ASSESSING THE RELIABILITY OF PSYCHOTHERAPY DATA: UNITS OF ANALYSIS AND SPECIFIC TESTS

Domenic V. Cicchetti, V.A. Hospital, West Haven, Connecticut

A. Introduction

The purpose of this presentation is to review an area which has received only minimal attention in the psychotherapy process literature. This concerns the choice of statistical methods used to assess the reliability of process psychotherapy variables. The major focus is upon two critical issues: first, the unit of analysis and second, the specific statistical tests used to assess the reliability of process psychotherapy variables.

B. The Unit of Analysis Problem

Chinsky & Rappaport (1970) introduce the unit of analysis problem by posing the question: Suppose five therapists have been rated on ten units each of some process psychotherapy variable (for example, accurate empathy (AE)). Is the sample size for independent reliability of judgments of raters 5 or 50? Truax has assumed that the number of interaction units, not the number of therapists, constitutes the number of independent observations. However, Chinsky & Rappaport show that reliability is increased by using few therapists and many patient responses (the typical manner of analyzing process psychotherapy variables), and is decreased by using many therapists and few patient responses (Chinsky & Rappaport, 1970; Rappaport & Chinsky, 1972). These investigators present the following supporting data:

> "Examination of 28 AE reliability coefficients reported by Truax & Carkhuff (1967) indicates that 15 of the 16 highest reliabilities ($r \geq .70$) were obtained when the number of therapists was 15 or less. In only one of the five ratings using more than 15 therapists did the reliability exceed .70." (Chinsky & Rappaport, 1970, p. 381).

This issue which can be labeled as one of inflated reliability coefficients appears to have been caused by a design problem. Specifically, Truax and associates have used actual tape recordings, rather than typescripts, in the assessment of AE by pairs of student clinical judges. In one study four therapists treated 10 patients (Truax, 1970). Each patient was rated six times so that each therapist was heard by the raters 60 times.

This problem was also encountered by Chinsky & Rappaport while training college students to rate accurate empathy responses:

> "In our own attempts to train raters in the use of the AE scale, the experience that stimulated our earlier critique, we found that when we had our raters rate the same therapists several times they reported to us that they could not help but remember how they had rated each therapist previously." (Rappaport & Chinsky, 1972, p. 404).

A recent study was undertaken to test further the contrary arguments presented by Truax on the one hand and Chinsky & Rappaport on the other. Beutler, Johnson, Neville & Workman (1973) reported that the reliability coefficient for therapists' AE behavior based upon the number of therapists was actually higher than that based upon the number of patients and, thereby, claimed support for the position of Truax and lack of support for that of Chinsky & Rappaport. However, their argument appears specious for the following reasons: (1) the authors used transcripts rather than tapes, thereby eliminating the design factor suspected of artificially inflating rater reliability coefficients; (2) Beutler, et al. used the number of patients rather than the usual method of using the number of therapist-patient interaction units, which must, per force, always be larger. This procedure would also mitigate against inflation of interrater reliability coefficients; and (3) finally, the difference between the two rater correlations (those based upon the number of therapists compared to those based upon the number of patients) was not statistically significant.

Truax (1972), in a rejoinder to the criticisms of Chinsky & Rappaport (1970), attempted to justify his approach to the reliability problem. He states that statistics such as the Ebel (1951) intraclass r control for the inflated reliability problem, when, in fact, no available statistical tests can control for the design problem assumed to cause inflated reliability coefficients. The solution offered by Chinsky & Rappaport is for researchers to use either larger numbers of therapists each of whom is rated once or a larger number of raters each of whom rates a given therapist only once. This solution is appropriate and will resolve the design problem which is caused by using actual tape recordings of patient-therapist interactions. If, on the other hand, the design problem is resolved either by using typescripts or a larger number of therapists, then simply presenting the number of psychotherapy variables to judges in a random order would also be expected to resolve this problem. The randomization solution is suggested, whenever possible, by Maxwell (1968, p. 805).

