and (2) the categorical and ordinal classification systems being considered all have an underlying metric or continuous scale format.

References

Aitken, R. C. B. (1969). Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine, 62,* 989–993.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19,* 3–11.

Bartko, J. J. (1974). Corrective note to: "The intraclass correlation coefficient as a measure of reliability." *Psychological Reports, 34,* 1–11.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83,* 762–765.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease, 163,* 307–317.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23,* 323–331.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129,* 452–456.

Cicchetti, D. V. (1981). Testing the normal approximation of minimal sample size requirements of weighted kappa when the number of categories is large. *Applied Psychological Measurement, 5,* 101–104.

Cicchetti, D. V., & Fleiss, J. L. (1977). Comparison of the null distributions of weighted kappa and the *C* ordinal statistic. *Applied Psychological Measurement, 1,* 195–201.

Cicchetti, D. V., Sharma, Y., & Cotlier, E. (1982). Assessment of observer variability in the classification of human cataracts. *Yale Journal of Biology and Medicine, 55,* 81–88.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency, 86,* 127–137.

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

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

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Cotlier, E., Fagadau, W., & Cicchetti, D. V. (1982). Methods for evaluation of medical therapy of senile and diabetic cataracts. *Opthalmologic Society of the United Kingdom, 102,* 416–422.

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

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651–659.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed). New York: Wiley.

Fleiss, J. L., & Cicchetti, D. V. (1978). Inference about weighted kappa in the non-null case. *Applied Psychological Measurement, 2,* 113–117.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323–327.

Gelder, M. G., & Marks, I. M. (1966). Severe agoraphobia: A controlled prospective trial of behavior therapy. *British Journal of Psychiatry, 112,* 309–319.

Hakama, M., Franssila, K., & Saxen, E. (1973). Reliability of histopathologic diagnosis of malignant lymphoma. *Annals of Clinical Research, 5,* 104–108.

Jenkins, G. D. Jr., & Taber, T. D. (1977). A monte carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62,* 392–398.

Kazdin, A. E. (1980). *Research design in clinical psychology.* New York: Harper and Row.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 4,* 987–995.

Koran, L. M. (1975a). The reliability of clinical methods, data and judgments. *New England Journal of Medicine, 293,* 642–644.

Koran, L. M. (1975b). The reliability of clinical methods, data and judgments. *New England Journal of Medicine, 293,* 695–701.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. In E. G. Borgatta (Ed.), *Sociological Methodology* (pp. 139–150). San Francisco: Jossey-Bass.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A monte carlo approach. *Journal of Applied Psychology, 60,* 10–13.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10,* 799–812.

Pirie, A. (1968). Color and solubility of the proteins of human cataracts. *Investigative Ophthalmology, 7,* 634–642.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38,* 513–533.

Remington, M., Tyrer, P. J., Newson-Smith, J., & Cicchetti, D. V. (1979). Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychological Medicine, 9,* 765–770.

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review, 22,* 17–25.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7,* 456–461.

Watson, J. P., Gaind, R., & Marks, I.M. (1971). Prolonged exposure: A rapid treatment for phobias. *British Medical Journal, 1,* 13–15.

Wing, J. K., Nixon, J. M., Mann, S. A., & Leff, J. P. (1977). Reliability of the PSE (9th ed.) used in a population study. *Psychological Medicine, 7,* 505–516.

Acknowledgments

This study was supported by the West Haven VA Medical Center (MRIS 1416). The authors gratefully acknowledge Joseph L. Fleiss for his many cogent suggestions and helpful critique of the manuscript.

Author's Address

Send requests for reprints or further information to Domenic V. Cicchetti, VA Medical Center, West Haven CT 06516, U.S.A.