C. Choice of a Statistical Test

1. Overview

An examination of the process psychotherapy literature reveals that a number of the earlier major content analysis systems consist of nominal (or categorical) variables. However, more recent systems are comprised of ordinal scales, as well, and, in fact, one is based upon interval scales. Thus, the gamut of different types of measurement scales, as defined by Stevens (1951), is represented in the scales used to measure process psychotherapy variables.

2. Nominal Psychotherapy Process Variables

A number of major psychotherapy content analysis systems contain one or more nominal variables. Examples include the variable 'type of therapeutic activity' and 'therapist dynamic focus' of Strupp's (1957 and 1966) multidimensional system. The components defining such variables can only be scored in terms of whether they are present or absent.

The typical procedure has been to use either chi square or the simple percentage of agreement for assessing the reliability of nominal psychotherapy process variables.

The major problem with the chi square statistic as a measure of rater reliability is that it measures association of any type and not specifically agreement. As shown by Fleiss (1973a, pp. 144-145) it is entirely possible to obtain a chi square value which is statistically significant in spite of the fact that the actual amount of agreement between raters is occurring at no better than chance expectancy!

Although the simple percentage of agreement obviates the basic problem caused by application of chi square, it used alone is also an inadequate measure because it fails to take into account the amount of agreement expected solely on the basis of chance. These criticisms have been voiced by Fleiss (1973a); Fleiss (1973b); Fleiss & Cohen (1973); Fleiss, Spitzer, Endicott & Cohen (1972); and Spitzer, Cohen, Fleiss & Endicott (1967). These authors recommend an alternative statistical approach which eliminates the serious problems introduced by chi square and the simple percentage of agreement. This statistic is kappa, introduced by Cohen (1960). A later study by Fleiss, Cohen & Everitt (1969) presents a revised, corrected, standard error for the kappa statistic. This latter statistical approach has the following desirable qualities: (1) it allows one to calculate the proportion of rater agreement; (2) it corrects for chance agreement; and (3) it allows one to determine the level of statistical significance of the observed amount of agreement. For all of these reasons, the kappa statistic should be used to assess interrater reliability when the clinical data are nominally scaled.

3. Ordinal and Interval Psychotherapy Process Variables

The statistics employed in assessing the reliability of interval data have not differed from those used for ordinal data.

Examples of ordinal process psychotherapy variables are found in such scales as the 'therapist activity level' scale due to Howe & Pope (1961) and Strupp's (1957, 1960, and 1966) 'depth directedness' and 'therapist initiative' scales. Each of these variables contains 3 or more scale points which are ordered from lower to higher degrees of the quality being measured. For example, Strupp's (1966) 'depth directedness' is measured on a 5 point ordinal scale, with specific criteria defining each point. The scale points are: 'noninferential' (0); 'mildly inferential' (1); 'moderately inferential' (2 or 3, depending upon extent); or 'highly inferential' (4).

The major statistical procedures which have been used to assess the reliability of ordinal process psychotherapy variables are the standard Pearsonian product moment correlation coefficient and Ebel's (1951) intraclass correlation coefficient. In addition, a few investigators have applied the generalized intraclass reliability coefficient due to Horst (1949). This statistic has, for instance, been applied by Bordin (1963) to assess the reliability of his 'free association' scales ('involvement', 'spontaneity', and 'freedom'); and was also used to obtain some of the reliability statistics presented for the Harway, et al. (1955) 'depth of interpretation' scale.

One system which is comprised of interval scale variables is the 'speech interaction' system due to Matarazzo, Saslow & Hare (1958). It consists of the patient and therapist variables mean 'speech duration' and mean 'speech latency'; and the patient or interviewee variable average 'percentage of interruptions'. The statistic used to obtain rater reliability on these scales has also been the Pearsonian product moment correlation coefficient.

The Pearsonian product moment correlation coefficient is perhaps the most widely applied statistic for assessing rater reliability with both ordinal and continuous data. This is especially true in psychological research. In fact, its usage has become common enough for some investigators to define observer agreement in terms of this statistic. As an example:

> "Interrater reliability is simply the product moment correlation between ratings by different individuals." (Overall & Gorham, 1962, p. 808).

In spite of its wide application in the behavioral sciences, the Pearsonian product moment correlation is inadequate as a measure of agreement. The argument was posed succintly by Robinson, as early as 1957, when he said:

> "The Pearsonian correlation is an inadequate measure of agreement because it measures the degree to which the paired values of the two variables are proportional (when expressed as deviations from their means) rather than identical." (Robinson, 1957, p. 19).

The implications of this argument were expressed recently by Cicchetti (1972). Briefly stated, the Pearsonian product moment correlation measures the degree of similarity in ordering of rankings between two independent judges and as such does not focus specifically upon agreement. What is not taken into account is the discrepancy between raters on individual pairs of measurements. As a consequence, slight shifts in ordering of ranks in one observer relative to another

can result in less agreement than between two other observers who may be much farther apart on individual rankings but who, nevertheless, tend to put their rankings in the same order.

The intraclass correlation coefficient, in one form or another, has been proposed as a measure of rater reliability for ordinal and interval data, not only by Ebel (1951) and Horst (1949), but by many other statisticians as well: (Bartko, 1966 and 1974; Burdock, Fleiss & Hardesty, 1963; Fleiss, 1973a and 1973b; Fleiss & Cohen, 1973; Guilford, 1950; Haggard, 1958; and Robinson, 1957). It should be stressed that the analysis of variance model (from which the intraclass correlation coefficient is derived) can be used for measuring agreement with both ordinal and interval data with, in fact, no concomitant violation of assumptions underlying these methods. (See most recently, for example, Gaito, 1974, p. 273).

It should be noted that the formula presented for the intraclass correlation coefficient by Ebel (1951) is inappropriate because the assumptions underlying the model were violated. (See Bartko, 1966, pp. 5-6). The formula given by Horst (1949) also appears inappropriate as a measure of interrater agreement because it does not distinguish between the variance due to differences between raters, on the one hand, and that due to differences among the subjects (e.g., therapists) being rated, on the other, a criticism made by Ebel as early as 1951.

The problem has been greatly simplified by the recent work of Fleiss (1973b). Following upon the work of Bartko (1966) and others, Fleiss defines three basic forms of the intraclass r depending upon what specific reliability questions are addressed by the clinical investigator. Of these three intraclass models the statistic of choice is one which allows the investigator to separately assess: the differences due to subjects (for this purpose, therapists); and the differences due to the variability between pairs of raters. This statistic has the advantages that (1) it is appropriate for both ordinal and continuous data; (2) it can be adjusted to fit the case of 3 or more ratings per subject; and (3) it is also applicable when the number of ratings varies from subject to subject.

A second statistic which is appropriate for ordinal data only is weighted kappa due to Cohen (1968). Its standard error was corrected by Fleiss, Cohen & Everitt (1969). Although it was originally used for weighted nominal data, it can be used with ordinal data by applying an ordered system of weights given by Cicchetti (1972); Cicchetti & Allison (1973); Cicchetti & Fleiss (1975); and Fleiss & Cicchetti (1975). Both Fleiss & Cohen (1973) and Krippendorf (1970) have demonstrated the conditions under which the appropriate form of r intraclass and kappa are mathematically equivalent.

In summary, a review of the psychotherapy process literature indicates that once the inflated rater reliability problem is obviated (following the suggestions of either Chinsky & Rappaport (1970) or of Maxwell (1968)) then the choice of valid statistical tests varies as a function of whether the data are nominal, ordinal, or continuous. Numeric examples of how these statistics can be applied are given in Tables 1-5.

References

Bartko, J.J. The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 1966, 19, 3-11.

Bartko, J.J. Corrective note to: "the intraclass correlation coefficient as a measure of reliability." Psychological Reports, 1974, 34, 418.

Beutler, L.E., Johnson, D.T., Neville, C.W. & Workman, S.N. Some sources of variance in "accurate empathy" ratings. Journal of Consulting and Clinical Psychology, 1973, 40, 167-169.

Bordin, E.S. Response to the task of free association as a reflection of personality. Paper read at Seventh International Congress for Scientific Psychology, Washington, D.C., August, 1963.

Burdock, E.I., Fleiss, J.L. & Hardesty, A.S. A new view of interobserver agreement. Personnel Psychology, 1963, 16, 373-384.

Chinsky, J.M. & Rappaport, J. Brief critique of the meaning and reliability of "accurate empathy" ratings. Psychological Bulletin, 1970, 73, 379-382.

Cicchetti, D.V. A new measure of agreement between rank ordered variables. Proceedings of the American Psychological Association, 1972, 7, 17-18.

Cicchetti, D.V. & Allison, T. Assessing the reliability of scoring EEG sleep records: an improved method. Proceedings and Journal of the Electrophysiological Technologists' Association, 1973, 20, 92-102.

Cicchetti, D.V. & Fleiss, J.L. A comparison of the null distributions of weighted kappa and the C ordinal statistic. Invited paper presented at the Joint Central Regional Meetings of the American Statistical Association, St. Paul, Minnesota, March, 1975.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

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

Fleiss, J.L. Statistical methods for rates and proportions. New York: Wiley, 1973 (a).

Fleiss, J.L. Measuring agreement between two

judges on the presence or absence of a trait. Paper presented at the Joint Meetings of the American Statistical Association, New York City, December, 1973 (b). (Also, in press, Biometrics).

Fleiss, J.L. & Cicchetti, D.V. The non-null distribution of weighted kappa. Invited paper presented at the Joint Central Regional Meetings of the American Statistical Association, St. Paul, Minnesota, March, 1975.

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

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

Fleiss, J.L., Spitzer, R.L., Endicott, J. & Cohen, J. Quantification of agreement in multiple psychiatric diagnosis. Archives of General Psychiatry, 1972, 26, 168-171.

Gaito, J. A review of F.N. Kerlinger's Foundations of Behavioral Research (2nd ed.). New York: Holt, Rinehart & Winston, 1973. In Psychometrika, 1974, 39, 273-274.

Guilford, J.P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1950.

Haggard, E.A. Intraclass correlation and the analysis of variance. Dryden Press, 1958.

Harway, N.I., Dittmann, A.T., Rausch, H.L., Bordin, E.S. & Rigler, D. The measurement of depth of interpretation. Journal of Consulting Psychology, 1955, 19, 247-253.

Horst, P. A generalized expression for the reliability of measures. Psychometrika, 1949, 14, 21-31.

Howe, E.S. & Pope, B. An empirical scale of therapist verbal activity level in the initial interview. Journal of Consulting Psychology, 1961, 25, 510-520.

Krippendorff, K. Bivariate agreement coefficients for reliability of data. In E.F. Borgatta (ed.). Sociological Methodology, San Francisco: Jossey-Bass, 1970.

Matarazzo, J.D., Saslow, G. & Hare, A.P. Factor analysis of interview interaction behavior. Journal of Consulting Psychology, 1958, 22, 419-429.

Maxwell, A.E. The effect of correlated errors on estimates of reliability coefficients. Educational and Psychological Measurement, 1968, 28, 803-811.

Overall, J.E. & Gorham, D.R. The brief psychiatric rating scale. Psychological Reports, 1962, 10, 799-812.

Rappaport, J. & Chinsky, J.M. Accurate empathy: confusion of a construct. Psychological Bulletin, 1972, 77, 400-404.

Rice, L.N. & Wagstaff, A.K. Client voice quality and expressive style as indexes of productive psychotherapy. Journal of Consulting Psychology, 1967, 31, 557-563.

Robinson, W.S. The statistical measurement of agreement. American Sociological Review, 1957, 22, 17-25.

Spitzer, R.L., Cohen, J., Fleiss, J.L. & Endicott, J. Quantification of agreement in psychiatric diagnosis. Archives of General Psychiatry, 1967, 17, 83-87.

Stevens, S.S. Mathematics, measurement, and psychophysics. In S.S. Stevens (ed.). Handbook of Experimental Psychology. New York: Wiley, 1951, pp. 1-49.

Strupp, H.H. A multidimensional system for analyzing psychotherapeutic techniques. Psychiatry, 1957, 20, 293-306.

Strupp, H.H. Psychotherapists in action: explorations of the therapist's contribution to the treatment process. New York: Grune & Stratton, 1960.

Strupp, H.H. A multidimensional system for analyzing psychotherapeutic communications: Manual, (2nd ed.). Chapel Hill: University of North Carolina, 1966 (Mimeo.).

Truax, C.B. Length of therapist response, accurate empathy, and patient improvement. Journal of Clinical Psychology, 1970, 26, 539-541.

Truax, C.B. The meaning and reliability of accurate empathy ratings. Psychological Bulletin, 1972, 77, 397-399.

Truax, C.B. & Carkhuff, R.R. Toward effective counseling and psychotherapy: training and practice. Chicago: Aldine, 1967.

TABLE 1
RATER AGREEMENT IN ASSESSING VOICE QUALITY[1]

Rater B	Rater A 1	2	3	4	Total ($p_{i.}$)
1	.65	.00	.00	.15	.80
2	.00	.10	.00	.00	.10
3	.00	.00	.05	.00	.05
4	.00	.00	.00	.05	.05
Total ($p_{.j}$)	.65	.10	.05	.20	1.00

Note. 1=emotional; 2=focused; 3=externalizing; 4=limited.

$$p_o = \sum_{i=j}^{k} p_{ij} = .8500$$

[1]As defined by Rice & Wagstaff (1967).

TABLE 2
EXPECTED RATER AGREEMENT IN ASSESSING VOICE QUALITY

Rater B	Rater A 1	2	3	4	Total ($p_{i.}$)
1	.5200	.0800	.0400	.1600	.80
2	.0650	.0100	.0050	.0200	.10
3	.0325	.0050	.0025	.0100	.05
4	.0325	.0050	.0025	.0100	.05
Total ($p.j$)	.65	.10	.05	.20	1.00

Note. 1=emotional; 2=focused; 3=externalizing; 4=limited.

$$p_c = \sum_{i=j}^{k} p_{i.}p_{.j} = .5425$$

$$\kappa = \frac{p_o - p_c}{1 - p_c} = .6721$$

$$\hat{Var}_{(\kappa)} = \frac{1}{N\,(1 - p_c)^2}\,[p_c + p_c^2 - \sum_{i=1}^{k} p_{i.}p_{.j}\,(p_{i.} + p_{.j})]$$

$$= .003729284$$

$$S.E._{(\kappa)} = \sqrt{.003729284} = .0611$$

$$Z_{(\kappa)} = \frac{\kappa}{S.E._{(\kappa)}} = \frac{.6721}{.0611}$$

$$= 11.00 \quad (p < .0001)$$

TABLE 3
HYPOTHETICAL DATA ON 50 THERAPISTS TO ILLUSTRATE THE ASSESSMENT OF RATER AGREEMENT WITH ORDINAL DATA

Rater B	Statistic	Rater A 1	2	3	4	5	$p_{i.}$	$\bar{w}_{i.}$
1	a	1	.75	.50	.25	0		.485
	b	.02	.02	0	.04	0	.08	
	c	.0032	.008	.0560	.0064	.0064		
	d	1.05	1.26	1.37	1.17	.92		
	e	.004199	.000066	.004761	.008593	.012188		
	f	.0025	.2601	.7569	.8464	.8464		
2	a	.75	1	.75	.50	.25		.715
	b	.02	.06	.06	.02	.04	.20	
	c	.008	.02	.14	.016	.016		
	d	1.28	1.49	1.60	1.40	1.15		
	e	.000110	.000144	.002391	.005271	.008154		
	f	.2809	.2401	.7225	.8100	.8100		
3	a	.50	.75	1	.75	.50		.895
	b	0	.02	.60	0	0	.62	
	c	.0248	.062	.434	.0496	.0496		
	d	1.46	1.67	1.78	1.58	1.33		
	e	.006368	.003283	.000520	.002162	.004122		
	f	.9216	.8464	.6084	.6889	.6889		
4	a	.25	.50	.75	1	.75		.725
	b	0	0	.04	.02	.04	.10	
	c	.0040	.01	.07	.008	.008		
	d	1.29	1.50	1.61	1.41	1.16		
	e	.011470	.007157	.002510	.000467	.000015		
	f	1.0816	1.0000	.7396	.1681	.1681		
5	a	0	.25	.50	.75	1		.515
	b	0	0	0	0	0	.00	
	c	0	0	0	0	0		
	d	1.08	1.29	1.40	1.20	.95		
	e	.016796	.011470	.005271	.000001	.005898		
	f	1.1664	1.0816	.8100	.2025	.0025		
	$p.j$	.04	.10	.70	.08	.08	1.00	
	$\bar{w}.j$	.565	.775	.885	.685	.435		

Note. As defined in Fleiss, et al. (1969), the six cell entries are as follows: a = w_{ij}; b = p_{ij}; c = $p_{i.}p_{.j}$; d = $\bar{w}_{i.} + \bar{w}.j$; e = $[w_{ij}\,(1 - p_c) - (\bar{w}_{i.} + \bar{w}.j) \times (1 - p_o)]^2$; and f = $[w_{ij} - (\bar{w}_{i.} + \bar{w}.j)]^2$

TABLE 4
DETERMINING RATER AGREEMENT FOR
ORDINAL DATA PRESENTED IN TABLE 3

w_{ij} = (k-1)/(k-1); (k-2)/(k-1); ... (k-k)/(k-1)
(ordinal weighting system)[1]

= 1; .75; .50; .25; and 0

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij} = .8800$$

$$p_c = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.j} = .8092$$

$$\kappa_w = \frac{p_o - p_c}{1 - p_c} \quad \frac{.8800 - .8092}{1 - .8092} = .8711$$

$$\hat{Var}(\kappa_w) = \frac{1}{N (1 - p_c)^2} \left[\sum_{i=1}^{k} \sum_{j=1}^{k} p_{i.} p_{.j} \right.$$

$$\left. x \left[w_{ij} - (\bar{w}_{i.} + w_{.j}) \right]^2 - p_c^2 \right]$$

= .006113506

$$S.E._{(\kappa_w)} = \sqrt{.006113506} = .0782$$

$$Z_{(\kappa_w)} = \frac{\kappa_w}{S.E._{(\kappa_w)}} = \frac{.3711}{.0782} = 4.75$$

[1]See Cicchetti, D.V. (1972).

TABLE 5
HYPOTHETICAL RESULTS OF AN INTRACLASS CORRELATION (R_I)
COMPARING TWO INDEPENDENT RATINGS OF THE PERCENTAGE OF
INTERVIEWEE INTERRUPTIONS DURING A PSYCHOTHERAPY INTERVIEW

Source	df	ss	ms	F	P of R_I
Subjects (S)	63	1800	28.571	30.01	< .001
Raters (O)	1	6	6.000		
Error (E or S x O)	63	60	.952		
Total (T)	127	1866			

$$R_I = \frac{1800 - 60}{1800 + 60 + 2\ (6)}$$ (Fleiss, 1973b)

= .93